Statistical intelligence: effective analysis of high-density microarray data

Sorin Draghici

Microarrays enable 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 review focusses 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 reviewed and compared.

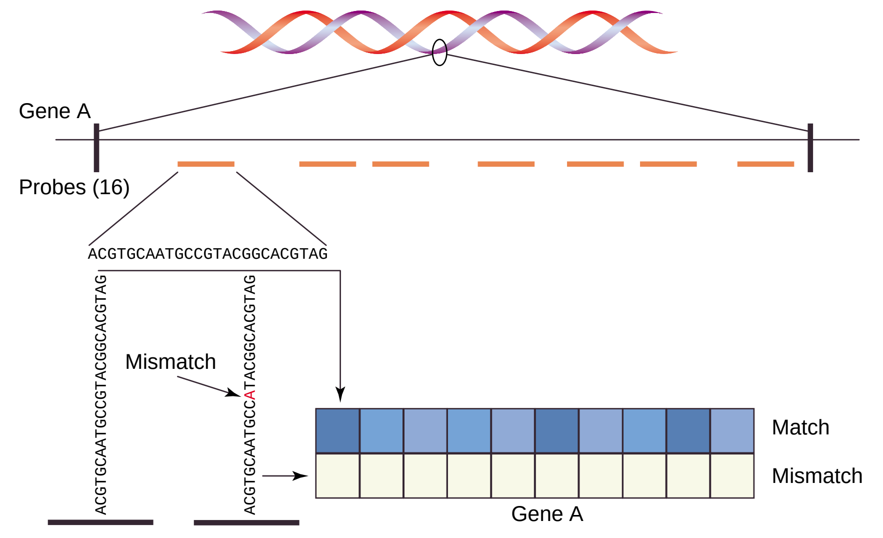
Sorin Draghici
431 State Hall
Dept of Computer Science
Wayne State University
Detroit
MI 48202, USA
tel: +1 313 577 5484
fax: +1 313 577 6868
e-mail: sod@cs.wayne.edu

▼ DNA microarrays are an effective tool for interrogating hundreds or thousands of genes simultaneously [1]. In many cases, the aim is to compare gene expression levels in two different samples. Typically, one sample is used as a reference or control, and the second is considered as the experiment. Obvious examples include comparisons between healthy and diseased tissues, or treated and untreated tissues. Sample comparisons can be done using hybridization on different arrays (e.g. oligonucleotide arrays) or by multiple color hybridizations on one array (e.g. cDNA arrays). In all such comparative studies, a crucial issue is to determine which genes are differentially expressed in the two samples compared. Although simple in principle, in reality, this problem becomes complicated because the measured intensity values are affected by numerous sources of fluctuation and noise [2–4]. For spotted cDNA arrays, there is a non-negligible probability (~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 (~10%) [5]).

The Affymetrix technology attempts to respond to this challenge of 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 (PM)] and one with a mismatch nucleotide in the middle of the fragment [mismatch (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, might result in an MM hybridization stronger than the PM hybridization for some probes. In this case, the calculated average difference might be negative. Such negative values introduce noise into the dataset 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. The methods outlined here are completely independent of the technology used to obtain the data (e.g. cDNA or Affymetrix). The only difference between the types of data is the pre-processing. Affymetrix data are pre-processed by combining the fluorescence levels of individual probes between match and mismatch to yield average differences and expression

1359-6446/02/$ – see front matter ©2002 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(02)02292-4


Gene A
Probes (16)
ACGTGCAATGCCGTACGGCACGTAG
ACGTGCAATGCCGTACGGCACGTAG
Mismatch
ACGTGCAATGCCATACGGCACGTAG
Match
Mismatch
Gene A

DDT: InfoBiotech supplement

Figure 1. The principle of the Affymetrix technology. The probes correspond to short oligonucleotide sequences thought to be representative for a 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). The hybridization occurs at a critical temperature – a single mismatch can totally inhibit hybridization. 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.

calls (present, absent, marginal, etc.). The cDNA data are usually processed by subtracting the background from the fluorescence values of the spots. Furthermore, when data from different arrays are compared, such comparisons must first be made meaningful by normalizing the arrays to 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, a log transformation is applied to improve the characteristics of the distribution of expression values.

The following sections will describe several methods that use the simple example of a comparison between two conditions: experiment and control.

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 to be significant if it is at least two- to three-fold [6–12]. Occasionally, this selection method is used in parallel with expression estimates provided by other techniques such as radioactive- and fluorescent-labeling [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, in experiments 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 that reflect the log fold change, so selecting differentially regulated genes can simply be done by setting thresholds on this axis and then selecting the genes outside such thresholds. For instance, to select genes that have a fold change of 4, the thresholds would have to be set 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 the control using a scatterplot (see Fig. 3), the genes selected will be at a distance of at least 2 from the line of origin (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 also has significant disadvantages. The most important drawback is that the fold threshold is chosen arbitrarily and might often be inappropriate. For instance, if one is selecting genes with at least a twofold change and the condition under study does not affect any genes to the point of inducing a twofold change, no genes will be selected, resulting in zero sensitivity. Equally, 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 amounts to a blind guess.

Another important disadvantage is related to the fact that the microarray technology tends to have a bad signal-to-noise ratio for genes with low expression levels. On a scatterplot, this is illustrated by a funnel shape in the distribution (see yellow curves in Fig. 3) resulting from a large variance in the values measured at the low end of the scale (to the left) and a low variance in the 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 further from the diagonal at a low expression level. As the fold change uses a constant threshold for all genes, false-positives will be introduced at the low end, thus reducing the specificity, and true-positives will be missed at the high end, thus reducing the sensitivity.

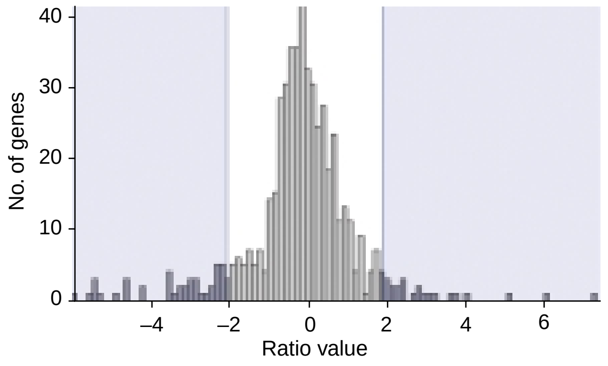
Unusual ratio

The second widely used selection method involves selecting the genes for which the ratio of the experiment and control value is a certain distance from the mean experiment:control ratio [14]. Typically, this distance is taken to be ±2 standard deviations; that is, the genes selected as being differentially regulated will be those genes having an experiment:control ratio of at least 2 standard deviations away from the mean experiment:control ratio. In practice, this can be achieved very simply by applying a z-transformation to the log ratio values. The z-transformation essentially subtracts the mean and divides by the standard deviation. As a 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 ±2 corresponds to selecting those genes that 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 varies considerably. Thus, the unusual ratio method uses thresholds that are based on the difference between the experiment:control ratio of a gene and the mean of all such ratios instead of thresholds based on the values of the ratios themselves. Regardless of how many genes are regulated and irrespective of 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 ±2 standard deviation, this method will select the 5% most regulated genes.

However, the unusual ratio method also has some important intrinsic drawbacks. For example, the method will report the 5% of differentially regulated genes *even if there are no differentially regulated genes*. This happens because in all microarray experiments there is a certain amount of variability owing to noise. Thus, if the same experiment is performed twice, the expression values measured for any particular gene will probably not be exactly the same. If the method is applied to study differentially 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 slightly owing to noise effects. The method will automatically calculate the mean and standard deviation of this distribution and will select those genes situated ±2 standard deviations away from the mean.

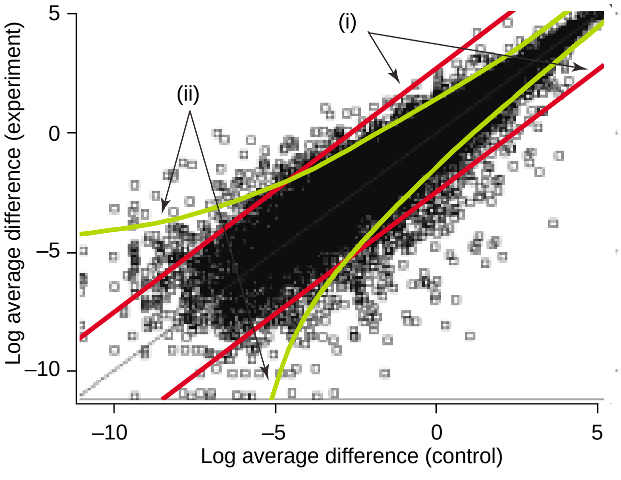
Furthermore, the method still selects 5% of the genes even if many more genes are regulated. Thus, while the fold method uses an arbitrary threshold and can provide too many or too few genes, the unusual ratio method uses a fixed proportion threshold that will always report a given proportion of the genes as being differentially regulated. On a scatterplot (such as 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.



DDT: InfoBiotech supplement

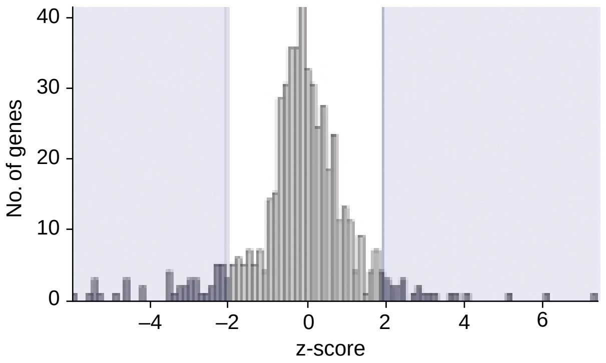
Figure 2. Experiment:control ratios can be plotted in a histogram showing the number of genes (vertical axis) for every ratio value (horizontal axis). The horizontal axis is graded in fold change units. Selecting differentially regulated genes based on fold change corresponds to setting thresholds at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

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 ($\hat{\sigma}$)



DDT: InfoBiotech supplement

Figure 3. A scatterplot representing experimental values plotted against 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 an equal distance from it to the minimum fold change requested (i). A better selection would be achieved with non-linear boundaries that adapt to the increased noise variance at low intensities (ii).



DDT: InfoBiotech supplement

Figure 4. Selecting by the unusual ratio method. The frequencies (vertical axis) of the z-scores of genes (horizontal axis) are plotted in a histogram. The horizontal axis is now graded in standard deviations. Setting thresholds at the ±2 mark corresponds to selecting genes with a ratio of >2 standard deviations from the mean log ratio.

computed for each gene, using array replicates. For duplicate experiments, the absolute difference has to be $>4.3\hat{\sigma}$ and $>22.3\hat{\sigma}$ for 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 5% and 1% significance levels, respectively. Several other *ad hoc* thresholding and selection procedures have also been used. For instance, Schena *et al.* [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 that exhibited at least a twofold change in expression. Although this appears to use the fold method, it can be shown that the combination of the duplicate consistency condition and the differentially regulated condition can be expressed in terms of mean and standard deviations, and it therefore falls under the scope of the unusual ratio method [15].

Confidence levels and corrections for multiple experiments

Another possible approach to gene selection is to use univariate statistical tests (e.g. the t-test) to select differentially expressed genes [15,18,19]. If, for example, 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 classified as differentially regulated (upregulated in this case). However, the measured log ratio could be there merely because of random factors such as noise. The probability of the measurement being there merely by chance is the *p*-value. In this case, classifying the gene 'differentially regulated' would be incorrect, and the *p*-value is the probability of making this type of error (Type I error).

Regardless of the particular test used (usually the t-test), one needs to consider that when many genes are analyzed at one time, some genes will appear as being significantly different merely by chance [20–24]. For example, a gene with a value (e.g. the log-ratio) *v* situated in the tail of the histogram of all such values, possibly indicates that the gene is regulated. The *p*-value provided by the univariate test is the probability that *v* is located here simply by chance. If this gene is classified as differentially regulated based on this value, and the value is there by chance, the classification would be incorrect. Therefore, *p* is the probability of making an error in this test. (From a statistical point of view, interrogating *R* genes at the same time, as on a microarray, is equivalent to running *R* parallel tests.) The probability of drawing the right conclusion in this one test will be $1 – p$. However, if there are *R* such tests, we would like to draw the right conclusion from *all* of them. The probability of this would be $prob(right) = (1 – p)^R$. The probability of making an error would be $prob(wrong) = 1 – prob(right) = 1\ (1 – p)^R$. This is the so-called Sidák correction [25]. Bonferroni [26,27] noted that for a small *p*, $1 – (1 – p)^R \approx Rp$, and proposed to correct the required *p*-value to $\tilde{p} = p/R$. Both the Bonferroni and Sidák corrections are unsuitable for gene expression analysis, because for large numbers of genes *R*, no gene will be below the corrected *p*-value (e.g. $\tilde{p} = p/R$ in the Bonferroni correction).

A family of methods that enables 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]. These methods order the genes by increasing *p*-value and make successive smaller adjustments.

Both the Bonferroni and Sidák corrections, and Holm's step-down adjustment assume that the variables are independent. However, this is not true for expression data because genes influence each other in complex interactions (P. D'haeseller, PhD Thesis, University of New Mexico) [29]. The Westfall and Young (W-Y) step-down is a more general method that 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 [24,30,31]. The method samples with replacement the pool of observations to create new datasets, and calculates *p*-values for all tests. For each dataset, the minimum *p*-value on the resampled datasets 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 [32,33]. Both bootstrap and permutation are computationally intensive.

ANOVA

A particularly interesting approach to both microarray data analysis and selecting differentially regulated genes is a method called ANalysis Of VAriance (ANOVA) [34–36]. The idea behind ANOVA is to build an explicit model about the sources of variance that affect the measurements, and then use the data to estimate the variance of each individual variable in the model.

For instance, Kerr and Churchill [37–39] proposed the following model to account for the multiple sources of variation in a microarray experiment:

$$\log(y_{ijkg}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ijkg} \qquad [1]$$

In this model, μ is the overall mean signal of the array, A_i is the effect of the i^{th} array, D_j is the effect of the j^{th} dye, V_k is the effect of the k^{th} variety (in this context, a variety is a condition such as healthy or diseased), 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 ε_{ijkg} represents the error term for array i, dye j, variety k and gene g. The error is assumed to be independent and have a mean of zero. Finally, $\log(y_{ijkg})$ is the measured log-ratio for gene g of variety k measured on array i using dye j.

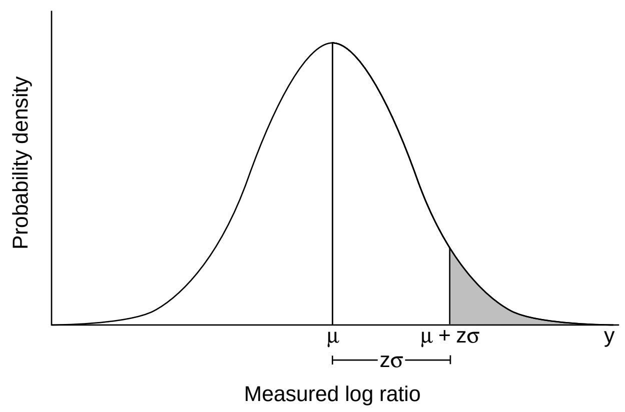
The advantage of ANOVA is that each source of variance is accounted for. Because of this, it is easy to distinguish between interesting variations, such as gene regulation, and side effects, such as differences caused by different dyes or arrays. The caveat is that ANOVA requires very careful experimental design [38,40] that must ensure a sufficient number of degrees of freedom. Thus, ANOVA cannot be used for experiments that have not been designed and executed in a manner consistent with the ANOVA model used.

Noise sampling

A full-scale ANOVA requires a design that blocks all controllable variables and randomizes the others. In most cases, this requires repeating several microarrays with various mRNA samples, and swapping dyes if a multichannel technology is used. A particular variation on the ANOVA idea can be used to identify differentially regulated genes by using spot replicates on single chips to estimate the noise and calculate confidence levels for gene regulation [2,41,42]. The Kerr–Churchill model is modified as follows:

$$\log R(g,s) = \mu + G(g) + \varepsilon(g,s) \qquad [2]$$

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



DDT: InfoBiotech supplement

Figure 5. For a given threshold and a given distribution, the p-value is the probability of the measured value being located 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 an error when classifying a gene as differentially regulated if its measured log ratio is in the shaded area.

In this model, an estimate $\hat{\mu}$ of the average log ratio μ can be calculated as follows:

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

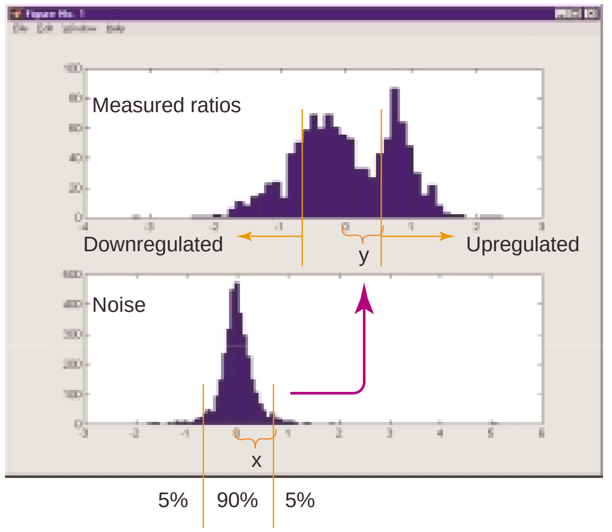
which is the sum of the log ratios for all genes and all spots, divided by the total number of spots (m replicates and n genes). An estimate $\widehat{G(g)}$ of the effect of gene g can also be calculated as:

$$\widehat{G(g)} = \frac{1}{m} \sum_{g} \log[R(g,s)] - \hat{\mu} \qquad [4]$$

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

$$\varepsilon\widehat{(g,s)} = \log[R(g,s)] - \hat{\mu} - \widehat{G(g)} \qquad [5]$$

This will provide a noise sample for each spot. The samples collected from all spots yield an empirical noise distribution. [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.] A given confidence level can be 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 x on the noise distribution can be put into correspondence to a distance y on the measured distribution (Fig. 6) by bootstrapping [24,41]. Furthermore, the dependency between intensity and variance can be taken into account by constructing



DDT: InfoBiotech supplement

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 down-regulated parts of the distribution if the noise is skewed.

several such models covering the entire intensity range, and constructing non-linear confidence boundaries similar to those shown 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 of noise for a given confidence level chosen by the user.

A full-blown ANOVA requires a special experimental design but provides error estimates for all variables considered in the model. The noise sampling method does not require such a special experimental design but only provides estimates for the log ratios of the genes. It has been shown that the noise sampling method provides better sensitivity than the unusual ratio method, and a much better sensitivity than the fold change method [41,42].

Other methods

Other statistically based methods for the selection of differentially regulated genes include model-based maximum likelihood estimation approaches [5,43,44]. A maximum likelihood estimation approach for two color arrays is described by Chen *et al.* [43]. 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 third degree polynomial approximation of the confidence interval as a function of the coefficient of variation *c*. This approach is interesting because it provides the means to deal with signals that are uncalibrated between the two colors through an iterative algorithm that compensates for the color difference. Sapir and Churchill [44] present a robust algorithm for estimating the subsequent 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 a component produced by the differential expression. An expectation maximization algorithm is used to deconvolve the mixture and provide an 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 then selects only those genes for which all replicates indicate that the gene is expressed [5]. In this particular study [5], the approach is used only to make the binary distinction between expressed and non-expressed genes; however, the approach can be extended to multichannel experiments for the detection of differentially expressed genes.

Two hierarchical models (Gamma-Gamma and Gamma-Gamma-Bernoulli) for the two channel (color) intensities are proposed by Newton *et al.* [45]. 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 by Roberts *et al.* [46], 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 equi-confidence curves (see the yellow non-linear boundaries in Fig. 3).

However, as Dudoit *et al.* point out [19], any gene selection method that does not use replication is critically sensitive to the typically large amount of noise associated with microarray data. They 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 a series of experiments with an expected expression profile [47,48]. 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 enable the computation of suitable cut-off thresholds for the degree of similarity. Clearly, these approaches can only be used in the context of large datasets, including several microarrays for each condition considered.

Other methods used for the selection of differentially regulated genes include gene shaving [49], assigning gene confidence [50] or significance [51], bootstrap [30,31], and Bayesian approaches [52–54]. Some methods also take into consideration that the variance depends on the intensity [45,46].

Finally, more elaborate methods for the analysis of gene expression data exist. Such methods include clustering [12,15,55–66]), principal component analysis [57,67,68], singular value decomposition [69], independent component analysis [70], gene shaving [49] and many others. The goals of such methods go well beyond the selection of differentially regulated genes and are, as such, outside the scope of this review.

Concluding remarks

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 widely used, the early methods of selection by fold change and unusual ratio are clearly inadequate. Using a fold change method without a clear biological justification is merely 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 severe drawbacks and their use as methods for selecting differentially regulated genes should be discontinued. However, studying the fold change of genes of known function is and will continue to be important. In order words, computing statistics as required by biological reasons is fully justified (e.g. 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 (e.g. data obtained from most commercial chips), Bonferroni should be taken as a sufficient but not a 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 to be differentially regulated after the Bonferroni correction, the gene might or might 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 present, the bootstrap method and W-Y 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 experimental 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 require involving a statistician from the very beginning, and designing the experiment in such a way that enough degrees of freedom are available to answer the relevant biological questions. The noise sampling method is a variation of ANOVA that enables the automatic computation of noise-dependent equi-confidence boundaries. Furthermore, the noise sampling method can be used in many instances in which data for a full-scale ANOVA are unavailable.

References

1 Schena, M. (2000) *Microarray Biochip Technology*, Eaton Publishing

2 Draghici, S. *et al.* (2001) Experimental design, analysis of variance and slide quality assessment in gene expression arrays. *Curr. Op. Drug Discov. Devel.* 4, 332–337

3 Schuchhardt, J. *et al.* (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res*. 28, e47i–e47v

4 Wildsmith, S.E. *et al.* (2000) Maximizing of signal derived from cDNA microarrays. *BioTechniques* 30, 202–208

5 Ting Lee, M-L. *et al.* (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. U. S. A.* 97, 9834–9839

6 Schultz, P.G. *et al.* (2001) The effects of aging on gene expression in the hypothalamus and cortex of mice. *Proc. Natl. Acad. Sci. U. S. A.* 98, 1930–1934

7 DeRisi, J.L. *et al.* (1996) User of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14, 457–460

8 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686

9 ter Linde, J.J.M. *et al.* (1999) Genome-wide transcriptional analysis of aerobic and anaerobic chemostat cultures of *Saccharomyces cerevisiae*. *J. Bacteriol.* 181, 7409–7413

10 Sudarsanam, P. *et al.* (2000) Whole-genome expression analysis of snf/swi mutants of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 3364–3369

11 Wellmann, A. *et al.* (2000) 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, 398–404

12 White, K.P. *et al.* (1999) Microarray analysis of *Drosophila* development during metamorphosis. *Science* 286, 2179–2184

13 Richmond, C.S. *et al.* (1999) Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Res.* 27, 3821–3835

14 Tao, H. *et al.* (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol.* 181, 6425–6440

15 Claverie, J-M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* 8, 1821–1832

16 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 467–470

17 Schena, M. *et al.* (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U. S. A.* 93, 10614–10519

18 Audic, S. and Claverie, J-M. (1998) Vizualizing the competitive recognition of TATA-boxes in vertebrate promoters. *Trends Genet.*14, 10–11

19 Dudoit, S. *et al.* (2000) Statistical models for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, University of California, Berkeley, CA, USA

20 Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statistics* 6, 65–70

21 Hochberg, Y. and Tamhane, A.C. (1987) *Multiple Comparison Procedures*, Wiley

22 Shaffer, J.P. (1986) Modified sequentially rejective multiple test procedures. *J. Am. Statis. Assoc.* 81, 826–831

23 Shaffer, J.P. (1995) Multiple hypothesis testing. *Ann. Rev. Psychol.* 46, 561–584

24 Westfall, P.H. and Young, S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, John Wiley & Sons

25 Sidák, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Statis. Assoc.* 62, 626–663

26 Bonferroni, C. E. (1935) *Il calcolo delle assicurazioni su gruppi di teste*. Chapter Studi in Onore del Professore Salvatore Ortu Carboni, pp. 13–60, Rome

27 Bonferroni, C.E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3–62

28 Holland, B. and Copenhaver, M.D. (1987) An improved sequentially rejective Bonferroni test procedure. *Biometrica*, 43, 417–423

29 D'haeseller, P. *et al.* (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 8, 707–726

30 Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39, 783–791

31 Kerr, M.K. and Churchill, G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. U. S. A.* 98, 8961–8965

32 Brown, C.C. and Fears, T.R. (1981) Exact significance levels for multiple binomial testing with application to carcinogenicity screens. *Biometrics* 37, 763–774

33 Heyse, J. and Rom, D. (19988) Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biometrical J.* 30, 883–896

34 Aharoni, A. *et al.* (1975) Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. *Plant Cell* 12, 647–661

35 Brazma, A. and Vilo, J. (2000) Gene expression data analysis. *FEBS Lett.* 480, 17–24

36 Hill, A.A. (2000) Genomic analysis of gene expression in *C. elegans*. *Science* 290, 809–812

37 Kerr, M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comp. Biol.* 7, 819–837

38 Kerr, M.K. and Churchill, G.A. (2001) Statistical design and the analysis of gene expression. *Genet. Res.* 77, 123–128

39 Kerr, M.K. and Churchill, G.A. (2001) Analysis of variance for gene expression microarray data. *J. Comp. Biol.* 7, 819–837

40 Kerr, M.K. and Churchill, G.A. (2001) Experimental design for gene expression analysis. *Biostatistics* 2, 183–201

41 Draghici, S. *et al.* Computational methods for the selection of differentially regulated genes in cell immortalization. *J.Comp. Biol.* (in press)

42 Wang, D. *et al.* Methods for selecting differentially regulated genes in microarrays: noise sampling vs. standard deviations. *Bioinformatics* (in press)

43 Chen, Y. *et al.* (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2, 364–374

44 Sapir, M. and Churchill, G.A. (2000) Estimating the posterior probability of differential gene expression from microarray data. Technical Report, Jackson Labs, Bar Harbor, ME, USA

45 Newton, M.A. *et al.* (1999) On differential variability of expression ratios: improving statistical inference about gene expresison changes from microarray data. Technical report, University of Wisconsin, Madison, WI, USA

46 Roberts, C.J. *et al.* (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287, 873–880

47 Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537

48 Galitski, T. *et al.* (1999) Ploidy regulation of gene expression. *Science* 285, 251–254

49 Hastie, T. *et al.* (2000) 'Gene shaving' as a method for indentifying distinct sets of genes with similar expression patterns. *GenomeBiology* 1, 1–21

50 Manduchi, E. *et al.* (2000) Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* 16, 685–698

51 Tusher, V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121

52 Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17, 509–519

53 Long, A.D. *et al.* (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *J. Biol. Chem.* 276, 19937–19944

54 West, M. *et al.* (2000) Bayesian regression analysis in the 'large p, small n' paradigm with application in DNA microarray studies. Technical report, Duke University, Durham, NC, USA

55 Aach, J. *et al.* (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.* 10, 431–445

56 Brazma, A. (1998) Mining the yeast genome expression and sequence data. *The BioInformer* 4, (http://bioinformer.ebi.ac.uk/newsletter/archives/4/lead_article.html)

57 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868

58 Ewing, R.M. *et al.* (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* 9, 950–959

59 Heyer, L.J. *et al.* (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* 9, 1106–1115

60 Pietu, G. *et al.* (1999) The genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. *Genome Res.* 9, 195–209

61 Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912

62 Tsoka, S. and Ouzounis, C.A. (2000) Recent developments and future directions in computational genomics. *FEBS Lett.* 23897, 1–7

63 van Helden, J. *et al.* (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.* 28, 1808–1818

64 Wang, M.L. *et al.* (1999) A cluster of ABA-regulated genes on *Arabidopsis Thaliana* BAC T07M07. *Genome Res.* 9, 325–333

65 Zhang, M.Q. (1999) Large-scaled gene expression data analysis: a new challenge to computational biologists. *Genome Res.* 9, 681–688

66 Zhu, J. and Zhang, M.Q. (2000) Cluster, function and promoter: analysis of yeast expression array. *Pac. Symp. Biocomp.* 479–490

67 Hilsenbeck, S.G. *et al.* (1999) Statistical analysis of array expression data as applied to the problem of Tamoxifen resistance. *J. Nat. Cancer Insti.* 91, 453–459

68 Raychaudhuri, S. *et al.* (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Proc. Pac. Symp. Biocomp*. 5, 452–463

69 Alter, O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106

70 Liebermeister, W. (2001) Independent component analysis of gene expression data. *Proc. German Conf. Bioinformatics* (http: / / www.bioinfo.de / isb / gcb01 / poster / index.html)

The best of drug discovery at your fingertips

www.drugdiscoverytoday.com

Stop at our new website for the best guide to the latest innovations in drug discovery

High quality printouts (from PDF files) and links to other articles, other journals and cited software and databases

All you have to do is:
Obtain your subscription key from the address label of your print subscription.
Go to http: / / www.drugdiscoverytoday.com
Click on the '**Claim online access**' button below the current issue cover image.
When you see the BioMedNet login screen, enter your BioMedNet username and password.
Once confirmed you can view the full-text of *Drug Discovery Today.*

See if you qualify to receive your own free copy, which will also entitle you to free full-text access online.
Simply click on the '**Get your FREE trial subscription**' tab at the top of the page.

If you get an error message please contact Customer Services (info@current-trends.com).
If your institute is interested in subscribing to the print journal, please ask them to contact ct.subs@qss-uk.com