Published online July 20, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 13 ***3807–3814***
doi:10.1093/nar/gkh718

Mathematical algorithm for discovering states of expression from direct genetic comparison by microarrays

Hassan M. Fathallah-Shaykh*, Bin He, Li-Juan Zhao and Aamir Badruddin

Department of Neurological Sciences, Section of Neuro-Oncology, Rush University Medical Center,
1725 West Harrison Street, Chicago, IL 60612, USA

Received May 11, 2004; Revised June 12, 2004; Accepted July 2, 2004

ABSTRACT

Highly specific direct genome-scale expression discovery from two biological samples facilitates functional discovery of molecular systems. Here, expression data from cDNA arrays are ranked and curve-fitted. The algorithm uses filters based on the derivatives (slopes) of the curve fits. The rules are set to (i) filter the largest number of artifactual ratios from same-to-same datasets and (ii) maximize discovery from direct comparisons of different samples. The unsupervised discovery is optimized without lowering specificity. The false discovery rates are significantly lower than other methods. The discovered states of genetic expression facilitate functional discovery and are validated by real-time RT–PCR. Better quality improves sensitivity.

INTRODUCTION

Several genomes have recently been sequenced and the development of DNA microarrays has facilitated genome-scale expression profiling (1,2). In a single assay, the transcriptional differences between phenotypes are measured (3–6). Furthermore, the idea that the global transcriptional response constitutes molecular phenotypes has only recently received attention (3,5,7–14). In this model, phenotypes are created by molecular systems in which single genes or molecules belong to rich networks of molecular interactions that include transcriptional regulation, signaling pathways, protein–protein and protein–nucleic acid interactions (15–18).

Current methods for microarray expression data analysis require numerous samples and yield low specificity (7,9,19,20). Thus, validation by other methods for measuring gene expression has become the ‘gold standard’ (21–24). However, biological samples are not always abundant, and validation of all discovered genes is a very expensive and time-consuming endeavor. The cost is prohibitive. The price tag for validating all the genes discovered by genome-scale expression profiling is in the order of tens of thousands of dollars per question or experiment. The cost to the whole biomedical community is astronomical.

Thus, a method that generates highly specific genome-scale expression discovery from two samples is not only cost effective but also very desirable. To be effective, the false discovery rate of such an algorithm should be ‘small enough’ to convey a high degree of confidence that the ‘discovered’ genes are truly differentially expressed between samples. This algorithm sets the stage for functional genomics by facilitating the discovery of molecular systems and the prediction of gene-to-gene interactions, signaling pathways and protein states behind phenotypes (25). The idea that quality controls sensitivity is rather intuitive; one expects high quality images to yield a sharper separation of true from false and to discover smaller expression ratios (higher sensitivity).

MATERIALS AND METHODS

Microarrays

Normal brain RNA is obtained by pooling RNA from human occipital lobes harvested and pooled from four individuals with no known neurological disease whose brains are frozen less than 3 h postmortem. The quality of RNA is assayed by gel electrophoresis and only high-quality RNA is processed. Total RNA (5–10 μg) is reverse transcribed and the cDNA products labeled by the amino-allyl method and hybridized to the 1.7K and 19K cDNA microarrays purchased from the Ontario Cancer Institute (Toronto, CA). The slides are scanned at 10 μm by a confocal scanner, (4000XL scanner; Packard Bioscience, Meriden, CT). Spot signals are quantified by the Imagene Software (Biodiscovery; Los Angeles, CA).

Real-time RT–PCR

Total RNA samples are analyzed by one-step, hot-start real-time RT–PCR (Qiagen, Valencia, CA; Cepheid, Sunnyvale, CA), and normalized to G3PDH as described elsewhere (26). Primer pairs are generated for each of the 21 genes as well as G3PDH (Supplementary Material).

Analysis

The mathematical analysis is performed using functions written in Matlab (Mathworks, Natick, MA).

*To whom correspondence should be addressed. Tel: +1 312 563 3563; Fax: +1 312 563 3562; Email: hfathall@rush.edu

RESULTS

Definitions

The 'state of genetic expression' of a spot in sample A versus sample B assayed by cDNA arrays is measured by the ratio of the background-subtracted intensities of Sample A/background-subtracted intensities of Sample B. A ratio > 1 ($\log_2 > 0$) implies up-regulation of the gene in sample A as compared to sample B; a ratio < 1 ($\log_2 < 0$) implies down-regulation in A as compared to B. We use the human 1.7K microarray chip to define the terms 'genes', 'spots', 'symmetrical', 'rank' and 'spot order'. These terms are also applicable to other microarrays. The 1.7K microarray chip contains 1920 cDNAs or controls, here referred to as 'genes', spotted in duplicates to a total of 3840 'spots'. The term 'symmetrical' refers to the two images, corresponding to the Cy3 and Cy5 fluorescent dyes, generated from a single microarray slide. Probe switching (dye swapping) refers to experiments where the Cy3 and Cy5 dyes are switched between the two samples to be compared; they are performed to annul confounding variables introduced by heterogeneous fluorescence of the Cy5 and Cy3 molecules.

To model the dynamic range, background-subtracted spot intensities are sorted in ascending order (y-axis) and plotted to generate the ranking curve $I(x)$ (Figure 1a). The x-axis of Figure 1a is a listing of the spots of a 1.7K dataset ranked in ascending order by their background-subtracted intensities; the x-axis coordinate corresponding to a specific spot is defined as its 'Rank'. For instance, a spot whose rank is 3000 has a higher background-subtracted spot intensity than all spots whose ranks are less than 3000. A microarray Spot Order (SO) is a listing of its spots sorted dby their ranks. Figure 1b is a plot of the log-transformation of the data in Figure 1a; it reveals a family of curves consisting of three parts: (i) an initial segment where spot intensities rise rapidly, (ii) a second almost 'linear' section associated with small increments and (iii) an 'exponentially growing' phase.

The datasets and rationale

The true negative datasets compare the same pool of brain RNA to itself (same-to-same). The goal of the same-to-same

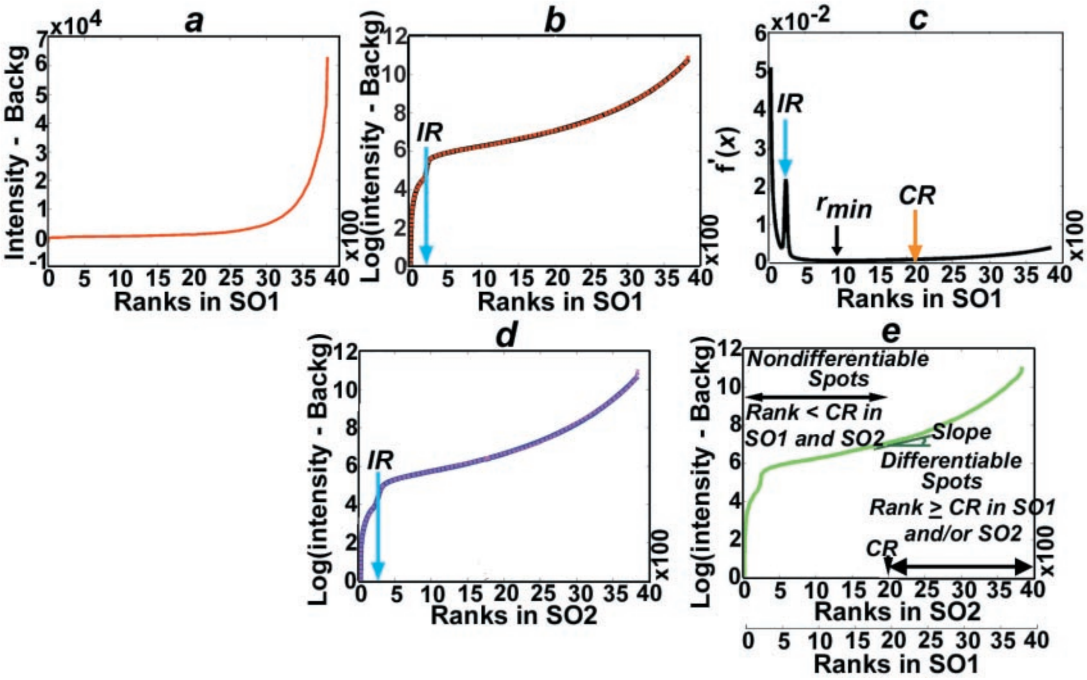

Figure 1. Curve fitting and normalization. Each spot generates two measurements of (i) the total intensity within the spot and (ii) the local background intensity defined as the total intensity within a small rim surrounding the spot. (**a**) Is a plot of the background-subtracted spot intensities (y-axis) versus spot ranks (x-axis) of a dataset acquired from the 1.7K-microarrays. (**b**) Red dashed line, is a log transformation of the dataset of (a). The dataset of (b) is curve-fitted to Equation 4:

$$f(x) = \left(\frac{a_1 * x}{x + a_2} + \frac{x}{ns - x + a_3} - a_4\right) * a_5 * \left(\frac{1}{1 + (a_7/x)^{a_6}} + \frac{a_8}{1 + (a_{10}/x)^{a_9}} - \frac{a_{11}}{x + a_{12}}\right) * \left(1 + \frac{a_{13}}{1 + |1 - (a_{15}/x)|^{a_{14}}}\right) + \left(\frac{1}{1 + (a_{17}/x)^{a_{16}}} - a_{18}\right) * a_{19} \qquad 4$$

where x refers to rank. (b, black line) shows the plot of Equation 4 that best fits the experimental dataset (red dashed). ns refers to the total number of spots in the array; $ns = 3840$ for the 1.7K microarrays. Parameters ($a_1, \ldots, a_{19}$) vary between individual curves. (**c**) Is a plot of the derivative corresponding to curve shown in (b) (Equation 1.2, Supplementary Material). (**d**) Shows the raw dataset (magenta dashed), curve-fit (blue) of the image symmetrical to the dataset shown in (b). (**e**) Illustrates the normalization of the dataset of (d) to model the curve of (b); the y-coordinates of the ranks of SO2 are transformed to become equal to the y-coordinates of equal ranks in SO1 (see Supplementary Material). The algorithm transforms the image with the lower CR to model the other. The cyan arrow points to the Inflection Rank (IR). $r_{\min}$ refers to the rank where the curve of the derivative reaches a minimum (c). The orange arrow transects the x-axis at the CR. Non-Differentiable Spots are those whose ranks are less than CR in *both* SO1 (b) and SO2 (d). Filter 1 excludes Non-Differentiable Spots.

comparisons is to collect experimental noise (artifacts) independent of biological heterogeneity. In this design, normalized expression ratios $\neq 1$ ($\log_2 \neq 0$) are false positive (noise) because the symmetrical images contain identical genetic information. The artifactual measurements may be caused by several factors including slide-to-slide differences, variations in the reverse transcription reactions, hybridization, labeling and laser. We perform 18 and 20 same-to-same experiments to generate a total of 9 and 10 probe-switching datasets using the human 1.7K and 19K microarrays, respectively. The experiments are paired by consecutive order. The goal of the algorithm is to filter the largest number of same-to-same expression ratios. Ideally, the algorithm is expected to filter all technical noise and discover no gene as being differentially expressed in same-to-same datasets.

The different-to-different datasets compare: (i) a meningioma sample to brain in probe switching experiments using 1.7K microarrays and (ii) 10 meningioma samples to normal brain using the 19K microarrays. The goal of the algorithm is to discover the largest number of genes differentially expressed between different samples. Ideally, all genes discovered from different-to-different datasets will be truly differentially expressed between meningioma and brain.

Curve fitting

The main objective of our curve fitting is to generate a differential equation that models the changes in slope versus rank. Unlike smooth curves, experimental data show point-to-point variations, which limit the accurate representations of the slopes. Here, we apply a stochastic global fit approach to construct a mathematical Equation 4 whose plots fit a smooth curve through the data points in such a way that the points are as 'close' to the curves as possible ($R^2 > 0.99$; Figure 1b). Equation 4 contains 19 parameters ($a_1, \ldots, a_{19}$) that are optimized within defined bounds to fit the heterogeneous members of this family of curves (see Supplementary Material). Equation 4 fits not only our data of 60/60 1.7K datasets, 200/200 human 19K datasets (38 400 spots on two separate slides P1 and P2), but also all 266 curves resulting from the 133 publicly available arrays from the lymphoma study by Alizadeh *et al.* (7) ($R^2 > 0.99$; see Supplementary Material). Each curve-fit generates a unique set of parameters ($a_1, \ldots, a_{19}$) determined by the function lsqcurvefit (MATHLAB, Optimization Toolbox), which uses the large-scale algorithm to solve the non-linear curve-fitting problem in the least-squares sense.

The complexity (19 parameters) of Equation 4 is not limiting because of the speed of current computers. Other equations of different forms may also be deduced. Nonetheless, having constructed an equation that fits the curves, its derivative is at hand (Equation 1.2, see Supplementary Material). Figure 1c plots the curve of the derivative $f'(x)$. Because the smooth curves of Equation 1.2 lack the fluctuations of biological data, they generate important tools that will be applied to filter technical noise and discover true states of genetic expression.

Normalization

Figure 1b and d show the datasets of two symmetrical images (CY3 and CY5). Some of the false-positive ratios are expected to be more than 1 and others less than 1. The idea that the majority of the genes are not differentially expressed between samples implies that the product of all the ratios is equal to 1. This idea leads to the derivation of a local normalization scheme, which transforms the curve of one dataset (Figure 1d) to model the other (Figure 1b and Supplementary Material).

The normalized curve is plotted in a graph having two separate *x*-axes corresponding to SO1 and SO2 (CY3 and CY5; Figure 1e). Each spot is ranked separately in the SO of each image (SO1 and SO2). The normalized expression ratio of a spot is computed as

$$\frac{\text{Normalized intensity of its rank in SO2}}{\text{Normalized intensity of its rank in SO1}} \qquad \mathbf{1}$$

Thus, if $g(x)$ is the log-function of the normalized curve, and if a and b are the symmetrical ranks of a single spot (see Supplementary Material), then the

$$\text{Normalized ratio} = \frac{e^{g(a)}}{e^{g(b)}} = e^{g(a) - g(b)} \qquad \mathbf{2}$$

Mathematical properties of the curves

Next, we study the slopes of the normalized curves. The first segment of the curve in Figure 1e rises rapidly; the rate of increase is maximal at the point of inflection that corresponds to the maximum of $f'(x)$ in that segment (Figure 1c). The rank (*x*-coordinate) of the point of inflection is defined as the Inflection Rank (IR). The 1.7K chips include 256 'buffer' spots containing no cDNA, which are expected to generate the lowest intensities caused by non-specific binding of the probes to glass or buffer. The *y*-coordinate at the IR corresponds to a small background-subtracted intensity, ranging from 50 to 150, most probably generated by non-specific probe binding. Interestingly, because the datasets of Alizadeh *et al.* (7) lack buffer spots, their curves show steep rise in the slope of the first segment reaching the IR very quickly (see Supplementary Material). We conclude that the majority of intensities whose ranks are smaller than the IR are likely caused by non-specific binding of the probe.

After reaching the inflection point, the curve of $f'(x)$ decreases to a minimum corresponding to a rank, $r_{\min}$, then increases again (Figure 1c and see Supplementary Material); $r_{\min}$ is very close to 0. To illustrate the applicability of the derivative, we study the specific example when the derivative of the equation is equal to 0. Here, the line is parallel to the *x*-axis. Because all background-subtracted intensities are equal, spots whose symmetrical ranks correspond to that line cannot be separated or differentiated based on their expression levels. In this specific example, because the line includes the ranks of about half the gene set and because overall gene expression is expected not to differ between the two samples, the predominant majority of genetic measurements whose ranks map to that line are expected to be false.

The numbers that follow pertain to the datasets acquired from the 1.7K microarrays; the same applies to the 19K microarrays after changing the total number of spots from 3840 to 192 000 per slide. If $g(x)$ is the log-function of the normalized curve, the Cutoff Rank (CR) is defined as a rank within the interval $[r_{\min},3840]$ such as

$$g'(\mathrm{CR}) = \delta_{\mathrm{optimal}} * \frac{g(3840)}{3840} \qquad 3$$

When completed, the algorithm computes $\delta_{\mathrm{optimal}}$ as a value within the interval [0.3, 0.4]. $\delta_{\mathrm{optimal}}$ is computed empirically to optimize sensitivity without lowering specificity (see Computing CR below). The algorithm computes an individual $\delta_{\mathrm{optimal}}$ for each dataset.

The CR maps to the junction of the second and third parts of the curve. Differentiable Spots are defined as those having at least one symmetrical rank larger than the CR. Non-Differentiable Spots are defined as those whose ranks are smaller than the CR in both symmetrical CY3 and CY5 images. Non-Differentiable Spots are filtered because their measurements may either be caused by non-specific probe binding to glass or buffer and/or a fall within the linear part of the curve where the slope is close to 0.

Filter 1

The same-to-same 19K and 1.7K datasets contain a total of 192 000 and 17 280 ratios, respectively. Histograms of unfiltered same-to-same ratios reveal that most $\log_2 \neq 0$ fall within the interval [−1,1]. However, they range from −10 to 10 and a large number are outside [−1,1] (Figure 2a).

Filter 1 (F1) is applied to individual 'spots' and is defined by the following rules.

F1 filters a spot by transforming its expression ratio to 1 ($\log_2 = 0$):

(i) If the spot is Non-Differentiable (see Figure 1) or
(ii) If its background intensity lies outside the mean ± 2SD of the background intensities of all spots.

F1-resistant spots must have at least one symmetrical rank larger than the CR. A spot, whose ranks are both less than the CR, is filtered.

Noise factor

Histograms of same-to-same datasets reveal the presence of F1-resistant noise with variable variance about the origin (Figure 2b and c). The notion that overall expression does not differ between two samples stipulates that the ranks of the predominant majority of spots in symmetrical SOs are similar. This idea should be especially true in same-to-same datasets where the symmetrical images contain identical genetic information. To study rank variability, spots from

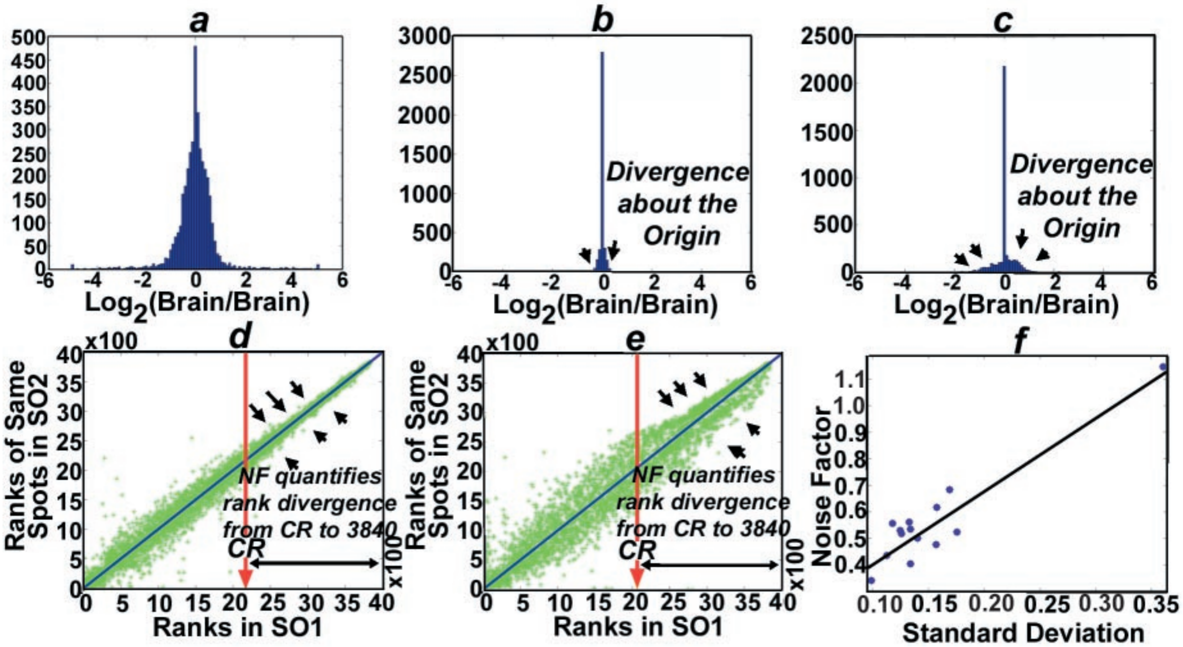

Figure 2. Linear correlation between the Noise Factor and SD of F1-resistant noise. (**a**) Shows a histogram of the unfiltered $\log_2$ transformed normalized ratios of an experiment where the Cy3 and Cy5 images correspond to the same RNA (technical noise). The histograms in (**b** and **c**) show F1-resistant false positive ratios from two distinct same-to-same experiments; (a and c) correspond to the same dataset. The SD of the data in (b and c) are 0.1 and 0.36, respectively. (**d** and **e**) Show the scatter plots of the symmetrical ranks of the experiments shown in (b and c), respectively. Red arrows transect the x-axis at the CRs. The degree of divergence (arrows) about $y = x$ differs between (d and e). The results suggest a relationship between the divergence of F1-resistant noise about the origin [(b and c) arrows] and the divergence of symmetrical ranks about the line $y = x$ [(d and e), arrows]. To study this idea, we define the Noise Factor (NF). Let (r_{1i}, r_{2i}) denote the coordinates of the n spots whose ranks in symmetrical Spot Orders are larger than the CR, then

$$\mathrm{NF} = \sqrt{\frac{\sum_{i=1}^{n}(r_{1i} - r_{2i})^2}{n}} * \frac{\alpha}{(3840 - \mathrm{CR})} \qquad 5$$

where α is any scalar; here we use α = 10. The Noise Factor quantifies the degree of divergence about the segment of $y = x$ that extends from the CR to 3840. (**f**) Is a scatter plot of the Noise Factors versus SD of F1-resistant noise in the 18 sets of 1.7K symmetrical images ($y = 2.75 * x + 0.12$, $R^2 = 0.823$).

the same-to-same datasets are plotted by their ranks in SO1 (x-axis) and SO2 (y-axis); as anticipated, the data scatter about the line $y = x$ ($R^2 > 0.9$; Figure 2d and e). However, the degree of divergence from $y = x$ varies between arrays despite the fact that each array compares the same RNA to itself (Figure 2d and e). This observation suggests the hypothesis that the degree of divergence of the symmetrical ranks about $y = x$ (Figure 2d and e) determines the 'margin' or variance of F1-resistant noise about the origin (Figure 2b and c).

To study this idea, we define the Noise Factor that quantifies the degree of divergence of the ranks in symmetrical SOs about the segment of $y = x$ that extends from the CR to 3840 (Figure 2d and e). Figure 2f reveals a linear correlation between the Noise Factor and the SDs of F1-resistant noise. The findings identify the Noise Factor as an important quality parameter. Thus, datasets whose Noise Factors are small ('better quality') are *not* likely to contain 'large' F1-resistant false-positive ratios (see Figure 2b and d). On the other hand, large F1-resistant ratios in 'lower quality' images may be inaccurate (see Figure 2c and e).

Filter 2

Next we set out to eliminate F1-resistant noise from same-to-same datasets regardless of the quality of the images. Probe-switching experiments generate 2 Noise Factors and 2 SDs, each corresponding to a set of symmetrical images. Because each microarray slide contains genes spotted in duplicate, the experiments generate four replicate ratios for each gene.

In the glioma study, we devised a noise model and a filter, f_4, and showed that f_4 generates a false negative rate of only 1.6% (26). f_4 is applied to four replicate ratios and includes a rule, named f_0, that requires all four replicate $\log_2$ (ratios) of resistant genes to be of the same sign and different from 0. We have also shown that the overwhelming majority of f_4-resistant noise vectors project onto the eigen space at distances from the origin that are within three SDs from the mean. Hence, we define the second filter.

Filter 2 (F2) is applied to F1-resistant 'genes' and is defined by the following rules:

An F1-resistant gene is filtered by transforming its expression ratio to 1 ($\log_2 = 0$) *unless* all four replicate $\log_2$ (ratios) are

(i) Of the same sign and different from 0 (consistently showing up- or down-regulation; same rule as f_0) and
(ii) At distances from the origin larger than 3× the largest SDs of the probe-switching experiments.

The algorithm outputs the mean values of the four replicate $\log_2$ (ratios) of F2-resistant genes.

Computing CR

Our goal is to find the optimal value of CR that maximizes sensitivity without lowering the specificity of discovery. Figure 3a–c shows the effects of varying δ (see Equation 3) on the false discovery rate when the algorithm is applied to the analysis of the 10 and 9 same-to-same 19K and 1.7K datasets, respectively. The results reveal that the specificity is high for values of δ within the interval [0.3, 0.4] (Figure 3b and c). Thus to optimize both sensitivity and specificity, the algorithm varies δ within the interval [0.3, 0.4] to determine a $\delta_{\mathrm{optimal}}$, which discovers the largest number of genes (Figure 3d). An individual CR is computed from the $\delta_{\mathrm{optimal}}$ of each dataset.

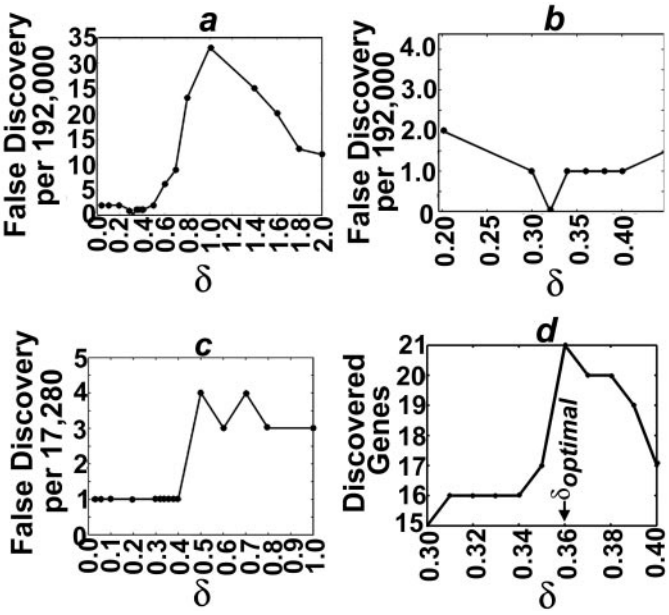

Figure 3. The CR is computed to optimize sensitivity without lowering specificity. To study the effects of varying δ on the false discovery rates, the complete algorithm is applied to the same-to-same 19K (**a** and **b**) and 1.7K (**c**) datasets (see Equation 3). (b) Plots the same data as (a) for δ between 0.2 and 0.5. (a–c) Illustrate the effects of varying δ on the false discovery rates. (b and c) Demonstrate that values of δ within the interval [0.3, 0.4] yield high specificity. Thus for each dataset, the algorithm varies δ within the interval [0.3, 0.4]. To illustrate the computation of $\delta_{\mathrm{optimal}}$, (**d**) plots the number of discovered genes versus δ for the 1.7K meningioma-to-brain dataset. In this specific example, $\delta_{\mathrm{optimal}} = 0.36$ discovers the largest number of genes. The algorithm computes a unique $\delta_{\mathrm{optimal}}$ for each dataset.

The algorithm is effective in filtering noise

Of all 9 probe-switching, same-to-same experiments using the 1.7K human microarrays, only 1 of the total of 17 280 (9*1920) genes is resistant to both F1 and F2 (Table 1). Changing the first rule of F2 to requiring only 3 $\log_2$ measurements instead of all 4 to be of the same sign, results in an 8-fold increase in the number of false positive genes. The complete algorithm is then applied to the data of the 10 probe-switching, same-to-same experiments (brain-Cy3 and same brain-Cy5) using the 19K chips. Only 1 of the 192 000 genes is resistant to both F1 and F2 (Table 1). The complete algorithm is also applied to analyze four same-to-same datasets of Rosenzweig *et al.* (27). Each dataset includes 710 'genes' spotted in duplicates to a total of 1420 spots (see Supplementary Material). The false discovery rate is 0 of 2840. Because the same-to-same arrays have a heterogeneous quality (Figure 2), the findings demonstrate the effectiveness of the algorithm in filtering noise regardless of the quality of the dataset.

To evaluate the false discovery rates in different-to-different datasets, we apply the algorithm to analyze the 19K datasets of the 10 meningioma samples. Each array includes 128 'genes' of Arabidopsis. In these experiments, the Arabidopsis cDNA genes serve as true negatives because both meningioma and brain RNAs are spiked by an equal amount of Arabidopsis RNA (1 ng). The false discovery rate of the algorithm is 0/1280. Thus, the high specificity of the algorithm is also true in different-to-different comparisons.

Discovered genetic expression states are validated

The algorithm discovers 21 genes from the 1.7K dataset comparing the meningioma sample versus brain (Table 1). Real-time RT–PCR, a semiquantitative method for comparing gene expression, confirms the states of genetic expression of all 21/21 genes (Figure 4a). The G3PDH-normalized ratios (Figure 4a) are corroborated by the expression profiling of the discovered genes in 10 other meningiomas using the 19K microarray chips (Figure 4b).

When applied to the analysis of the 19K datasets comparing 10 meningiomas to normal brain, the algorithm discovers 364 as being consistently up- or down-regulated in a minimum of 5/10 meningiomas. The discovered states of genetic expression combined with current knowledge in biological chemistry accurately predict activation of signaling pathways and opposing molecular functions behind phenotypes. For example, the data predict activation of the Wnt, ERK and Akt pathways in meningiomas and reveal opposing molecular functions behind the phenotype of enhanced transcription, growth, remodeling of the cytoskeleton and extracellular matrix, angiogenesis and immunological evasiveness (25).

Table 1. The mathematical algorithm is effective in filtering same-to-same technical noise

Comparison	Discovered genes	No. of genes	Microarray
Brain RNA versus brain RNA	1	192 000	19K
Brain RNA versus brain RNA	1	17 280	1.7K
Meningioma versus brain RNA	21	1920	1.7K

Only 1 of the total of 17 280 genes analyzed was not excluded in the 9 probe switching experiments comparing normal brain RNA to itself (1.7K chip). In addition, an only 1 of 192 000 genes analyzed in 10 probe-switching, same-to-same experiments was not filtered (19K chip). The algorithm discovers 21 genes out of 1920 from the comparison of meningioma versus brain.

Sensitivity is dependent on quality

To generate different-to-different datasets of heterogeneous quality, the experiment, comparing a meningioma RNA versus brain using the 1.7K microarrays, is repeated four times using aliquots from the same meningioma and brain RNAs. Figure 5 reveals a negative correlation between the Noise Factor and sensitivity; the higher the Noise Factor, the lower the number of discovered genes. Therefore, the Noise Factor is a quality parameter that predicts sensitivity.

High specificity as compared to other methods

To compare the specificity of the algorithm to others, we have analyzed both the 1.7K and 19K same-to-same datasets by (i) our algorithm and (ii) the TIGR MIDAS software (release August 2003, http://www.tigr.org/software) (28,29) (see Table 2). For the same-to-same 19K datasets, the false discovery rates of our algorithm, standard MIDAS, and high stringency MIDAS are 1/192 000, 1347/192 000 and 932/

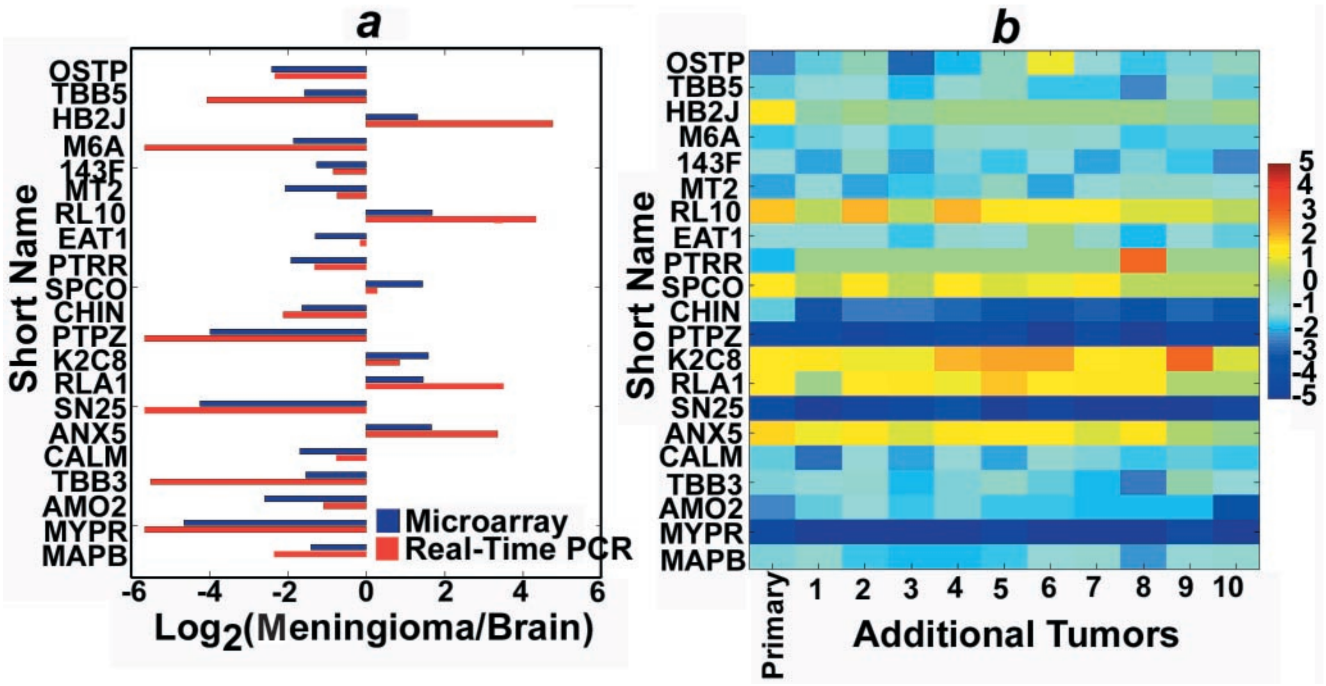

Figure 4. The mathematical algorithm discovers highly specific states of genetic expression. Real-time RT–PCR validates all 21/21-discovered genes [(**a**) and Supplementary Material]; the expression ratios (meningioma/normal brain) are capped at 50- and 0.02-fold. (**b**) Shows the $\log_2$ measurements discovered by the algorithm from the profiling of a meningioma against normal brain by the 1.7K chips (Primary), and the $\log_2$ transformed normalized but unfiltered ratios in 10 other meningiomas profiled by the 19K microarray chips also against normal brain (Additional Tumors 1–10). Here, colors other than green ($\log_2 \neq 0$) indicate that all four measurements consistently show either up- or down-regulation (rule f_0).

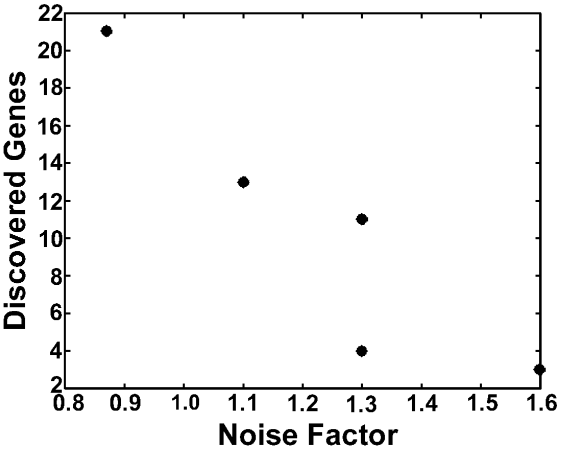

Figure 5. Sensitivity is negatively correlated with the Noise Factor. The *x*-axis plots the Noise Factors of replicate 1.7K meningioma-to-brain datasets having heterogeneous quality. The *y*-axis plots the number of discovered genes.

Table 2. High specificity as compared to others

Comparison	Our algorithm	Standard MIDAS	High stringency MIDAS	Array
Brain versus brain	1/192 000	1347/192 000	932/192 000	19K
Brain versus brain	1/17 280	170/17 280	91/17 280	1.7K

A comparative study of the false discovery rates and specificity of the algorithm, standard MIDAS, and highly stringent MIDAS in analyzing microarray profiling of the same-to-same RNA (brain versus brain). The standard configuration applies: (i) Locfit (LOWESS) normalization (33,34), (ii) SD regularization (35), (iii) low intensity filter, and (iv) flip dye consistency checking (33,35). The high stringency configuration consists of the following operations in order: (i) Locfit (LOWESS) normalization, (ii) iterative linear regression normalization (33), (iii) iterative log mean centering normalization (36), (iv) ratio statistics normalization and confidence interval checking (confidence range at 99%) (37), (v) SD regularization, vi) low-intensity filter, (vii) slice analysis (33,34), and (viii) flip dye consistency checking.

192 000, respectively (Table 2). For the same-to-same 1.7K datasets, the false discovery rates of our algorithm, standard MIDAS, and high stringency MIDAS are 1/17 280, 170/17 280, and 91/17 280, respectively (Table 2).

To evaluate the sensitivity of the algorithm and compare it to MIDAS, we study four independent spike-in 1.7K datasets where 1 ng of Arabidopsis RNA was added to tumor RNA but *not* brain RNA (26). Each dataset includes 64 'genes' of Arabidopsis cDNAs, which are expected to be true positives. Figure 5 shows that sensitivity is dependent not only on the analytical method but also on quality. Thus to dilute the effects of slide-to-slide variations in quality on the sensitivity of the analytical method, we report the best single-experiment sensitivity computed from the four datasets. The algorithm, standard and high stringency MIDAS discover 26/64, 26/64 and 20/64 Arabidopsis genes as differentially expressed, respectively. Receiver operating characteristic (ROC) analysis is the standard approach to evaluate the sensitivity and specificity of diagnostic procedures (30). The algorithm, standard and high stringency MIDAS generate the empiric ROC areas 0.703, 0.698 and 0.654, respectively (Table 3). The accuracy rates are 99.8, 98.8 and 99.2%, respectively.

Table 3. ROC analysis

	Sensitivity	Specificity	Empiric ROC area	Accuracy (%)
Algorithm	26/64	1/17 280	0.703	99.8
Standard MIDAS	26/64	170/17 280	0.698	98.8
High stringency MIDAS	20/64	91/17 280	0.654	99.2

ROC estimates a curve, which describes the inherent tradeoff between sensitivity and specificity of a diagnostic test. The area under the ROC curve is important for evaluating diagnostic procedures because it is the average sensitivity over all possible specificities (38–40). J. Eng (n.d.). ROC analysis: web-based calculator for ROC curves. Retrieved on June 10, 2004 from http://www.rad.jhmi.edu/roc.

DISCUSSION AND CONCLUSIONS

We conclude that the mathematical algorithm optimizes sensitivity without lowering the high specificity of discovery (Figure 3). Furthermore, sensitivity is a function of measurable quality parameters; specifically, sensitivity is negatively correlated with the Noise Factor (Figure 5). For instance, the array whose F1-resistant genes and Noise Factor are shown in Figure 2b and d is more likely to discover smaller differences in gene expression than the datasets shown in Figure 2c and e. These results are similar to Raffelsberger *et al.* who report that quality parameters have an impact on the efficient detection of low level, regulated genes (31). This paper does not address the question of accuracy of fold changes in gene expression levels; however, the results offer a solution to the problem of discovering highly specific states of genetic expression directly from two biological samples.

Fitting smooth curves through the data generates the differential equations that set the rules of F1 and the means to compute CR (see Equation 3). We have chosen a global fit approach and have shown that the curve-fits are statistically significant for small and large datasets acquired in different laboratories ($R^2 > 0.9$). Other strategies for curve fitting that generate differential equations include piece-wise polynomial functions and least-squares approximation. Piece-wise curve fitting partitions the input space into regions, each with its own polynomial equation (spline) whose parameters are estimated by least-square approximation. Each spline is fitted to a small number of data points, while at the same time ensuring that the joints between one part of the curve and another are continuous (32).

The algorithm generates unbiased and unsupervised highly specific, genome-scale expression discovery of states of genetic expression between phenotypes. High specificity facilitates the analysis of genomic comparisons by microarrays because the datasets contain both a predominant majority of true negatives (genes not differentially expressed) as well as a small fraction of true positives (differentially expressed genes). The high degree of certainty facilitates functional discovery of molecular systems by generating testable hypotheses in gene-to-gene and gene-to-protein interactions. Perturbations of the discovered genes may be designed to study the dynamical behavior of the molecular system as a whole and to discover new functions and novel targets for therapeutic interventions (11). The algorithm has numerous applications in biology and medicine.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work is supported by grants from the National Cancer Institute (R01-CA81367 and R29-CA78825) to H.M.F.-S. We are grateful to Roger Rosenberg for a critical review of this paper.

REFERENCES

1. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
2. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H., *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
3. DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
4. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
5. Holstege,F.C., Jennings,E.G., Wyrick,J.J., Lee,T.I., Hengartner,C.J., Green,M.R., Golub,T.R., Lander,E.S. and Young,R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.
6. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
7. Alizadeh,A.A., Eisen,M.B., Davis,E.R, Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse late B-cell lymphomas identified by gene expression profiling. *Nature*, **403**, 503–511.
8. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **8**, 14863–14868.
9. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
10. Gray,N.S., Wodicka,L., Thunnissen,A.M., Norman,T.C., Kwon,S., Espinoza,F.H., Morgan,D.O., Barnes,G., LeClerc,S., Meijer,L. *et al.* (1998) Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science*, **281**, 533–538.
11. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
12. Marton,M.J., DeRisi,J.L., Iyer,V.R., Meyer,M.R., Roberts,C.J., Stoughton,R., Burchard,J., Slade,D., Dai,H., Bassett,D.E.,Jr *et al.* (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.*, **4**, 1293–1301.
13. Perou,C.M., Jeffrey,S.S., Rees,C.A., Eisen,M.B., Ross,D.T., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lee,J.C., Lashkari,D. *et al.* (2000) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.
14. Roberts,C.J., Nelson,B., Marton,M.J., Stoughton,R., Meyer,M.R., Bennett,H.A., He,Y.D., Dai,H., Walker,W.L., Hughes,T.R. *et al.* (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, **287**, 873–880.
15. Hood,L. (2003) Leroy Hood expounds the principles, practice and future of systems biology. *Drug Discov. Today*, **8**, 436–438.
16. Hood,L. (2003) Systems biology: integrating technology, biology, and computation. *Mech. Ageing Dev.*, **124**, 9–16.
17. Ideker,T., Galitski,T. and Hood,L. (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.*, **2**, 343–372.
18. Fathallah-Shaykh,H. (2002) Darts in the dark cure animal but not human brain tumors. *Arch. Neurol.*, **59**, 721–724.
19. Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.
20. Bittner,M., Meltzer,P., Chen,C., Jiang,H., Seftor,E.A., Hendrix,M., Radmacher,M., Simon,R., Yakhini,Z., Ben-Dor,A. *et al.* (2001) Molecular classification of cutaneous melanoma by gene expression profiling. *Nature*, **406**, 536–539.
21. Nielsen,T.O., Hsu,F.D., O'Connell,J.X., Gilks,C.B., Sorensen,P.H., Linn,S., West,R.B., Liu,C.L., Botstein,D., Brown,P.O. *et al.* (2003) Tissue microarray validation of epidermal growth factor receptor and SALL2 in synovial sarcoma with comparison to tumors of similar histology. *Am. J. Pathol.*, **163**, 1449–1456.
22. Tan,P.K., Downey,T.J., Spitznagel,E.L.,Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.
23. Nishizuka,S., Chen,S.T., Gwadry,F.G., Alexander,J., Major,S.M., Scherf,U., Reinhold,W.C., Waltham,M., Charboneau,L., Young,L. *et al.* (2003) Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling. *Cancer Res.*, **63**, 5243–5350.
24. Lee,R.T. (2000) Use of microarrays to identify targets in cardiovascular disease. *Drug News Perspect.*, **13**, 403–406.
25. Fathallah-Shaykh,H.M., He,B., Zhao,L.J., Engelhard,H., Cerullo,L., Lichtor,T., Byrne,R., Munoz,L., Von Roenn,K., Rosseau,G. *et al.* (2003) Genomic expression discovery predicts pathways and opposing functions behind phenotypes. *J. Biol. Chem.*, **278**, 23830–23833.
26. Fathallah-Shaykh,H.M., Rigen,M., Zhao,L.J., Bansal,K., He,B., Engelhard,H., Cerullo,L., Von Roenn,K., Byrne,R., Munoz,L. *et al.* (2002) Mathematical modeling of noise and discovery of genetic expression classes in gliomas. *Oncogene*, **21**, 7164–7174.
27. Rosenzweig,B.A., Pine,P.S., Domon,O.E., Morris,S.M., Chen,J.J. and Sistare,F.D. (2004) Dye bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ. Health Perspect.*, **112**, 480–487.
28. Dudoit,S., Gentleman,R.C. and Quackenbush,J. (2003) Open source software for the analysis of microarray data. *Biotechniques*, Suppl., 45–51.
29. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.
30. Swets,J.A. and Pickett,R.M. (1982) *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, NY.
31. Raffelsberger,W., Dembele,D., Neubauer,M.G., Gottardis,M.M. and Gronemeyer,H. (2003) Quality indicators increase the reliablity of microarray data. *Genomics*, **80**, 385–394.
32. De Boor,C. (2001) *A Practical Guide to Splines*. Springer-Verlag, Berlin, Germany.
33. Quackenbush,J. (2002) Microarray data normalization and transformation. *Nature Genet.*, **32**, Suppl., 496–501.
34. Yang,I.V., Chen,E., Hasseman,J.P., Liang,W., Frank,B.C., Wang,S., Sharov,V., Saeed,A.I., White,J., Li,J. *et al.* (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.*, **3**, research0062.
35. Yang,Y.H., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
36. Causton,H.C., Quackenbush,J. and Brazma,A. (2003) *Microarray Gene Expression Data Analysis: A Beginner's Guide*. Blackwell Publishing, Oxford, pp. 55–56.
37. Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, **2**, 364–374.
38. Swets,J.A. (1979) ROC analysis applied to the evaluation of medical imaging techniques. *Invest. Radiol.*, **14**, 109–121.
39. Metz,C.E. (1986) *Methodology in Radiologic Imaging*. *Invest. Radiol.*, **21**, 720–733.
40. Obuchowski,N.A. (2003) Receiver operating characteristic curves and their use in radiology. *Radiology*, **229**, 3–8.