Thresholding rules for recovering a sparse signal from microarray experiments

Sabatti, Chiara
Karsten, Stanislav L.
Geschwind, Daniel

2001

Thresholding rules for recovering a sparse signal from microarray experiments

Chiara Sabatti*, Stanislav L. Karsten†, and Daniel Geschwind†

∗ UCLA Departments of Human Genetics and Statistics; † UCLA Department of Neurology;

Running head: Thresholding in Microarrays.

Key words: Minimax; False Discovery Rate; Empirical Bayes; Sparsity; Contaminated data.

UCLA Statistics

Technical Report # 304

[*] Address for correspondence: Dr. Chiara Sabatti, Departments of Human Genetics and Statistics, UCLA, 695 Charles Young Drive South, Los Angeles, CA 90095-7088. e-mail: csabatti@mednet.ucla.edu; FAX: (310) 794-5446; Phone: (310) 794-9567.

Abstract

We consider array experiments that compare expression levels of a high number of genes in two cell lines with few repetitions and with no subject effect. We develop a statistical model that illustrates under which assumptions thresholding is optimal in the analysis of such microarray data. The results of our model explain the success of the empirical rule of 2-fold change. We illustrate a thresholding procedure that is adaptive to the noise level of the experiment, the amount of genes analyzed, and the amount of genes that truly change expression level. This procedure, in a world of perfect knowledge on noise distribution, would allow reconstruction of a sparse signal, minimizing the false discovery rate. Given the amount of information actually available, the thresholding rule described provides a reasonable estimator for the change in expression of any gene in two compared cell-lines.

1 Introduction

DNA-based microarrays represent one of the most exciting technological advances of the last decade. They offer the opportunity of gathering experimental data on the expression levels of thousands of genes in different cells or in different stages of cell cycle. With time, the accuracy of the technology is improving. In the current state, however, there is a significant amount of noise contaminating the results of such experiments. We analyze the simplest form of array experiment, where mRNA from two cell-lines are compared on one slide to identify differences in the expression levels of some of their genes. Additionally, we restrict our attention to the type of comparisons where a small fraction of the genes is expected to exhibit differential expression. In such a context, it is common practice to consider a gene differentially expressed if its expression measurements in the two samples differ by two fold (a log-ratio with absolute value bigger then 0.3). Smaller differences are attributed to random noise. Such thresholding rule is grounded on some empirical investigation conducted in at least one laboratory (De Risi et al., 1997). There are obviously some difficulties in extending the same threshold to the description of experiments carried out in other laboratories, using other materials and instruments that would lead to a different noise level. Moreover such threshold cannot be considered as indicative of statistically significant changes, but has at best the value of a descriptive statistic, capturing the amount of variation between gene expressions observed in the specific experiment.

Accurate statistical analysis of microarray data has just started. Important contributions are due to Dudoit (2000), Tseng (2001), Newton (2001), Kerr (2000). Wong and Newton develop some parametric models that capture some relevant aspects of array experiments, while Kerr suggests a combination of ANOVA analysis and bootstrap resampling of the residuals to obtain confidence intervals of the differences in expression levels for each gene in the two considered samples. While important for a better understanding of the nature of variability in array experiments, these techniques have the inherent limit that they produce gene specific confidence levels, ignoring the important multiple comparison problem that arises when we are considering thousands of genes at the same time. Dudoit describes this problem in detail and suggests the adoption of a step-down correction of the p-values based on permutations.

In this paper, we intend to pursue the following: 1) develop a statistical model that explains the threshold strategy in wide use. 2) Identify a "data-driven" procedure that defines the threshold value in an experiment-specific way. Such a threshold can be effectively used for Exploratory Data Analysis (EDA) purposes. 3) Describe under which conditions such a threshold corresponds to a statistical test and effectively identifies genes whose expression changes with a given level of significance.

The framework that we will use for such attempt is the problem of estimation of a sparse Gaussian mean vector. This is deeply connected with hypothesis testing in the context of multiple comparisons. The rest of the paper is organized as follows: Section 2 gives a description of array experiments, identifies the type of data we are going to study, and discusses possible distributional assumptions. Section 3 presents some recent (Abramovich et al. 2000) theoretical work about asymptotic minimax estimation of a sparse mean vector for multivariate Gaussian. Section 4 illustrates the relevance of this theory in the context of microarray experiments and gives a detailed description of our thresholding rule. Section 5 gives an example of application of our methodology to a novel dataset characterized by high noise level and sparse signal.

2 Gene expression array and statistical properties of the recorded signal

The realization of a microarray experiment spans over a significant amount of time. Our study focuses on the data resulting from the final image analysis. However, in order to understand the adequacy of various distributional assumptions, it is necessary to briefly recall the complex nature of the procedure (see, for example, Geschwind 2000). At any given moment, a cell expresses only a fraction of the genes that are coded by its DNA. These genes are transcribed into mRNA and subsequently translated into proteins. The essence of a microarray experiment consists of measuring the amount of mRNA present at a given moment in time in a cell, obtaining a "snapshot" of the relative expression levels of all the genes. This is done exploiting the complementarity of DNA in an hybridization procedure. A number of DNA fragments, corresponding to genes in the particular cell line under study, are attached in different predetermined spots on a glass slide. Fluorescent labeled complementary DNA, synthesized from the mRNA extracted from a cell, is hybridized onto the slide. Each labelled cDNA molecule will migrate on the surface by diffusion to find the spot that contains its complement and attach there. After subsequent washes, by measuring the fluorescence of all the spots, one can get a grasp of the quantity of mRNA that was present in the cell. The cDNAs obtained from two different cell lines are labeled with different dyes and co-hybridized onto the slide: this allows to control for non uniform quantity of spotted DNA and differential success of the hybridization, and allows gathering data on expression levels of two cell-lines or tissues using only one slide. The ratio of the signal intensities of the two dyes at any given spot is taken as a measure of the relative abundance of the mRNA for that gene in the two cell-lines.

If the above represents the road-map of a microarray experiment, the actual journey is quite more complex. We can distinguish two "preparatory phases" and the final hybridization with data collection. One of the preparatory phases involves the construction of the array: the identification of DNA fragments corresponding to various genes, the preparation of solutions that contain these fragments, each with the same concentration, and the spotting of these fragments on an array. There is room for variability in all these stages. Consider for example the spotting: it is typically sensitive to the printer type, the pen type, the length of the printing cycle, the quality of maintenance of the instruments. It is quite common that arrays are printed in batches, so that similar "production defects" may be shared by all the arrays in the batch, in addition to slide specific factors. Also, it is quite common that the slides are divided in quadrants for printing purposes: the same pen prints all the spots in a quadrant, moving along them. This may introduce quadrant effects (one pen behaves differently from the others) and spatial effects within the quadrant (efficiency of a pen in a specific position). Printing of arrays is generally done in specialized laboratories, and consistent efforts are devoted to minimize all the sources of variability.

The other preparatory phase consists in obtaining labeled cDNA from the two cell lines to be compared. This process is most often conducted by the end-users of microarrays: laboratories interested in the study of particular cell types or processes that desire to acquire gene expression data. Just because there are much more laboratories that use microarrays than laboratories that produce them, this process is characterized by higher variability. Here, we provide only few examples of the steps involved. mRNA has to be extracted from cells: often the quantity of cells available doesn't lead to enough mRNA to conduct the experiment and it is necessary to conduct an amplification process, which may skew the relative abundance of mRNA fragments (typically some low-abundance mRNAs may not get amplified). Additionally, the cDNA obtained from the mRNA has to be labeled and the success of this process is quite variable.

The final stage of the experiment consists in the hybridization and data collection. Sources of variability include, but are not limited to: temperature and duration of the hybridization; different buffers to prevent non specific binding; successive scans of the same slide, that can lead to images that are not perfectly overlapping; software analysis packages, which will produce different background and signal measurements and different quality checks (depending, for example, on the segmentation technique used to define the spots).

The final product, which will be the target of our analysis, is a measure of signal for each spot and each dye and a corresponding measure of local background. Let g_i be the green intensity level measured for spot i in a given array: $g_i \geq 0$ and by inspecting the distribution of intensities (Figure 1 gives one example), one can notice that it is heavy tailed, consistent with the hypothesis that intensities grow exponentially with the number of molecules bound to the spots. It is then natural to analyze the logarithm of g_i to obtain a measure that grows linearly with effective amount of expressed genes. However, attention has to be paid to what is the appropriate "origin" of the data: the measured amounts are the results of multiple contributions, some of which change slowly if at all, while one has a substantial value, almost constant across the data, what is appropriately called background. As Tukey (1977) very clearly points out, omitting to correct for background will introduce an artificial curvature in the logarithm of the signal values. For this reason, a background correction is an appropriate standard practice. However, caution should be paid with regard to how background values are obtained and, in particular, to the noise level of their estimates. An informative discussion on the strategies of background and signal estimation from image files can be found in Yang (2000). A last preprocessing step, often called normalization, relates to the fact that the two dyes commonly employed have differential incorporating power, so that a difference in

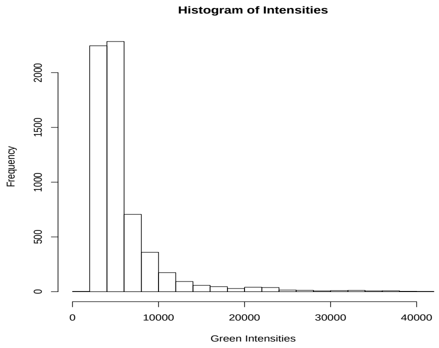

Figure 1: Histogram of the green intensities from the calibration experiment in De Risi et al. (1997)

intensity measured at a spot may be due to a dye effect, rather than to a difference in the mRNA amounts. Normalization is of crucial importance especially in the presence of a signal amplification procedure, which increases the amount of dye incorporated in a molecule (as in the data we report about in section 5). Dudoit (2000) and Schadt (2000) provide useful suggestions for normalization. After this last step, one can calculate the log-ratios of expression for each spot on the array, pairing the data to correct for differences in amount of printed and hybridized cDNA across spots. Let us indicate with y_{ik} the log-ratio of the intensities of spot i on array k. Our statistical analysis will be based on y_{ik}. Some groups have taken a different approach, working directly with the signals coming from the green and red dye and normalize and pair the data while estimating changes in expression (Kerr et al. (2000), Newton et al. (2001)). With reference to Kerr's ANOVA model, we found the dye effect too complex to be effectively described by an additive term and we prefer a preliminary non-linear normalization. We find valuable the suggestion of Newton that there is information in the size of the signal: the normalization technique that we follow uses signal size as one of the inputs.

We want to construct a statistical model for the log-ratios y_{ik} of the background corrected and normalized signals. Even the schematic description of microarry experiments given above demonstrates that there are numerous possible sources of errors in the procedure, and more specifically that the procedure alternatively involves operations on scales of different sizes. Therefore, it is quite possible that a source of experimental error ends up having a very big effect, rather that acting as minor random noise. Equivalently, whatever statistical model one wishes to use, one will have to acknowledge the presence of outliers. Detecting and eliminating outliers in this context is rather difficult, as they are completely confounded with the signal.

In general, the distribution Φ_{ik} of each log-ratio y_{ik} will depend on the gene i and the comparison k; because we are working with normalized, background corrected and paired signals, we can assume that other factors as spot position, dye etc. have only a second order influence on the distribution and can be excluded from the model. With regard to the dependency of Φ_{ik} on k, in the present work, we are restricting ourselves to the analysis of cases where there is no reason to hypothesize subject effects as all the cells compared descend from the same ancestor; we additionally assume that background correction, normalization and pairing of the signals allowed us to control slide effects, so that $y_{ik} \sim \Phi_i$. The smaller the number of parameters needed to describe the differences between the Φ_is, the higher the power of any statistical procedure. Clearly, each gene i will have a different location parameter θ_i, capturing the different amounts of expression. Additionally, there is no reason a priori to assume, for example, that the variances should be the same across genes. However, given the scarce amount of data available (k is at most 5-6), one has to be careful in expanding the details of the parametric description of Φ_i. One valuable source of information on the distribution of y_{ik} is what is often called a calibration experiment: cDNA obtained by the same sample and independently labeled with the two dyes is hybridized onto the same array, leading to log ratios x_ik that have expected value equal to zero. If we assume that the Φ_i differ from each other only because of a location parameter $\Phi_i = \Phi(\theta_i)$, then one can use the calibration experiment to estimate even nonparametrically the distribution Φ. If one assumes a particular functional form for Φ, one can use the calibration data to estimate nuisance parameters of Φ_i. We are interested in two specific models

1. contaminated normal:

$$y_{ik} \sim \mathcal{N}(\theta_i, \sigma_i)(1 - \varepsilon) + \varepsilon F,$$

   where F represent the contamination component and ϵ the contamination probability.

2. nonparametric:

$$y_{ik} \sim \Phi(\theta_i),$$

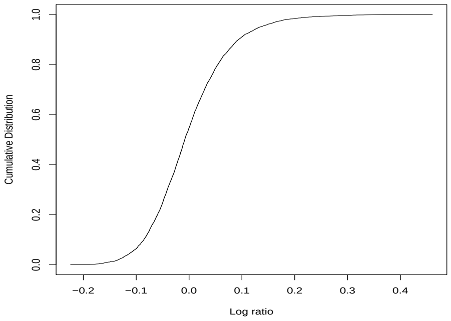

Figure 2: Empirical cumulative distribution function of log-ratios from calibration experiment in De Risi et al. (1997)

where θ_i is a location parameter.

The normal model with contamination offers a flexible description of the data, allowing for a different variance for each gene. It is, however, useful only when the contamination is small, so that the normal distribution still approximates satisfactorily the data. In both situations, we are interested in the recovery of θ_i. A crucial observation is that in the context of many array experiments, it is possible to establish a priori the fact that most of the θ_is are truly zero, that is the parameter vector of interest is sparse. This hypothesis links the estimation problem we described to the simultaneous testing of multiple hypothesis $H_0 : \theta_i = 0$ and to the general problem of model selection. In the following section we will illustrate a beautiful theory that tackles the estimation of a sparse mean vector when the data are multivariate Gaussian. This will provide a theoretical framework that explains the effectiveness of thresholding and suggests heuristic procedures for microarray analysis.

3 Asymptotically approximate minimax estimation of a sparse mean vector for a Gaussian distribution

Let us consider $y_i \sim N(\theta_i, \sigma_i)$, with $i = 1, \ldots, n$. We assume σ_i known and want to recover θ_i. The vector θ of means is said to be sparse, if there is a relative small proportion of non zero coordinates. There are various ways of making this notion precise, the results that we will quote adapt to various sparsity definitions; to fix ideas we will consider θ sparse if there exists a constant η such that $(\sum_{i=1}^{n} |\theta_i|^p/n)^{\frac{1}{p}} \leq \eta^p$ with $p < 1$. We are interested in considering estimation procedures that are simultaneously asymptotically approximately minimax. That is, we want to minimize the squared loss $\sum_i(\theta_i - \hat{\theta}_i)^2$ and we are interested in estimators whose risk approach that of a minimax estimator as $n \longrightarrow \infty$. This criteria to select estimators has been advocated by Donoho and Johnstone (1994a) to whom much of the results that we are going to quote are due. Note that the asymptotics does not refer to multiple observations on one random variable, but to an increasing number of random variables and parameters. Donoho and Johnstone (1994a) proved that the following thresholding estimator satisfies the requirements:

$$\hat{\theta}_i = \begin{cases} y_i & |y_i| > \sigma\sqrt{2\log(n)} \\ 0 & \text{otherwise} \end{cases} .$$

They also showed (Donoho and Jonhstone, 1994b) that if the degree of sparsity in the vector was known, one would obtain better results defining a threshold level that depends on such degree. For example, say that it is known that there are $n_0 = n^\beta$ with $0 < \beta < 1$ non zero components of θ, then, a threshold value of $\sigma\sqrt{2(1-\beta)\log(n)}$ would be appropriate. However, the degree of sparsity of a problem is generally not known in advance. Abramovich et al. (2000) show that an adaptive thresholding procedure constructed to control the false discovery rate is asymptotically minimax and adapts to the degree of sparsity (the false discovery rate–FDR–is the expected value of the proportion of null hypothesis that are erroneously rejected over the total number of rejected null hypothesis; this approach to multiple testing was introduced by Benjamini and Hochberg, 1995) . Let $|y|_{(1)} \geq |y|_{(2)} \geq \cdots \geq |y|_{(n)}$ be the ordered magnitudes of the observations, Let t_k be the right tail quantile of a normal distribution $t_k = \sigma z(q/2k/n)$, where q is the desired FDR. Let k^*_{FDR} be the largest index k for which $|y|_{(k)} > t_k$; then the thresholding scheme

$$\hat{\theta}_k^{FDR} = \begin{cases} y_k & |y_k| > t_{k_{FDR}} \\ 0 & \text{otherwise} \end{cases}$$

is inherently adaptive to unknown degree of sparsity. Note that as this adaptive thresholding scheme is connected with the problem of simultaneously testing various hypothesis controlling the FDR, the universal threshold of $\sqrt{2\log(n)}$ can be shown to be loosely connected to a Bonferroni approach to multiple comparison.

These results in minimax estimation of a sparse mean vector for multivariate normal data have interesting implications in the statistical analysis of microarrays. They suggest that thresholding is a winning strategy and give a situation where it is definitely preferable to linear estimators. They clearly indicate how the thresholding level should depend on the number of random variables, and they indicate one possible strategy for adaptive determination of the threshold. We turn now to examine the practical relevance of this approach.

4 Thresholding for EDA and signal recovery in microarray experiments

The results of an array experiment are often used as inputs of various statistical procedures: clustering of genes based on the similarity of their expression, classification of cell-types, identification of regulatory mechanism. Obviously, it is important that the values of expressions used in such procedures are as noise free as possible. By setting to zero the θ_i's corresponding to the y_i's that do not show enough evidence of variation, one avoids spurious over-fitting. This simple consideration argues in favor of a simultaneous estimation of the θ_i. Moreover, simultaneous estimation of a sparse mean vector is deeply connected with the simultaneous testing of the hypothesis $H_0 : \theta_i = 0 \quad \forall i$, which is also a very natural question in array experiments. The asymptotic minimax paradigm outlined in the previous section also corresponds to the specific characteristics of array experiments where a large number of genes n is studied, with very few observations per gene. By doing asymptotics on n, the number of genes, rather than on the number of replicates, this paradigm exploits the high dimensionality of the problem.

One first application of the model we have described can be found in interpreting the 0.3 cut-off value (2 fold change) that is wildly used in the literature. At this level, we are simply advocating the use of such threshold methods for exploratory data analysis purposes. If log-ratios from microarray experiments are normal with constant variance, and if we take a representative number of 6000 genes per slide, the above universal cut-off value corresponds to a standard deviation of 0.072. A

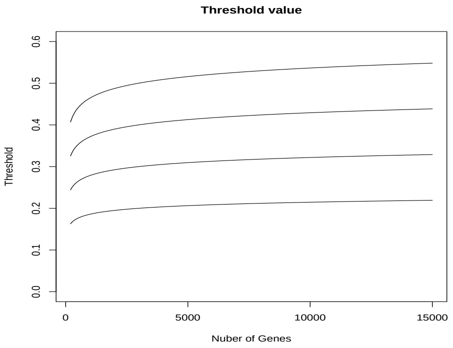

Figure 3: Values of the universal threshold as a function of the number of compared genes and the noise levels. The standard deviation values considered are 0.05, 0.075, 0.10, 0.125.

reasonable estimate of the noise level of the particular experiment at hand can be gathered from the calibration experiments using the sample standard deviation of $(x_1, \dots x_n)$, or some more robust estimators, as the one based on the median absolute deviation. Indeed, if we take the data coming from the original paper of De Risi et al. (1997) and look at their calibration experiments (whose distribution function was shown in figure 2), where $n = 6153$ and σ can be estimated to be 075, leading to an universal threshold of .31. Hence, the parametric expression of the cutoff value that we suggested produces a result which is remarkably close to the the one established heuristically on this set of arrays. The parametric expression of the cutoff value, allows us to modify it as the number of genes spotted on the array and the noise level of the experiment vary. Figure 3 illustrates how this universal cutoff varies with n and σ. Variations in the number of genes spotted have a very mild effect. Variations in σ have a much bigger impact. The variability of expression values depends on the quality of the spotted array, the preparation of cDNA used for the hybridization, the type of buffer, the skills of the researcher: it is quite possible that σ varies considerably between laboratories and experiment types. If the scientist's goal is to summarize the amount of variation observed in the experiment in a way that is indicative of the real amount of change of expression and yet easy to communicate and compare with others, the use of the data-dependent parametric form of the thresholds seems particularly convenient.

Once some signal is detected, one needs assess how much of it is real and how much due to random noise. This has been a formidable problem in microarray experiments: the amount of repetition is so small and the sources of variations are so many that it is difficult to construct a model that realistically describes the data and then have enough information to estimate its parameters. Newton et al. (2001) suggest modeling the signals of the two dies directly with gamma distributions assuming constant coefficient of variation across genes. Kerr et al. (2000) also propose working with the separate signals data and estimate dye and array effects together with gene and cell-line effects in an anova model using the bootstrap to obtain confidence intervals. Tseng et al. (2001) suggest using the signals from the two dies to achieve an efficient normalization and propose for log-ratios a Gaussian random effect model, using calibration experiments to empirically identify the values of a-priori parameters in a Bayesian setting. A common problem of the above methods (that make distributional assumptions similar to ours in substance) is that they all produce gene-specific confidence intervals or significant values that need to be successively corrected for multiple comparison. Dudoit et al. (2000) takle this problem in detail. Given the high number of genes this results in considerable re-scaling of the results and possible loss of power. We believe that a successful statistical model for microarray should take into account from the beginning the high dimensionality of the problem and possibly take advantage of it, rather than hopelessly fight it (see Donoho, 2000; and van der Laan and Bryan (2000) for a different exploitation of high-dimensionality in microarry data analysis). The FDR-thresholding method we described leads in this direction. The procedure as outlined in section 3 would be optimal if the log-ratio of expression of different genes were normal with known variance and independent from each other. It is easy to argue that this is not the case and that at least one of these assumptions fails. We heuristically adapt the procedure to account for unknown variance and non-normality of the distributions. We maintain the assumption that expression measurements are independent across genes. This is not to say that we do not recognize that group of genes are co-regulated, but that we assume that the noise level (not the signal) is independent across genes. This assumption may be questioned, given that there are experimental factors that affect groups of genes (printing tip, spatial location in the array, non-specific binding, etc.); however, given the pre-processing of the data that includes normalization and pairing of signals at the same spot, it does not appear as an unreasonable working assumption. Here is a description of our procedure. We have two models under consideration:

1. $y_{ik} \sim \mathcal{N}(\theta_i, \sigma_i)$

2. $y_{ik} \sim \Phi(\theta_i)$, with θ_i location parameter.

We are also going to assume that the randomness associated with a calibration experiment is equal to the one associated with comparison experiments (in terms on the second model, $x_{ik} \sim \Phi(0)$). For this to be the case, the cDNA labelled with 2 different dyes in the calibration has to be processed separately from the beginning of the experiment, so that the distance between the two samples is similar to the distance between the samples hybridized in a comparison experiment. A simplified version of model 1. assumes constant variance, which, as we have seen, can be estimated from calibration data efficiently, using the median absolute deviation. To estimate different variances, one needs to have a number of repetitions of the experiment. Given that these are generally small, one may want to use an empirical-Bayes type of estimator that weights the sample variances for each y_i and the variance estimated from the calibration experiment. It is also possible to make assumptions on the dependence of variance of y_i on the size the the signals from which it is obtained. Once an estimate of the variance $\hat{\sigma}_i$ has been obtained, one can standardize the data obtaining $y'_i = \bar{y}_{i\cdot}/(\hat{\sigma}_i/\sqrt{K_i})$, where K_i is the number of replicates for gene i. Such data should then be thresholded using the quantiles of a Gaussian distribution, as described in section 3.

If the Gaussian does not satisfactory approximate the data, then revert to model 2, paying the price that all genes are now required to have the same distribution, except for a location parameter. Empirical quantiles of the calibration experiment provide an estimate of the quantiles of this nonparametric distribution that can be used for adaptive thresholding. As the variance of these quantile estimates can be quite high, the resulting estimator of θ is not guaranteed to control FDR at the desired level. Let $t_k = F(qk/n)$ the upper quantiles of the distribution of $|x|_i$, with q the desired FDR. Let k_{FDR} be the first k for which $|y|_{(k)} < t_k$, then the proposed estimator has the form

$$\hat{\theta_k}^{FDR} = \begin{cases} y_k & |y_k| > t_{k_{FDR}} \\ 0 & \text{otherwise} \end{cases}$$

In case there are repetitions, one wants to threshold some summary of $(y_{i1}, \ldots, y_{iK_i})$ using the appropriate reference distribution (of the same statistics) derived from the calibration experiments. As there are various possible adaptations of the procedure, we prefer to leave a detailed description to the discussion of our example, confident that this will illustrate not only the thresholding rule, but also the necessary pre-processing of the data and model checking.

An example of data analysis

The data analyzed in this section was produced using slides containing a representative subset of the mouse genome (estimated to contain a total of 30-100,000 genes). Each slide had 9504 spots, organized in 8 quadrants, each of 33 rows and 36 column. In each quadrant, the 72 spots in the upper left corner of the first 4 rows were left empty and immediately at their right, some control spots were printed: 4 spots contain buffer, 8 contain glyceraldehyderphosphate dehydrogenase (GAPHD), a housekeeping gene whose expression level is supposed to be constant in all cell types, 4 spots contain Mouse Cot1 DNA, a low complexity DNA fragment, 4 spots Salmon Sperm DNA and 4 spots poly(dA) (for further detail refer to the UCLA Microarray Core website `http://www.medsch.ucla.edu/som/humgen/nelsonlab/Mouse&human.htm`). These last three spots serve as a control for non-specific binding: the first two, being a very low complexity form of DNA, are highly likely to generate non specific binding, while the last generates non specific binding because a high fraction of genes has a poly-A tail. Figure 4 illustrates the structure of the array.

The experiments included 3 comparisons and 3 calibrations. In comparison experiments, mRNA from neuronal stem cell colonies or neurospheres (NS) (reference cell lines) is hybridized against mRNA of differentiated cells (DC), descending from the same initial population, so that there is no genetic subject difference. From the original pool of neurospheres 2 flasks were isolated and grown separately to be reference samples in the array experiment: NS7 and NS8. From the same original population 3 flasks were isolated and let differentiate for 2 weeks and were used as study sample. Table 1 describes the design of the experiment. Pictures that follow refer to the red signal in the first comparison experiment. It can be noticed that two of the calibration experiments involve the hybridization of two separately grown cell lines NS7 and NS8, with a dye reversal. A primary goal of the experiment was to identify genes that are preferentially expressed in neurospheres vs differentiated cells and vice-versa. A secondary goal was to investigate the feasibility of large array comparison using an enzymatic signal amplification kit that is commercially available and increases the amount of dye attached to any cDNA molecule, allowing to conduct experiments with smaller initial amounts of mRNA. Because of this experimental procedure, we expected a higher noise level and a significant dye effect, compared to direct labeling.

Images were scanned using Imagene 4.1 analysis software (Biodiscovery). Before turning to

Array number	Red Dye	Green Dye	Log-ratio
exp1	NS7	DC4	$y_{\cdot 1}$
exp2	DC3	NS8	$y_{\cdot 2}$
exp3	DC2	NS8	$y_{\cdot 3}$
cal1	NS7	NS8	$x_{\cdot 1}$
cal2	NS8	NS7	$x_{\cdot 2}$
cal3	NS8	NS8	$x_{\cdot 3}$

Table 1: Experiment Design

the analysis of the distribution of log ratios, a considerable amount of preprocessing of the data was carried out, in the attempt to eliminate spurious observations. Inspection of the background revealed the presence of outliers: points with extremely high background that did not correspond to high signal. We set such extreme values equal to the .999-percentile of the background. We then obtained a smooth version of the background data using 4-nearest-neighbors averages. An interesting feature was noticed when comparing the smoothed background (actually it suffices to eliminate the outliers, without proceeding to the smoothing) and the signal values: there seems to be a leap between the spots whose signal value corresponds to the background and the ones that have a signal definitely bigger than the background (see figure 5). The mechanism with which slide fluorescence is read is probably responsible for this: the laser captures variations from the background only if they have a certain amplitude. This separation is also useful to statistically classify spots as having a signal value indistinguishable from background, using a simple logistic model: the resulting separating line is depicted in figure 5.

Figure 5 also shows the location of the various control spots in the signal-background plane: the empty, buffer, and poly(dA) spots were all in the signal≈background cloud, GAPDH showed high values of signal and background, as well as Mouse Cot1 and Salmon Sperm DNA. Indeed, through out the arrays, the region containing control spots has very high signal and background values. Our current hypothesis is that the amount of DNA in the control spots is actually higher than in the remaining ones, resulting in stronger hybridization and generic increase of fluorescence in the area. After eliminating the spots where the signal was not significantly higher than background and the control spots (that seemed to have been spotted with higher content of cDNA), we proceed to inspect the distribution of the log base 10 of the background corrected signal. Extreme outliers are eliminated (more than 2.5 times the interquartile range away from the hinges). We choose to eliminate only such extreme outliers as we want to make sure that we do not attribute to contamination what is true signal, that, because of the sparsity assumption, is also going to appear in the form of outliers. The signal measurements coming from the two fluorescence sources are then normalized using a non-linear rank-based normalization, that borrows from suggestions of Dudoit et al. (2000) (we use a smoothing spline on the $(\log(g)+\log(r))/2, \log(g)-\log(r)$ space) and from Tseng et al. (2001) (for the normalization of comparison experiments, we use using only those genes whose rank position does not vary for more than 5 places between red and green signal).

Log-ratios were then obtained and inspection of their distributions carried out. We started analyzing results of the calibration experiment NS8-vs-NS8: their distribution appears reasonably symmetric, but with heavier tails than the Gaussian (see Figure 6). When comparing the quantiles of the first two calibration experiments (both obtained subtracting NS7 from NS8 values), we note that they agree well, indicating that we succeeded in correcting dye effects. As NS7 and NS8, being independently grown, may have a small subject difference, we decided to use $x_{\cdot 1} - x_{\cdot 2}$, which would definitely have expected value zero. We compared the distribution of $x_{\cdot 1} - x_{\cdot 2}$ with the convolution of the the one of $x_{\cdot 3}$ and with the distribution of the difference of log-ratios $y_{\cdot 2} - y_{\cdot 3}$ from two comparison experiments: again we see a reasonable agreement, indicating that there is no major slide effect, we can pool data from different experiments, and the noise level in calibration and comparison experiments is similar. (Figure 7).

We then applied our thresholding procedures. To obtain a universal threshold based on normality, we estimated the standard deviation value from calibration experiments averaging the median absolute deviation estimates derived from $(x_{\cdot 1} - x_{\cdot 2})/\sqrt{2})$ and the one derived from $x_{\cdot 3}$. We obtained a standard deviation value of .13, which, given the number of analyzed genes, translates in a descriptive threshold of .54 for the log-ratio of one single experiment. Considering each comparison experiment, we recorded, respectively, a total of 348, 272 and 271 genes passing this threshold.

Because the distribution of log-ratio in calibration experiments did not appear to be Gaussian, we decided to use the nonparametric approach to gain a better estimate of the number of genes that exhibit differential expression. We used the mean of the three comparison experiments as summary statistics. We then derived the distribution of the convolution of three calibration experiments by convolving $x_{\cdot 1} - x_{\cdot 2}$ with $x_{\cdot 3}$ and obtained the upper quantiles of the absolute values of the corresponding averages, setting $q = .05$. The plot of ordered magnitudes from the average of the three experiments and the quantiles from the calibration distribution is given in figure 8. The threshold value resulted in 0.4, corresponding to an average fold change of 2.5. There were 309 genes whose average difference in expression passed the 0.4 cutoff, corresponding to 4% of the entire set of genes analyzed. The laboratory that conducted these array experiments had previously investigated differential gene expression between NS and DC with a combination of different techniques (Geschwind, 2001). On the arrays that we analyzed there were spotted 41 genes that these previous studies suggested to be enriched in NS. The histogram of their log-ratio values is given in figure 9: significantly, they are almost always positive and 20% (7 out of the 37 that passe our quality checks) of them is greater than 0.4, marking a sharp change from the 4% of the overall dataset.

5 Conclusions

The analysis of microarray experiments is one example of the numerous high-dimensional problems that statistics is facing. The coming century may very well be the century of data (Donoho, 2000). This certainly the case in the field of genetics. The theory of statistics was developed in a world of scarce data and in many aspects needs to adapt to this new reality of data abundance. Some of its conceptual achievements, however, are more than relevant today. We believe that the simultaneous asymptotic minimax estimation of a sparse mean vector and controlling the false discovery rate are paradigms that can be successfully applied in recovery of signal from microarray experiment. In the present work, we have set the grounds for their most straight-forward application, obtaining a “data-driven” thresholding rule that is adaptive to the noise level of the experiment, the number of genes studied and the percentage of genes whose expression truly changes in the analyzed conditions. One obvious limitation in our model is that we assume the same distribution for each gene (or the same form in the case of the Gaussian with different variances). There is no empirical evidence to support this assumption, but also little to oppose it, as the number of replicates that have been obtained for array experiments are most often very limited, so that comparison of distributions is difficult. Certainly, not assuming identical distribution across genes would be safer and more flexible. Dudoit (2000) describes how to use permutation tests in this context. The fact remains, however, that with the small number of observations from each gene that are typically available, power to detect differences in distributions and use them efficiently is quite small. Multiple issues remain to be investigated:

1. We estimated a distribution for log-ratios under the null hypothesis using only calibration experiments. Comparison experiments should be also used to avoid bias in the case the two experiments differ substantially. More robust inferential procedures to reconstruct quantiles of a given distribution should be studied.

2. We considered situations where no subject effect needs to be estimated; however, slides effect are often present and many context of analysis require the inclusion of a subject effect. The procedure we described needs modifications to be amenable to such cases.

3. We have parametrized the problem using one distinct expression value θ_i for each gene. In reality, these mean values are dependent from each other: group of genes are under the influence of the same regulating factors and their expressions vary jointly. Incorporating this information in the model will certainly lead to more efficient estimating procedures. Unfortunately, at this stage the biological information on co-regulation is very limited, so that it is not possible to take advantage of it in modeling.

4. As we estimate the common distribution Φ from the data, the procedure outlined does not necessarily control FDR. We are in the process of conducting simulation studies to assess empirically the performance of our estimator under a variety of hypothesis.

Acknowledgements

The first author thanks Emmanuel Candes for suggesting the relevance of FDR procedures in this context; additionally she thanks organizers and participants of the IPAM session on functional genomics during the fall of 2000 for their generosity in sharing ideas. We also thank Stan Nelson and Zhun Chen of the UCLA microarray core for array fabrication. Grant support from NIHM grants #MH62800 and #MH60233 (DHG) and NIA P50 AG16570 (SK and DHG) is acknowledged for microarray work.

References

Abramovich, F., Y. Benjamini, D. Donoho, and I. Johnstone, (2000) "Adapting to Unknown Sparsity by controlling the False Discovery Rate," Stanford Statistics Department, Technical Report # 2000-19

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

De Risi, J. et al. (1997) "Exploring the metabolic and genetic control of gene expression on a genomic scale" *Science* **278**:680-686

Donoho, D. (2000) "High-Dimensional Data Analysis: The Curse and Blessing of Dimensionality" Lecture delivered at the Conference "Math Challenges of the 21st Century", August 2000. Aide-Memoire available at `http://www-stat.stanford.edu/d̃onoho/Lectures/AMS2000/AMS2000.html`

Donoho, D., and I. Johnstone (1994a) "Ideal spatial adaptation by wavelet shrinkage," *Biometrika* **81**:425-55

Donoho, D., and I. Johnstone (1994b) "Minimax risk over l_p-balls for l_q-error", *Probability Theory and Related Fields* **99**:277-303.

Dudoit, S., Y. H. Yang, M. Callow, and T. Speed (2000) "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," Berkely Statistics Department, Technical Report # 578

Geschwind, D. (2000),"Mice, microarrays, and the genetic diversity of the brain", Proc. Natl. Acad. Sci. USA, **97**: 10676-10678.

Geschwind, D. et al. (2001) "A Genetic Analysis of Neural Projenitor Differentiation," *Neuron* **29**: 325-339

Kerr, M., M. Martin, and G. Churchill, (2000) "Analysis of Variance for Gene Expression Microarray Data" to appear in *Journal of Computational Biology.*

Newton, M. et *al.* (2001) “On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data” *Journal of Computational Biology* **8**: 37-52.

Schadt, E., Li, C., Su C., and W.H. Wong (2000) “Analyzing high-density oligonucleotide gene expression array data” *J. Cell Biochem.* **80**:192-202.

Tseng, G.C., M.-K.Oh, L Rohlin, J, C. Liao, and W.H. Wong (2001), ”Issues In cDNA Microarray Analysis: Quality Filtering, Channel Normalization, Models of Variations and Assessment of Gene Effects” *Nucleic Acid Research, in print*

Tukey (1977) *Exploratory Data Analysis*, Addison-Wesley.

van der Laan, M., and Bryan, J. (2000) “Gene Expression Analysis with the Parametric Bootstrap”, to appear in *Biostatistics*

Yang, Y., M. Buckley, S. Dudoit, and T. Speed (2000) “ Comparison of methods for image analysis on cDNA microarray data”, Berkeley Statistics Department, Technical report # 584.

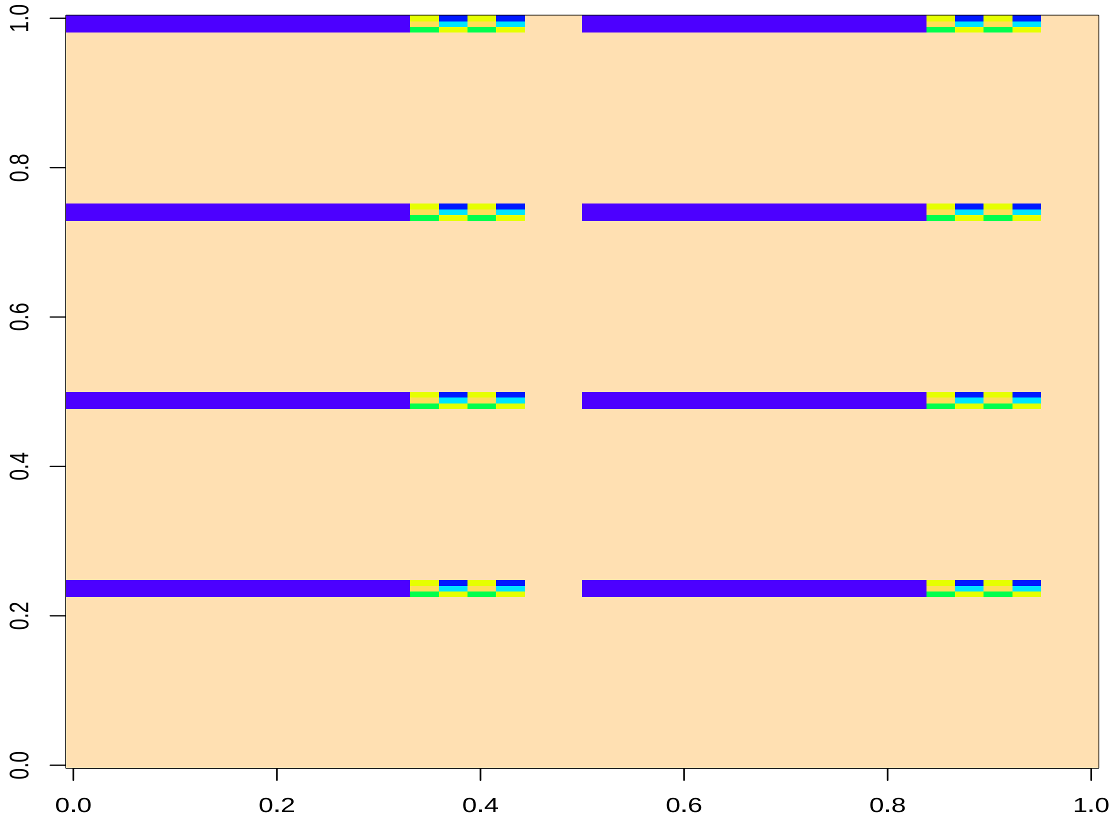

Figure 4: Schematic illustration of the division in quadrants and location of control spots on the array.

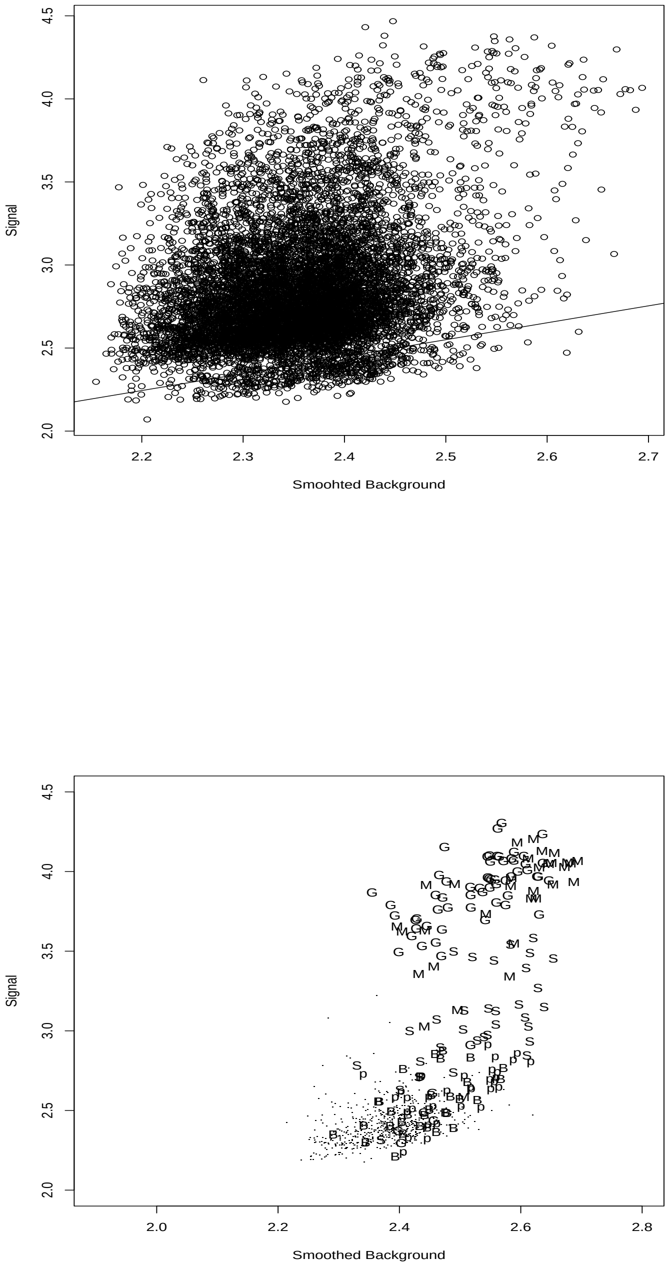

Figure 5: Signal versus background for the red signal in the first comparison experiment. On the top, scatterplot of all the points: notice the “gap” between spots with signal values corresponding to the background one and spots where signal is definitely above background. In the bottom, the position of signal from control spots is identified: G stands for GAPH, S for Salmon Sperm DNA, M for Mouse Cot1, p for Poly(da), B for Buffer, a dot for an empty spot.

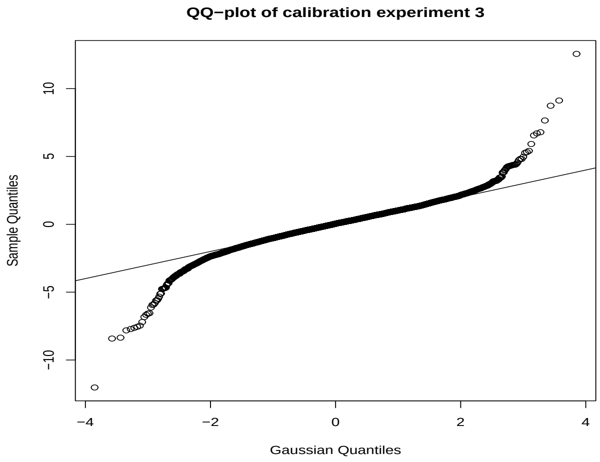

Figure 6: Quantiles of the standardized log-ratio from calibration experiment cal3 vs quantiles of the standard Gaussian distribution.

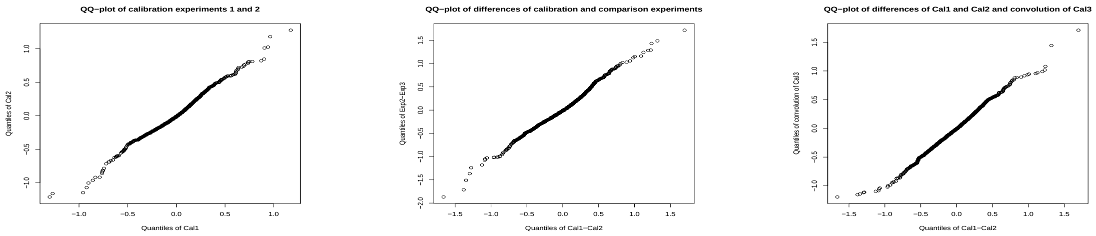

Figure 7: Comparison of quantiles of the first two calibration experiments. On the left hand side, quantiles of cal1 vs quantiles of cal2. In the center, quantiles of the difference of cal1 and cal2 versus quantiles of the difference of exp2 and exp3; on the right hand side, quantiles of the difference of cal1 and cal2 versus quantiles of the convolution of cal3.

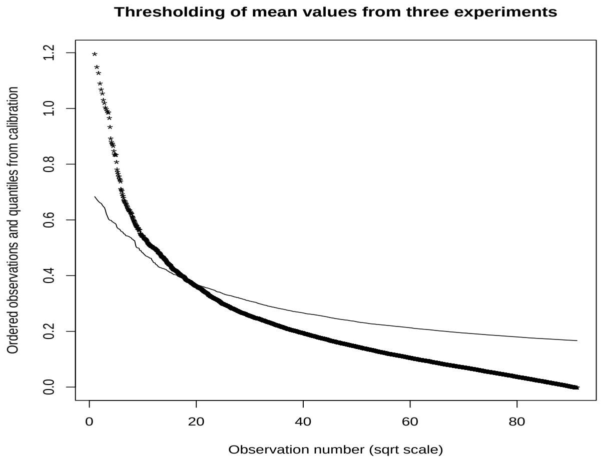

Figure 8: Plot of the ordered magnitude of the average log-ratios and quantile curve for adaptive thresholding. Note that the x-axis is the square root of the rank order.

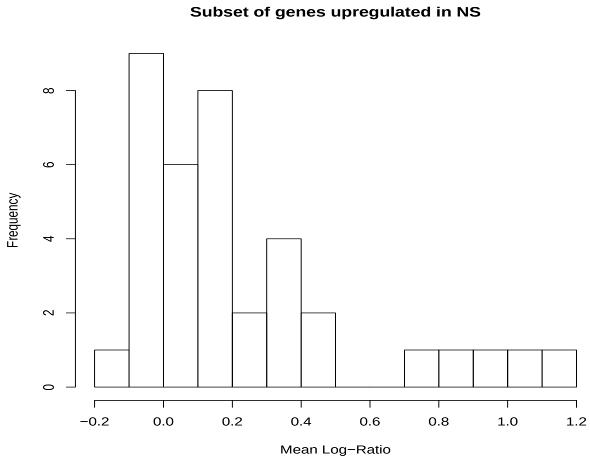

Figure 9: Histogram of average log-ratio values for the set of 37 genes suspected to be enriched in neurospheres.