

# Statistical intelligence: effective analysis of high-density microarray data

Sorin Draghici

431 State Hall, Dept. of Computer Science

Wayne State University, Detroit, MI 48202

Tel: (313) 577-5484, Fax: (313) 577-6868,

Email: sod@cs.wayne.edu

April 5, 2002

## Abstract

Microarrays allow researchers to interrogate thousands of genes simultaneously. A crucial step in data analysis is the selection of subsets of interesting genes from the initial set of genes. In many cases, especially when comparing genes expressed in a specific condition to a reference condition, the genes of interest are those which are differentially regulated. This paper focuses on the methods currently available for the selection of such genes. Fold-change, unusual ratio, univariate testing with correction for multiple experiments, ANOVA and noise sampling methods are presented and compared to each other.

**Keywords:** fold change, unusual ratio, ANOVA, Bonferroni correction, statistical significance.

## 1 Introduction

DNA microarrays are used as a very effective method to interrogate hundreds or thousands of genes simultaneously [1]. In many cases, the purpose is to compare the gene expression levels in two different specimens. In most cases, one sample is considered the reference or control and the other one is considered the experiment. Obvious examples include comparing healthy vs. disease or treated vs. untreated tissues. Sample comparison may be done using different arrays (eg. oligonucleotide arrays) or multiple channels on the same array (eg. cDNA arrays). In all such comparative studies, a very important problem is to determine those genes that are differentially expressed in the two samples compared. Although simple in principle, this problem becomes complicated in reality because the measured intensity values are affected by numerous sources of fluctuation and

noise [2, 3, 4]. For spotted cDNA arrays, there is a non-negligible probability (about 5%) that the hybridization of any single spot containing complementary DNA will not reflect the presence of the mRNA or that a single spot will provide a signal even if the mRNA is not present (about 10%) [5]).

The Affymetrix technology tries to respond to the challenge of a poor reliability for single hybridizations by representing a gene through a set of probes. The probes correspond to short oligonucleotide sequences thought to be representative for the given gene. Each oligonucleotide sequence is represented by two probes: one with the exact sequence of the chosen fragment of the gene (perfect match or PM) and one with a mismatch nucleotide in the middle of the fragment (mismatch or MM). For each gene, the value that is usually taken as representative for the expression level of the gene is the average difference between PM and MM (see Fig. 1). In principle, this value is expected to be positive because the hybridization of the PM is expected to be stronger than the hybridization of the MM. However, many factors including non-specific hybridizations and a less than optimal choice of the oligonucleotide sequences representative for the gene may determine a MM hybridization stronger than the PM hybridization for some probes. In this case, the calculated average difference may be negative. Such negative values introduce noise in the data set and make the gene selection task difficult even for Affymetrix data.

In this context, distinguishing between genes that are truly differentially regulated and genes that are simply affected by noise becomes a real challenge. All methods discussed here are completely independent of the technology used to obtain the data (e.g. cDNA or Affymetrix). The only difference between the different types of data is the pre-processing. The

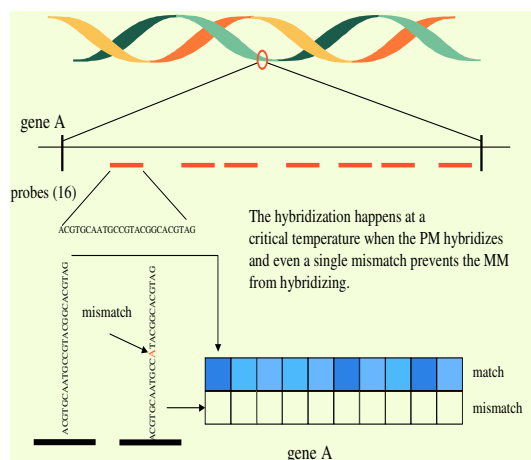


Figure 1: The principle of the Affymetrix technology. The probes correspond to short oligonucleotide sequences thought to be representative for the given gene. Each oligonucleotide sequence is represented by two probes: one with the exact sequence of the chosen fragment of the gene (perfect match or PM) and one with a mismatch nucleotide in the middle of the fragment (mismatch or MM). For each gene, the value that is usually taken as representative for the expression level of the gene is the average difference between PM and MM.

Affymetrix data will be pre-processed by combining the fluorescence levels of individual probes between match and mismatch to yield average differences and expression calls (present, absent, marginal, etc.). The cDNA data is usually processed by subtracting the background from the fluorescence values of the spots. Furthermore, when data from different arrays are compared, such comparison must be first made meaningful by bringing the arrays at comparable levels of intensity. This is usually done by some global normalization such as dividing the values on each array by their mean over the whole array. Finally, in most cases, one would like to apply a log transform in order to improve the characteristics of the distribution of expression values.

In the following, we will exemplify several methods using the very simple example of a comparison between two conditions: experiment and control.

## 1.1 Fold change

The simplest and most intuitive approach to finding the genes that are differentially regulated is to consider their fold change between control and experiment. Typically, a difference is considered as significant if it is at least 2 or 3 fold [6, 7, 8, 9, 10, 11, 12].

Sometimes, this selection method is used in parallel on expression estimates provided by several techniques such as radioactive and fluorescent labelling [13]. A convenient way to select by fold change is to calculate the ratio between the two expression levels for each gene. Such ratios can be plotted as a histogram (Fig. 2).

Typically for an experiment involving many genes, most genes will not change. Thus, the experiment/control ratio of most genes will be grouped around 1 and their logs will be grouped around 0. The horizontal axis of such a plot is graded in units reflecting the log fold change so selecting differentially regulated genes can be simply done by setting thresholds on this axis and selecting the genes outside such thresholds. For instance, in order to select genes that have a fold change of 4, one would set the thresholds at  $+/- 2$  (assuming the log was taken in base 2). If the log expression levels in the experiment are plotted against the log expression levels in control in a scatterplot (see Fig. 3), the genes selected will be at a distance of at least 2 from the diagonal that corresponds to the expression being the same in control and experiment.

The fold change method is often used because it is simple and intuitive. However, the method has important disadvantages. The most important drawback is that the fold threshold is chosen arbitrarily and may often be inappropriate. For instance, if one is selecting genes with at least 2 fold change and the condition under study does not affect any genes to the point of inducing a 2 fold change, no genes will be selected resulting in zero sensitivity. Reciprocally, if the condition is such that many genes change dramatically (or if the threshold is lowered), the method will select too many genes and will have a low specificity. In this respect, the fold change method is nothing but a blind guess. Another important disadvantage is related to the fact that the microarray technology tends to have a bad signal/noise ratio for genes with low expression levels. On a scatterplot this is illustrated by the funnel shape of the distribution (see yellow curves 2 in Fig. 3) determined by a large variance of the values measured at the low end of the scale (to the left) and a low variance of values measured at the high end of the scale (to the right). A gene that is closer to the diagonal at a high expression level might be more reliable than a gene that is a bit further from the diagonal at a low level. Since the fold change uses a constant threshold for all genes, it will introduce false positives at the low end thus reducing the specificity, while missing true positives at the high end and thus reducing the sensitivity.

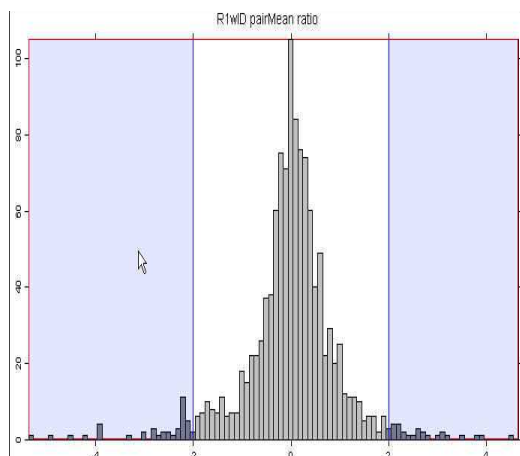


Figure 2: Experiment-control ratios can be plotted in a histogram showing the number of genes (vertical axis) for every ratio value (horizontal axis). The horizontal axis is graded in fold change units. Selecting differentially regulated genes based on fold change corresponds to setting thresholds at the desired minimum fold change and selecting the genes in the tails of the histogram (blue areas).

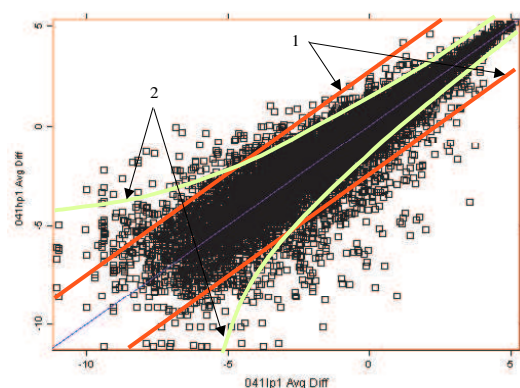


Figure 3: A scatterplot representing the experiment values plotted against the control values. Unchanged genes will appear on the diagonal as the two values are similar. Selecting genes with a minimum fold change is equivalent to setting linear boundaries parallel to the  $y=x$  diagonal at a distance from it equal to the minimum fold change requested (1). A better selection would be achieved with non-linear boundaries that adapt to the increased noise variance at low intensities (2).

## 1.2 Unusual ratio

The second widely used selection method involves selecting the genes for which the ratio of the experiment and control values is a certain distance from the mean experiment/control ratio [14]. Typically, this distance is taken to be  $\pm 2$  standard deviations. In other words, the genes selected as being differentially regulated will be those genes having an experiment/control ratio at least 2 standard deviations away from the mean experiment/control ratio. In practice, this can be achieved very simply by applying a z-transform to the log ratio values. The z-transform essentially subtracts the mean and divides by the standard deviation. In consequence, a histogram of the transformed values will still be centered around 0 (most genes will still have a ratio of 1 corresponding to a log ratio of 0) but the horizontal axis will be graded in units of standard deviation (see Fig. 4). Thus, setting thresholds at  $\pm 2$  will correspond to selecting those genes which have an unusual ratio, situated at least 2 standard deviations away from the mean ratio.

This method is superior to the fold change method while still simple and intuitive. The advantage of the unusual ratio method is that it will automatically adjust the cut-off threshold even if the number of genes regulated and the amount of regulation vary considerably. Thus, the unusual ratio method uses thresholds on how different the experiment/control ratio of a gene is with respect to the mean of all such ratios instead of thresholds on the values of the ratios themselves. No matter how many genes are regulated and no matter by how much, this method will always pick the genes that are affected most. In particular, if the ratio distribution is close to a normal distribution and the thresholds are set at  $\pm 2$  standard deviation, this method will select the 5% most regulated genes.

However, the unusual ratio method still has important intrinsic drawbacks. Thus, the method will report as differentially regulated the 5% of the genes *even if there are no differentially regulated genes*. This happens because in all microarray experiments there is a certain amount of variability due to noise. Thus, if the same experiment is performed twice, the expression values measured for any particular gene will likely not be exactly the same. If the method is applied to study differential regulated genes in two control experiments, the result will still contain about 5% of the genes. This is because different measurements for the same gene will still vary a little bit due to the noise. The method will dutifully calculate the mean and standard deviation of this distribution and will select those genes situated  $\pm 2$  standard devia-

tions away from the mean.

Furthermore, the method will still select 5% of the genes even if much more genes are in fact regulated. Thus, while the fold method uses an arbitrary threshold and can provide too many or too few genes, the unusual ratio method uses a fixed proportion threshold that will always report a given proportion of the genes as being differentially regulated. On a scatterplot (such that in Fig. 3), the ratio method continues to use cut-off boundaries parallel to the diagonal which will continue to overestimate the regulation at low intensity and underestimate it at high intensity.

A variation of the unusual ratio method selects those genes for which the absolute difference in the average expression intensities is much larger than the estimated standard error ( $\hat{\sigma}$ ) computed for each gene using array replicates. For duplicate experiments the absolute difference has to be larger than  $4.3\hat{\sigma}$  and  $22.3\hat{\sigma}$  for the 5% and 1% significance levels, respectively [15]. For triplicate experiments the requirements can be relaxed to  $2.8\hat{\sigma}$  and  $5.2\hat{\sigma}$  for the 5% and 1% significance levels, respectively. A number of other *ad hoc* thresholding and selection procedures have also been used. For instance, [16, 17] only considered genes for which the difference between the duplicate measurements did not exceed half their average. Furthermore, the genes considered as differentially regulated were those genes which exhibited at least a 2-fold change in expression. Although this seems to use the fold method, it can be shown [15] that the combination of the duplicate consistency condition and the differentially regulated condition can be expressed in terms of mean and standard deviations and therefore it falls under the scope of the unusual ratio method.

### 1.3 Confidence levels and corrections for multiple experiments

Another possible approach to gene selection is to use univariate statistical tests (e.g. t-test) to select differentially expressed genes [18, 15, 19]. Let us consider that the log ratios follow a distribution like the one in Fig. 5. For a given threshold and a given distribution the confidence level or p-value is the probability of the measured value being in the shaded area by chance. The thinking is that a gene whose log ratio falls in the shaded area is far from the mean log ratio and will be called differentially regulated (upregulated in this case). However, the measured log ratio may be there just due to random factors such as noise. The probability of the measurement being there just by chance is the p-value. In this case, calling the gene differentially regulated will be a mistake and the p-value is

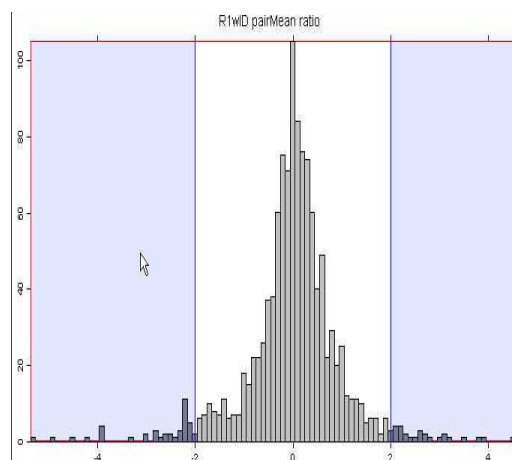


Figure 4: Selecting by the unusual ratio criterion. The frequencies (vertical axis) of the z-scores of the genes (horizontal axis) are plotted in a histogram. The horizontal axis is now graded in standard deviations. Setting thresholds at the  $\pm 2$  marks corresponds to selecting genes with a ratio more than 2 standard deviations away from the mean log ratio.

the probability of making this mistake.

Regardless of the particular test used (e.g. t-test if normality is assumed), one needs to consider the fact that when many genes are analyzed at one time, some genes will appear as being significantly different just by chance [20, 21, 22, 23, 24]. Let us assume we are considering a gene with a value (e.g. log-ratio)  $v$  situated in the tail of the histogram of all such values, possibly indicating that the gene is regulated. The  $p$  value provided by the univariate test is the probability that  $v$  is where it is just by chance. If we call this gene differentially regulated based on this value and the value is there by chance, we will be making a mistake. Therefore,  $p$  is the probability of making a mistake in this test<sup>1</sup>. The probability of drawing the right conclusion in this one test will be  $1 - p$ . If there are  $R$  such tests, we would like to draw the right conclusion from *all* of them. The probability of this will be  $prob(right) = (1 - p)^R$ . The probability of making a mistake will be  $prob(wrong) = 1 - prob(right) = 1 - (1 - p)^R$ . This is the so-called Sidák correction [25]. Bonferroni [26, 27] noted that for small  $p$ ,  $1 - (1 - p)^R \approx Rp$  and proposed to correct the required p-value to  $\hat{p} = p/R$ . Both Bonferroni and Sidák corrections are unsuitable for gene expression analysis because for large number of genes  $R$ , no gene will be below the corrected  $p$  value (e.g.

<sup>1</sup>From a statistical point of view, interrogating  $R$  genes at the same time, as on a microarray, is equivalent to running  $R$  parallel tests.

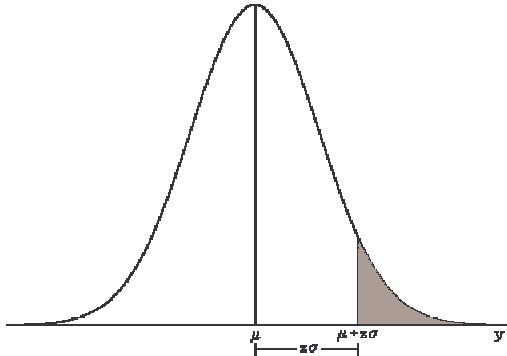


Figure 5: For a given threshold and a given distribution the p-value is the probability of the measured value being in the shaded area by chance. If the method is applied to the distribution of the log ratios, the p-value is the probability of making a mistake when calling a gene differentially regulated if its measured log ratio is in the shaded area.

$\tilde{p} = p/R$  for Bonferroni).

A family of methods that allow less conservative adjustments of the  $p$ -values without the heavy computation involved in resampling is the Holm step-down group of methods [20, 22, 28, 21]. These methods order the genes in increasing order of their  $p$ -value and make successive smaller adjustments.

Bonferroni, Sidák and Holm’s step-down adjustment assume the variables are independent. This is not true for expression data since genes influence each other in complex interactions [29, 30]. The Westfall and Young (W-Y) step-down is a more general method which adjusts the  $p$ -value while taking into consideration the possible correlations. Duplication, together with a univariate testing procedure (e.g.  $t$ -test or Wilcoxon) followed by a W-Y adjustment for multiple testing [24] are proposed in [19]. Another technique that considers the correlation is the bootstrap method [31, 32, 24]. The method samples with replacement the pool of observations to create new data sets and calculates  $p$ -values for all tests. For each data set, the minimum  $p$ -value on the resampled data sets is compared with the  $p$ -value on the original test. The adjusted  $p$ -value will be the proportion of resampled data where the minimum pseudo- $p$ -value is less than or equal to an actual  $p$ -value. Bootstrap used with sampling without replacement is known as the permutation method [33, 34]. Both bootstrap and permutation are computationally intensive.

## 1.4 ANOVA

A particularly interesting approach to microarray data analysis and selecting differentially regulated genes is the ANalysis Of VAriance (ANOVA) [35, 36, 37]. The idea behind ANOVA is to build an explicit model about the sources of variance that affect the measurements and use the data to estimate the variance of each individual variable in the model.

For instance, Kerr and Churchill [38, 39, 40] proposed the following model to account for the multiple sources of variation in a microarray experiment:

$$\log(y_{ijk g}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ijk g} \quad (1)$$

In this model,  $\mu$  is the overall mean signal of the array,  $A_i$  is the effect of the  $i^{th}$  array,  $D_j$  represents the effect of the  $j^{th}$  dye,  $V_k$  is the effect of the  $k^{th}$  variety<sup>2</sup>,  $G_g$  is the variation of the  $g^{th}$  gene,  $(AG)_{ig}$  is the effect of a particular spot on a given array,  $(VG)_{kg}$  represents the interaction between the  $k^{th}$  variety and the  $g^{th}$  gene and  $\epsilon_{ijk g}$  represents the error term for array  $i$ , dye  $j$ , variety  $k$  and gene  $g$ . The error is assumed to be independent and of zero mean. Finally,  $\log(y_{ijk g})$  is the measured log-ratio for gene  $g$  of variety  $j$  measured on array  $i$  using dye  $j$ .

The advantage of ANOVA is that each source of variance is accounted for. Because of this, it is easy to distinguish between interesting variation such as gene regulation and side effects such as differences due to different dyes or arrays. The caveat is that ANOVA requires a very careful experiment design [39, 41] that must ensure a sufficient number of degrees of freedom. Thus, ANOVA cannot be used if the experiments have not been designed and executed in a manner consistent with the ANOVA model used.

## 1.5 Noise sampling

A full blown ANOVA requires a design that blocks all variables under control and randomizes the others. In most cases, this requires repeating several microarrays with various mRNA samples and swapping dyes if a multi-channel technology is used. A particular variation on the ANOVA idea can be used to identify differentially regulated genes using spot replicates on single chips to estimate the noise and calculate confidence levels for gene regulation [2, 42, 43]. The Kerr-Churchill model is modified as follows:

$$\log R(gs) = \mu + G(g) + \epsilon(g, s) \quad (2)$$

<sup>2</sup>In this context a variety is a condition such as healthy or disease.

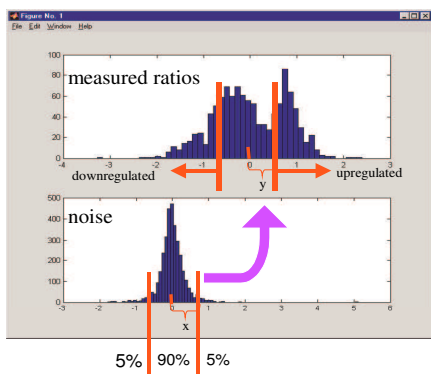


Figure 6: Using the empirical noise distribution to find differentially expressed genes. The confidence intervals found on the noise distribution (lower panel) can be mapped onto confidence intervals on the distribution of the expression values (upper panel). The confidence interval can be different between the up and downregulated parts of the distribution if the noise is skewed.

where  $\log R(gs)$  is the measured log ratio for gene  $g$  and spot  $s$ ,  $\mu$  is the average log ratio over the whole array,  $G(g)$  is a term for the differential regulation of gene  $g$  and  $\epsilon(g, s)$  is a zero-mean noise term.

In the model above, one can calculate an estimate  $\hat{\mu}$  of the average log ratio  $\mu$ :

$$\hat{\mu} = \frac{1}{n \cdot m} \sum_{g,s} \log(R(g, s)) \quad (3)$$

which is the sum of the log ratios for all genes and all spots divided by the total number of spots ( $m$  replicates and  $n$  genes). An estimate  $\widehat{G}(g)$  of the effect of gene  $g$  can also be calculated as:

$$\widehat{G}(g) = \frac{1}{m} \sum_s \log(R(g, s)) - \hat{\mu} \quad (4)$$

where the first term is the average log ratio over the spots corresponding to the given gene. Using the estimates above, one can now calculate an estimate of the noise as follows:

$$\widehat{\epsilon}(g, s) = \log(R(g, s)) - \hat{\mu} - \widehat{G}(g) \quad (5)$$

This will provide a noise sample for each spot. The samples collected from all spots yield an empirical noise distribution<sup>3</sup>. A given confidence level can be

<sup>3</sup>Note that no particular shape (such as Gaussian) is assumed for the noise distribution or for the distribution of the gene expression values, which makes this approach very general.

associated with a deviation from the mean of this distribution. To avoid using any particular model, the distance from the mean can be calculated by numerically integrating the area under the distribution. This distance on the noise distribution can be put into correspondence to a distance  $y$  on the measured distribution (Fig. 6) by bootstrapping [42, 24]. Furthermore, the dependency between intensity and variance can be taken into account by constructing several such models covering the entire intensity range and constructing non-linear confidence boundaries similar to those in Fig. 3. The method has the important advantage that its non-linear selection boundaries adapt automatically both to various amounts of regulation and different amounts on noise for a given confidence level chosen by the user.

A full blown ANOVA requires a special experimental design but provide error estimates for all variables considered in the model. The noise sampling method does not require such a special experiment design but only provides estimates for the log ratios of the genes. It has been shown that the noise sampling method provides a better sensitivity than the unusual ratio method and a much better sensitivity than the fold change method [42, 43].

## 1.6 Other methods

Other statistically based methods for the selection of differentially regulated genes include model based maximum likelihood estimation approaches [44, 45, 5]. A maximum likelihood estimation approach for two color arrays is described in [44]. This approach is based on the hypothesis that the level of a transcript depends on the concentration of the factors driving its selection and that the variation for any particular transcript is normally distributed and in a constant proportion relative to most other transcripts. This hypothesis is then exploited by considering a constant coefficient of variation  $c$  for the entire gene set and constructing a 3-rd degree polynomial approximation of the confidence interval as a function of the coefficient of variation  $c$ . This approach is also interesting because it provides the means to deal with signals uncalibrated between the two colors through an iterative algorithm that compensates for the color difference. Sapir et al. [45] present a robust algorithm for estimating the posterior probability of differential expression based on an orthogonal linear regression of the signals obtained from the two channels. The residuals from the regression are modeled as a mixture of a common component and component due to differential expression. An expectation maximization algorithm is used to deconvolve

the mixture and provide estimate of the probability that each gene is differentially regulated as well as estimates of the error variance and proportion of differentially expressed genes.

Another approach uses replicates and a maximum likelihood approach to calculate the probability of a particular gene being expressed and selecting only those genes for which all replicates indicate that the gene is expressed [5]. Note that although in [5] this approach is used only to make the binary distinction between expressed and not expressed genes, the approach can be extended to a multichannel experiment to detect differentially expressed genes.

Two hierarchical models (Gamma-Gamma and Gamma-Gamma-Bernoulli) for the two channel (color) intensities are proposed in [46]. One advantage of such an approach is that the models constructed take into consideration the variation of the posterior probability of change on the absolute intensity level at which the gene is expressed. This particular dependency is also considered in [47] where the values measured on the two channels are assumed to be normally distributed with a variance depending on the mean. Such intensity dependency reduces to defining some curves in the green-red plane corresponding to the two channels and selecting as differentially regulated the genes that fall outside the equiconfidence curves.

However, as pointed out in [19] any gene selection method that does not use replication is critically sensitive to the typically large amount of noise in microarrays. Dudoit et al. [19] propose a univariate testing procedure (e.g t-test or Wilcoxon) followed by a Westfall and Young adjustment for multiple testing [24]. Another multiple experiment approach is to identify the differentially expressed genes by comparing their behavior in series of experiments with an expected expression profile [48, 49]. The genes can be ranked according to their degree of similarity to a given expression profile. The number of false positives can be controlled through random permutations that allow the computation of suitable cut-off thresholds for the degree of similarity. Clearly, these approaches can only be used in the context of a large data set including several microarrays for each condition considered.

Other methods used for the selection of differentially regulated genes include gene shaving [50], assigning gene confidence [51] or significance [52], bootstrap [31, 32] and Bayesian approaches [53, 54, 55]. A few methods also take into consideration that the variance depends on the intensity [46] and [47].

Finally, more elaborate methods for the data analysis of gene expression data exist. Such methods in-

clude clustering [56, 57, 15, 58, 59, 60, 61, 62, 63, 64, 65, 12, 66, 67]), principal component analysis [58, 68, 69], singular value decomposition [70], independent component analysis [71], gene shaving [50] and many others. The goals of such methods go well beyond the selection of differentially regulated genes and, as such, they are outside the scope of the present paper.

## 2 Conclusions

A plethora of refined methods is available for the analysis of microarray data and in particular for the selection of differentially regulated genes. Although still in widespread use, the early methods of selection by fold change and unusual ratio are clearly inadequate. Using a fold change without a clear biological justification is just a blind guess. The unusual ratio method will always report some genes as regulated even if two identical tissues are studied (false positives). These two methods suffer from severe drawbacks and their use should be discontinued as methods for selecting differentially regulated genes. However, studying the fold change of genes of known function is and will continue to remain important. In order words, computing statistics as required by biological reasons is fully justified (eg. how do apoptosis related genes change in immortalization?). However, drawing biological conclusions based on an arbitrary choice of fold change is not (e.g. concluding that gene X is relevant to immortalization because it has a fold change of 2).

When using univariate statistical tests for hundreds or thousands of genes (eg. with data coming from most commercial chips), Bonferroni should be taken as a sufficient but not necessary condition. In other words, if a gene still appears to be differentially regulated after applying the Bonferroni correction, the gene is indeed so. However, if a gene does not appear as differentially regulated after the Bonferroni correction, the gene may or may not be so. Univariate tests such as the t-test followed by a Bonferroni correction can be used effectively if the number of genes on the array is relatively low (tens to a few hundreds). At this time, the bootstrap and Westfall and Young families of methods appear to provide the most accurate correction for multiple experiments.

Current statistical methods offer a great deal of control and the possibility of selecting genes within a given confidence interval. However, all such methods rely essentially on a careful experiment design and the presence of replicate measurements. A good way to obtain reliable results is arguably some version of

the ANOVA method. However, in most cases, this will probably mean involving a statistician from the very beginning and designing the experiment in such a way that enough degrees of freedom are available in order to answer the relevant biological questions. The noise sampling method is a variation of ANOVA and allows the automatic computation of noise dependent equiconfidence boundaries. The noise sampling method can be used in many instances in which data for a full blown ANOVA is not available.

## References

- [1] Mark Schena. *Microarray Biochip Technology*. Eaton Publishing, 2000.
- [2] Sorin Draghici, Alexander Kuklin, Bruce Hoff, and Soheil Shams. Experimental design, analysis of variance and slide quality assessment in gene expression arrays. *Current Opinion in Drug Discovery and Development*, 4(3):332–337, 2001.
- [3] Johannes Schuchhardt, Dieter Beule, Eryc Wol-ski, and Holger Eickhoff. Normalization strategies for cDNA microarrays. *Nucleic Acids Research*, 28(10):e47i–e47v, 2000.
- [4] S. E. Wildsmith, G. E. Archer, A. J. Winkley, P. W. Lane, and P. J. Bugelski. Maximizing of signal derived from cDNA microarrays. *BioTechniques*, 30:202–208, 2000.
- [5] Mei-Ling Ting Lee, Frank C. Kuo, G. A. Whitmore, and Jeffrey Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci.*, 97(18):9834–9839, 2000.
- [6] Peter G. Schultz Cecilia H. Jiang, Joe Z. Tsien and Yinghe Hu. The effects of aging on gene expression in the hypothalamus and cortex of mice. *PNAS*, 98(4):1930–1934, February 13,2001.
- [7] J. L. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, and J.M. Trent. User of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14(4):457–460, 1996.
- [8] Joseph L. DeRisi, Vishwanath R. Iyer, and Patrick O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.
- [9] J. J. M. ter Linde, H. Liang, R. W. Davis, H. Y. Steensma, J. P. Van Dijken, and J. T. Pronk. Genome-wide transcriptional analysis of aerobic and anaerobic chemostat cultures of *Saccharomyces cerevisiae*. *Journal of Bacteriology*, 181(24):7409–7413, 1999.
- [10] Priya Sudarsanam, Vishwanath R. Iyer, Patrick O. Brown, and Fred Winston. Whole-genome expression analysis of snf/swi mutants of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.*, 97(7):3364–3369, 2000.
- [11] Axel Wellmann, Catherine Thieblemont, Stefania Pittaluga, Akira Sakai, Elaine S. Jaffe, Paul Seibert, and Mark Raffeld. Detection of differentially expressed genes in lymphomas using cDNA arrays: identification of *clusterin* as a new diagnostic marker for anaplastic large-cell lymphomas. *Blood*, 96(2):398–404, 2000.
- [12] Kevin P. White, Scott A. Rifkin, Patrick Hurban, and David S. Hogness. Microarray analysis of *Drosophila* development during metamorphosis. *Science*, 286:2179–2184, 1999.
- [13] Craig S. Richmond, Jeremy D. Glasner, Robert Mau, Hongfan Jin, and Frederick R. Blattner. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Research*, 27(19):3821–3835, 1999.
- [14] Han Tao, Christoph Bausch, Craig Richmond, Frederick R. Blattner, and Tyrrell Conway. Functional genomics: Expression analysis of *Escherichia coli* growing on minimal and rich media. *Journal of Bacteriology*, 181(20):6425–6440, 1999.
- [15] Jean-Michael Claverie. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics*, 8(10):1821–1832, 1999.
- [16] M. Schena, D. Shalon, R. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.
- [17] M. Schena, D. Shalon, R. Heller, A. Chai, P. Brown, and R. Davis. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. National Academy of Science USA*, 93:10614–10519, 1996.
- [18] S. Audic and Jean-Michael Claverie. Vizualizing the competitive recognition of TATA-boxes

- in vertebrate promoters. *Trends in Genetics*, 14:10–11, 1998.
- [19] S. Dudoit, Y. H. Yang, M. Callow, and T. Speed. Statistical models for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, University of California, Berkeley, 2000.
- [20] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.
- [21] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley and Sons, Inc., New York, 1987.
- [22] J. P. Shaffer. Modified sequentially rejective multiple test procedures. *Journal of American Statistical Association*, 81:826–831, 1986.
- [23] J. P. Shaffer. Multiple hypothesis testing. *Annual Reviews in Psychology*, 46:561–584, 1995.
- [24] P. H. Westfall and S. S. Young. *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley, New York, 1993.
- [25] Z. Sidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62:626–633, 1967.
- [26] C. E. Bonferroni. *Il calcolo delle assicurazioni su gruppi di teste*, chapter Studi in Onore del Professore Salvatore Ortu Carboni, pages 13–60. Rome, 1935.
- [27] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
- [28] B. Holland and M. D. Copenhaver. An improved sequentially rejective Bonferroni test procedure. *Biometrika*, 43:417–423, 1987.
- [29] Patrick D’haeseller. *Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering*. PhD Thesis, University of New Mexico, 2000.
- [30] Patrick D’haeseller, S. Liang, and R. Somogyi. Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics*, 8(16):707–726, 2000.
- [31] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791, 1985.
- [32] M. Kathleen Kerr and Gary A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. [www.jax.org/research/churchill/pubs/index.html](http://www.jax.org/research/churchill/pubs/index.html), 2001.
- [33] C. C. Brown and T. R. Fears. Exact significance levels for multiple binomial testing with application to carcinogenicity screens. *Biometrics*, 37:763–774, 1981.
- [34] J. Heyse and D. Rom. Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biometrical Journal*, 30:883–896, 1988.
- [35] A. Aharoni, L. C. P. Keizer, H. J. Bouwneester, Z. Sun, M. Alvarez-Huerta, H. A. Verhoeven, J. Blaas, A.M.M.L. van Houwelingen, R. C.H. De Vos, H. van der Voet, R. C. Jansen, M. Guis, J. Mos, R. W. Davis, M. Schena, A. J. van Tunen, and A. P. O’Connell. Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. *The Plant Cell*, 12:647–661, May 1975.
- [36] Alvis Brazma and Jaak Vilo. Gene expression data analysis. *Federation of European Biochemical Societies Letters*, 480(23893):17–24, 2000.
- [37] A. A. Hill, C. P. Hunter, B. T. Tsung, G. Tucker-Kellogg, and E. L. Brown. Genomic analysis of gene expression in *C. elegans*. *Science*, 290:809–812, 2000.
- [38] M. Kathleen Kerr, Mitchell Martin, and Gary A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837, 2000.
- [39] M. Kathleen Kerr and Gary A. Churchill. Statistical design and the analysis of gene expression. *Genetical Research*, [www.jax.org/research/churchill/pubs/index.html](http://www.jax.org/research/churchill/pubs/index.html), 77:123–128, 2001.
- [40] M. Kathleen Kerr and Gary A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, [www.jax.org/research/churchill/pubs/index.html](http://www.jax.org/research/churchill/pubs/index.html), 2001.

- [41] M. Kathleen Kerr and Gary A. Churchill. Experimental design for gene expression analysis. *Biostatistics*, [www.jax.org/research/churchill/pubs/index.html](http://www.jax.org/research/churchill/pubs/index.html), (2):183–201, 2001.
- [42] Sorin Draghici, Olga Kulaeva, Anton Petrov, Bruce Hoff, Alexander Kuklin, Soheil Shams, and Michael Tainsky. Computational methods for the selection of differentially regulated genes in cell immortalization. *Submitted to Journal of Computational Biology*, 2002.
- [43] Dai Wang, Sorin Draghici, Anton Petrov, Bruce Hoff, Alexander Kuklin, and Soheil Shams. Methods for selecting differentially regulated genes in microarrays: noise sampling vs. standard deviations. *Submitted to Bioinformatics*, 2001.
- [44] Yidong Chen, Edward R. Dougherty, and Michael L. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2(4):364–374, 1997.
- [45] Marina Sapir and Gary A. Churchill. Estimating the posterior probability of differential gene expression from microarray data. Technical Report <http://www.jax.org/research/churchill/pubs/>, Jackson Labs, Bar Harbor, ME, 2000.
- [46] M.A. Newton, C.M. Kendzierski, C.S. Richmond, F. R. Blattner, and K.W. Tsui. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Technical report, University of Wisconsin, <http://www.biostat.wisc.edu/geda/eba.html>, 1999.
- [47] Christopher J. Roberts, Bryce Nelson, Mathew J. Marton, Roland Stoughton, Michael R. Meyer, Holly A. Bennett, Yudong D. He, Hongyue Dia, Wynn L. Walker, Timothy R. Hughes, Mike Tyers, Charles Boone, and Stephen H. Friend. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, 287:873–880, 2000.
- [48] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.
- [49] T. Galitski, A. J. Saldanha, C. A. Styles, E. S. Lander, and G. R. Fink. Ploidy regulation of gene expression. *Science*, 285:251–254, 1999.
- [50] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P. Brown. ‘Gene shaving’ as a method for indentifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2):1–21, 2000.
- [51] E. Manduchi, G. R. Grant, S. E. McKenzie, G. C. Overton, S. Surrey, and C. J. Stoeckert. Generation of patterns from gene expression data by assigninig confidence to differentially expressed genes. *Bioinformatics*, 16(8):685–698, 2000.
- [52] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.*, 98(9):5116–5121, 2001.
- [53] P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.
- [54] A.D. Long, H.J. Mangalam, B.Y.P. Chan, L. Toller, G. W. Hatfield, and P. Baldi. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *J. Biol. Chem.*, 276(23):19937–19944, 2001.
- [55] M. West, J.R. Nevins, J.R. Marks, R. Spang, C.A.B. Blanchette, and H. Zuzan. Bayesian regression analysis in the “large p, small n” paradigm with application in DNA microarray studies. Technical report, Duke University, 2000.
- [56] John Aach, Wayne Rindone, and George M. Church. Systematic management and analysis of yeast gene expression data. *Genome Research*, 10:431–445, 2000. <http://arep.med.harvard.edu/ExpressDB>.
- [57] Alvis Brazma. Mining the yeast genome expression and sequence data. *The BioInformer*, (4), 1998. [http://bioinformer.ebi.ac.uk/newsletter/archives/4/lead\\_article.html](http://bioinformer.ebi.ac.uk/newsletter/archives/4/lead_article.html).

- [58] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Sciences*, volume 95, pages 14863–14868, 1998.
- [59] Rob M. Ewing, Alia Ben Kahla, Oliver Poirot, Fabrice Lopez, Stephane Audic, and Jean-Michel Claverie. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Research*, 9:950–959, 1999.
- [60] Laurie J. Heyer, Semyon Kruglyak, and Shibu Yooseph. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, 9:1106–1115, 1999.
- [61] Genevieve Pietu, Regine Mariage-Samson, Nicole-Adeline Fayein, Christiane Matingou, Eric Eveno, et al. The genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. *Genome Research*, 9:195–209, 1999.
- [62] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. In *Proc. Natl. Acad. Sci.*, volume 96, pages 2907–2912, 1999.
- [63] Sophia Tsoka and Christos A. Ouzounis. Recent developments and future directions in computational genomics. *Federation of European Biochemical Societies Letters*, (23897):1–7, 2000.
- [64] Jacques van Helden, Alma F. Rios, and Julio Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28(8):1808–1818, 2000.
- [65] Ming Li Wang, Stephen Belmonte, Ulandt Kim, Maureen Dolan, John W. Morris, and Howard M. Goodman. A cluster of ABA-regulated genes on *Arabidopsis Thaliana* BAC T07M07. *Genome Research*, 9:325–333, 1999.
- [66] Michael Q. Zhang. Large-scaled gene expression data analysis: A new challenge to computational biologists. *Genome Research*, 9:681–688, 1999.
- [67] J. Zhu and M.Q. Zhang. Cluster, function and promoter: Analysis of yeast expression array. In *Pacific Symposium on Biocomputing*, pages 476–487, 2000.
- [68] S.G. Hilsenbeck, W.E. Friedrichs, R. Schiff, P. O’Connell, R.K. Hansen, C.K. Osborne, and S.A. W. Fuqua. Statistical analysis of array expression data as applied to the problem of Tamoxifen resistance. *Journal of the National Cancer Institute*, 91(5):453–459, 1999.
- [69] S. Raychaudhuri, J. M. Stuart, and R.B. Altman. Principal components analysis to summarize microarray experiments: Application to sporulation time series. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 5, pages 452–463, 2000.
- [70] O. Alter, P.O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.*, 97(18):10101–10106, 2000.
- [71] W. Liebermeister. Independent component analysis of gene expression data. In *Proc. of German Conference on Bioinformatics GCB’01*, 2001. <http://www.bioinfo.de/isb/gcb01/poster/index.html>.