JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 6, 2001
Mary Ann Liebert, Inc.
Pp. 585–614

Bayesian Estimation of Fold-Changes in the Analysis of Gene Expression: The PFOLD Algorithm

JOACHIM THEILHABER,[1] STEVEN BUSHNELL,[1] AMANDA JACKSON,[2] and RAINER FUCHS[3]

ABSTRACT

A general and detailed noise model for the DNA microarray measurement of gene expression is presented and used to derive a Bayesian estimation scheme for expression ratios, implemented in a program called PFOLD, which provides not only an estimate of the fold-change in gene expression, but also confidence limits for the change and a P-value quantifying the significance of the change. Although the focus is on oligonucleotide microarray technologies, the scheme can also be applied to cDNA based technologies if parameters for the noise model are provided. The model unifies estimation for all signals in that it provides a seamless transition from very low to very high signal-to-noise ratios, an essential feature for current microarray technologies for which the median signal-to-noise ratios are always moderate. The dual use, as decision statistics in a two-dimensional space, of the P-value and the fold-change is shown to be effective in the ubiquitous problem of detecting changing genes against a background of unchanging genes, leading to markedly higher sensitivities, at equal selectivity, than detection and selection based on the fold-change alone, a current practice until now.

Key words: gene expression data analysis, microarray noise modeling, gene expression profiles, Bayesian estimation and detection.

1. INTRODUCTION

CURRENTLY, THERE IS A GROWING FIELD in molecular biology that revolves around the use of DNA microarrays (Fodor *et al*., 1993; Schena *et al*., 1995; Lockhart *et al*., 1996; Wodicka *et al*., 1997; Eisen *et al*., 1998; Cho *et al*., 1998; Iyer *et al*., 1999; Chu *et al*., 1999; DeRisi *et al*., 1997), a technology which makes possible the measurement of gene expression in biological systems for thousands of genes simultaneously. A ubiquitous problem in the analysis of expression data is the estimation of the fold-change in the expression level of a gene in one context relative to its expression in another context, typically to infer context-dependent, differential regulation. Given two raw measurements, the simplest approach has been to take the arithmetic ratio of the values as an estimate of the fold-change. While for very strong signals this leads to a meaningful estimate of the fold-change in the underlying mRNA concentrations, for weaker signals the results are much more ambiguous because of contamination by noise. Furthermore,

[1]Aventis Pharmaceuticals, Cambridge Genomics Center, 26 Landsdowne Street, Cambridge, MA 02139.
[2]CuraGen Corporation, 555 Long Wharf Drive, New Haven, CT 06511.
[3]Biogen Inc., 14 Cambridge Center, Cambridge, MA 02142.

for technologies based on differential signal intensities (e.g., Affymetrix [Lockhart *et al*., 1996]), the values assigned to expression levels can even be negative, leading to the awkward situation of negative or undefined expression ratios.

A work around for the problem of weak or negative expression levels is to "floor" the values at some threshold, typically thought to reflect the level of noise in the experiment, and below which values are no longer meaningful. While this approach is reasonable, it has the drawback of being heuristic, and in itself neither provides confidence limits for the estimate nor a measure of significance for a change.

In the following, we circumvent heuristics by using a simple deductive approach grounded in a Bayesian framework (Van Trees, 1978; Duda and Hart, 1973) and on an underlying model of the noise inherent in the microarray measurements. Rather than immediately seek a point estimate of the fold-change, we first derive a mathematical formula for the a posteriori distribution of *all* the fold-changes which can be inferred from the given measurements. From this distribution, we then obtain several statistics, including an estimator for the fold change, confidence limits for the fold change at any given confidence level, and a P-value for assessing the statistical significance of the change. In particular, we can assign fold-change estimates and confidence limits even to signal pairs where both signals are zero or negative, without resorting to heuristic thresholds. Indeed, the mathematical framework unifies estimation for all signals.

The computer implementation of this scheme has been called "PFOLD." For any pair of intensities (and associated measures of noise), PFOLD provides a two-dimensional representation of data based on both fold-change and P-value. This representation can be very useful for the basic task of discriminating the genes with significant change from the usually much larger background of unchanging genes. For instance, at a given level of selectivity, using the P-value as a statistic for discriminating changing versus non-changing genes yields markedly higher sensitivity than using the fold-change alone, the latter being a current practice until now. Note that, to some extent, the approach embodied in PFOLD, with intensities, is analogous to that of the sequence alignment tool BLAST (Altschul *et al*., 1990), where both a score and a P-value are provided for the two sequences being matched.

In what follows, we first present the noise model for microarray measurements that is the underpinning of the entire approach. We then give the derivation of the mathematical model underlying PFOLD and follow with validation work based both on Monte Carlo simulations and cRNA spiking experiments.

1.1. Related work

There are similarities between the model presented here and the one derived in the important work by Chen *et al*. (1997), which also addresses the problem of quantifying expression ratios. However, there are also basic differences between the two approaches: the noise model assumed by Chen *et al*. only accounts for a coefficient of variation as source of the noise (cf. Section 2.5), while PFOLD includes background and cross-hybridization terms as well; perhaps more fundamentally, the approach adopted by Chen *et al*. is not Bayesian and seeks to quantify the significance of change by assuming that the bulk of the ratio distribution represents the null hypothesis of no change against which a P-value is estimated. As such, the authors do not obtain a point estimate of the actual fold-change or a confidence interval for that estimate (although a confidence interval is obtained for the region of validity of the no-change hypothesis).

2. FORMULATION OF THE NOISE MODEL FOR MICROARRAY MEASUREMENTS

2.1. Microarray technologies

To give some of the context of the noise model, we very briefly review how DNA microarrays are used to measure mRNA transcript abundance and hence quantify the level of expression of a gene (Fodor *et al*., 1993; Schena *et al*., 1995; Lockhart *et al*., 1996; Nature Genetics, 1999). The microarray itself consists of a glass slide or "chip" on which up to several thousand "target" features have been created,[1] each feature

[1] In this paper we adhere to the convention that the nucleic acid strands immobilized on the chip are called the "targets," and the nucleic acid strands in the sample are called the "probes," with the entire sample at times simply referred to as "the probe."

consisting of a large number of identical DNA strands, which are complementary to a specific transcript. The experimental procedure begins with the processing of a sample of total RNA, obtained from the biological source of interest, into a "probe" containing labeled and fragmented cDNA or cRNA[2] copies of the mRNA transcripts, which are then hybridized to the targets in the features on the microarray. Because the probe is labeled with a fluorescent dye, the amount of material that hybridizes to a specific feature is measured when the entire chip is scanned by a laser, and the fluorescing intensities are captured into an image, providing the data for the downstream analysis that is discussed here. Beyond this general sketch of the procedure, the measurement process has details which are specific to each of the two major microarray technologies currently in use, based on either cDNA (Schena *et al.*, 1995) or on oligonucleotides targets (Lockhart *et al.*, 1996). In what follows, we focus on the oligonucleotide microarrays, although with an appropriate noise model, the methodology applies to cDNA arrays as well.

It should be emphasized that with the current microarray technologies, obtaining an absolute measurement of concentration is difficult, because the large sequence-dependent variation in hybridization affinity of cRNA probes to their DNA targets introduces a proportionality constant between the intensity and the concentration of the transcript in solution that is generally unknown (unless determined by independent and labor-intensive titration experiments). The emphasis in the field is therefore to quantify relative, rather than absolute levels, of expression in the form of a "fold-change" which is the ratio of expression levels measured in two different experiments.

2.2. *The general noise model*

For a given gene, the model of noise is based on the expression

$$x = Cn_t + \epsilon_b + \epsilon_c, \tag{1}$$

where x denotes the measured intensity, n_t is the physical concentration of the mRNA transcripts in solution (it could be expressed, for instance, as a molarity in pM), C is a proportionality constant specific to the gene, and ϵ_b and ϵ_c are noise terms accounting for background and cross-hybridization effects, respectively. In this model, C, ϵ_b, and ϵ_c are all considered random variables.

Equation (1) automatically embodies linearity, because it is assumed that for a given realization of C, the signal part of the intensity, Cn_t, is simply proportional to the concentration n_t. While we have indications of the breakdown of simple proportionality at high concentrations, we have found that the effect is moderate and affects only a small proportion of the total gene population in a typical experiment. For instance, for the scans of Affymetrix chips considered here, saturation is seen as a flattening of the curve through a scatter plot when a very bright chip, as measured by median chip intensity, is compared to a very dim chip; the onset of flattening typically occurs at bright chip intensities $x \sim 13,000$, with maximum intensities on either chip rarely exceeding $x \sim 24,000$. However, in a population of 224 scans of Affymetrix Mu19KsubA chips, even for the brightest chip, less than 10% of the genes called *present*[3] had intensities in the range $x \gtrsim 13{,}000$, and this fraction fell to about 3% for the chips with median brightness. Furthermore, this form of signal saturation occurs precisely for genes with a very high signal-to-noise ratio, for which a detailed noise analysis is less crucial (although some form of nonlinear rescaling, outside the scope of this paper, may be necessary for proper processing of the data). We shall therefore assume a linear model in all that follows.

Equation (1) is very general in that it assumes that the intensity x can be obtained in a number of different ways, depending on the microarray technology and the image-processing methods used. For the Affymetrix oligonucleotide (Lockhart *et al.*, 1996), typically 40 nonredundant features are assigned to each gene. The features are organized in pairs, consisting of a "perfect match" feature and a paired "mismatch"

[2]For Affymetrix chips, an additional amplification step using in vitro transcription is applied. This results in a probe consisting of cRNA, rather than cDNA copies of the original transcripts.

[3]For the oligonucleotide array data presented in this paper, all of the initial image processing, the extraction of intensities in the form of "average differences," and the assignment of *present* or *absent* calls to individual genes on the chip, were done using the GeneChip™ software built by Affymetrix, Inc., 3380 Central Expressway Santa Clara, California. We have simplified the *absent/present* decision process by declaring so-called *marginal* calls to be equivalent to *present* calls.

feature, for which the oligonucleotides are either perfect matches to the targeted sequence or contain a single mismatch at the central position, respectively. A trimmed mean of the 20 differential intensities between pair members is then taken, and this defines a single, effective intensity x (the so-called average difference), which is then assigned to the gene of interest. It should be noted that the intensity x can be negative because it is based on several differential measurements. This will occur, for instance, when a large number of the mismatch features are brighter than the corresponding perfect match features, typically because of strong cross-hybridization by a component of the probe which has a sequence closely related to that of the gene.

There are global sources of variation which affect all features on a chip equally and which result in overall "brightness" or "dimness" of the scan image; some of the sources are chip-to-chip variation in the efficiency of feature synthesis when the chips are constructed, or variation in overall hybridization conditions or in total probe concentration from one experiment to the next. These global effects can be accounted for by rescaling all of the intensities in a given scan by multiplication by a single rescaling factor, chosen so that after rescaling the mean or median brightness of all the scans in a given data set are adjusted to the same value. In what follows, we assume that such a global rescaling has been done, so that all of the sources of variation to be modeled by Equation (1) are local in nature.

2.3. *Sources of noise*

In Equation (1) we write the proportionality constant as a constant plus variable part,

$$C = C_0 + \delta C \tag{2}$$

where C_0 is the mean of C and $\delta C \equiv C - C_0$. If we define

$$\epsilon_{cv} = \delta C n_t, \tag{3}$$

then Equation (1) can be rewritten as

$$x = C_0 n_t + \epsilon, \tag{4}$$

where now the composite noise term ϵ is the only random variable in the equation and ϵ is the sum of three terms

$$\epsilon = \epsilon_{cv} + \epsilon_b + \epsilon_c, \tag{5}$$

each arising from a distinct physical mechanism. In what follows, we assume that all the noise terms in Equation (5) are normally distributed ("Gaussian"), with zero mean (Feller, 1966, p. 45). This assumption is an approximation; we have found that in most instances the distribution of the noise is composite, with exponential tails attached to an approximately Gaussian, central distribution. However, we believe that the impact of non-Gaussian behavior on the resulting distribution of expression ratios is limited (cf. Section 4.1), and we pursue the Gaussian model in all that follows.

The first contribution in Equation (5), ϵ_{cv}, is due to the coefficient of variation (cv) of the proportionality constant C and results in a noise term that is proportional to the signal. Using Equation (3) we write the standard deviation σ_{cv} of ϵ_{cv} as

$$\sigma_{cv} = \alpha C_0 n_t, \tag{6}$$

where α is the coefficient of variation of C,

$$\alpha = \frac{\sigma_{\delta C}}{C_0}. \tag{7}$$

The second term in Equation (5), ϵ_b, accounts for background contributions to the noise. These contributions arise, for instance, from fluorescent dye that is dispersed on the chip but not part of hybridized complexes and from optical noise arising in the actual imaging process.

Table 1. Typical Values for the Noise Parameters for Mu19KsubA Affymetrix Chips[a]

x_m	σ_b	σ_c	σ_{bc}	σ_{cv}^m	σ_ϵ^m	x_m/σ_ϵ^m
685	37 ± 13	256 ± 48	259 ± 45	171	311 ± 40	2.2 ± 0.25

[a]Seven samples derived from mouse C3H10T1/2 cells lines were hybridized to seven chips from the same lot; the samples were biologically similar but not strictly equivalent, as they correspond to treatments with different growth factors; the preparation protocol used antibody-enhanced fluorescence. The intensities on each chip were rescaled so that the median intensity x_m of *present* genes is constant across all chips and equal to the average x_m of the raw median intensities, $x_m = 685$. The values given in the table are the averages across the seven scans, with standard deviations indicated by the $\pm$ terms; σ_b = standard deviation of background noise; σ_c = standard deviation of cross-hybridization noise; σ_{bc} = standard deviation of combined background and cross-hybridization noise; x_m = median intensity for all the *present* genes on a chip; σ_{cv}^m = standard deviation of noise due to the coefficient of variation α, computed for median intensity and for $\alpha = 0.25$; σ_ϵ^m = standard deviation of the sum of all noise terms, at median intensity; x_m/σ_ϵ^m = median signal-to-noise ratio.

The third term in Equation (5), ϵ_c, models all cross-hybridization effects. It is conceptually a superposition of effects,

$$\epsilon_c = \sum_{s \neq \text{target}} C_s n_s, \tag{8}$$

where the sum is extended over *all* transcript species present in the sample, excepting the gene of interest, and where n_s is the concentration of transcript s, and C_s the measure of its affinity to the feature under consideration. In what follows, we do not attempt any detailed modeling of the terms in Equation (8), but rather evaluate the variance of ϵ_c, lumped with that of ϵ_b, in a direct, semi-phenomenological manner described in the next section.

2.4. Estimation of the noise parameters

The actual estimation of the characteristics of the noise terms in Equation (5) is technology dependent and proceeds as follows for the Affymetrix oligonucleotide arrays.

An estimate of the standard deviation σ_b of the background noise ϵ_b can be directly obtained by collecting the intensities of groups of features on the chip which have no oligonucleotides and computing a sample standard deviation for these intensities.[4] A typical value is indicated in Table 1. However, for the method that follows, it is convenient to group background and cross-hybridization terms into a single noise term ϵ_{bc},

$$\epsilon_{bc} = \epsilon_b + \epsilon_c. \tag{9}$$

In computing the standard deviation σ_{bc} of ϵ_{bc}, our intent is not to establish σ_{bc} on a gene-by-gene basis, but rather to obtain a value representative of all the genes on a chip, a somewhat easier task. To this end, we have adopted an "on-the-fly" method which proceeds as follows: for a single, given scan, of all the genes represented on the chip we separately consider those labeled *present* (P) and those labeled *absent* (A) by the Affymetrix GeneChip™ decision algorithm.[5] This results in the two distinct intensity distributions, shown in Figs. 1a,b. For the *absent* genes, we can set $n_t \approx 0$ in Equation (1), so that for this population $x \approx \epsilon_{bc}$ (this is the reason for grouping the two noise terms together). As a consequence, the standard deviation σ_{bc}^A of ϵ_{bc} for the *absent* genes is given by

$$\sigma_{bc}^A \equiv \underset{A}{\text{stdev}}(x), \tag{10}$$

[4]For instance, we have used strips of nine features at the borders of the four "landmarks" which define the corners of the Mu19K Affymetrix chips.

[5]In Lockhart *et al.* (1996), see the section "Quantitative analysis of hybridization patterns and intensities" under "Experimental Protocol." See also footnote 3, page 587.

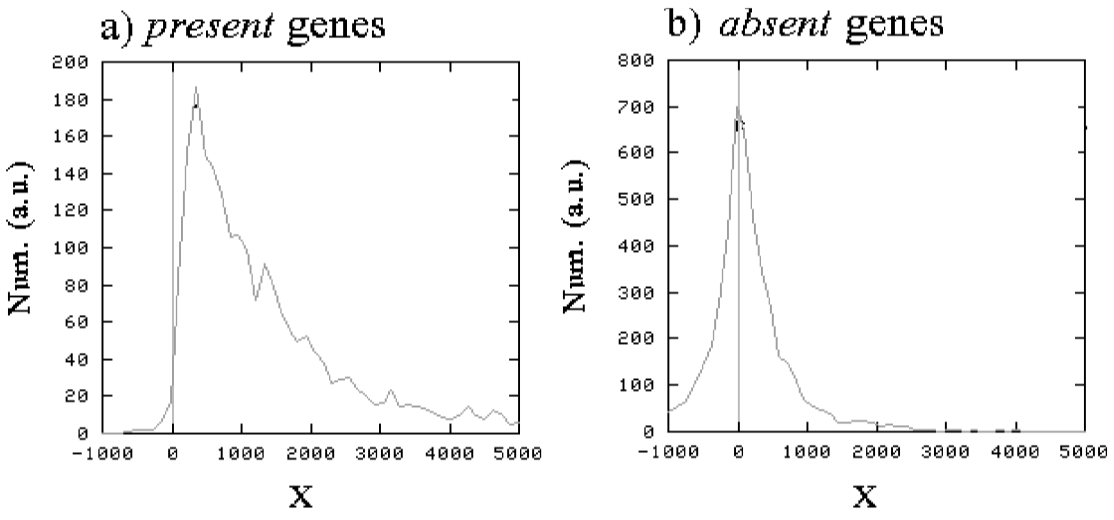

FIG. 1. Distributions of the intensities for a murine intestine tissue sample hybridized to an Affymetrix Mu19KsubA chip. Out of the total of 7,045 genes featured on the chip, 2,340 were signaled *present* and 4,705 signaled *absent* by the Genechip decision algorithm. The resulting intensities distributions are separately shown for the *present* genes in a) (median intensity = 1,200), and for the *absent* genes in b) (median intensity ≈ 0). The histograms were constructed using 50 bins of equal intensity intervals. The standard deviation of the intensities of the *absent* genes, which here is $\sigma_{bc} = 560$, is the measure of the combined background and cross-hybridization noise. Note that the distribution of the *present* genes is approximately log-normal.

where the "stdev" operator on the right-hand side computes the standard deviation of the *absent* gene intensities (i.e., of the distribution shown in Fig. 1b). The final step in computing σ_{bc} is to assume that the result obtained in Equation (10) for the *absent* genes applies to all the genes on the chip, so that we make the assignment

$$\sigma_{bc} = \sigma_{bc}^{A}. \tag{11}$$

This assumption appears reasonable because biological samples hybridized to high density Affymetrix chips, which have features for about 7,000 genes, typically result in about 4,500 *absent* calls of the total of 7,000, and this large number provides a broad statistical sampling of cross-hybridization. In effect, we are saying that the intensity distribution of the *absent* genes (Fig. 1b) is implicitly present as a random, additive term in the intensity distribution of the *present* genes (Fig. 1a) as well.

The Affymetrix GeneChip™ decision algorithm is proprietary, but is known to depend on a voting procedure which uses sets of scores defined for each of the 20 perfect match and mismatch pairs of features. A conceptual, "dummy" outline of the process might be as follows: for each feature pair, a plus score is assigned if the perfect match intensity exceeds the mismatch intensity by a threshold proportional to σ_b, and conversely, a minus score is assigned if the mismatch intensity exceeds the perfect match intensity by the same threshold. A *present* call is then assigned to the gene if a majority of feature pairs have positive scores, and an *absent* call is assigned to the gene otherwise. While the actual GeneChip™ decision process is different, its basis is nonetheless a geometric argument that requires consistency of the hybridization pattern across many feature pairs (perfect matches consistently brighter than mismatches), and which does not require a priori knowledge of the cross-hybridization noise itself. This avoids a circular argument in the use of Equation (10) to derive an estimate of the standard deviation of the combined background and cross-hybridization noises. Finally, note that in using Equation (10) we are ignoring the effects of misclassification errors in the GeneChip™ decision process itself, and thus Equation (10) should certainly be regarded as no better than an approximation.

To estimate the coefficient of variation α, Equation (7), we analyzed two data sets, each capturing aspects of technological variation of Affymetrix chips. The first data set was obtained by hybridizing probes derived from a single biological sample (mouse C2C12 cell line) to seven Mu19KsubA chips all derived from the

same chip lot.[6] A total of three separate probes, corresponding to three separate preparations starting from the initial sample of total RNA, were generated and hybridized to 1, 3, and 3 chips, respectively. After scanning, the resulting numerical data sets were normalized to each other, and all genes with a majority of *present* calls across the seven scans (2,097 genes out of 7,045) were then ranked according to average intensity. The average value for the standard deviation of the combined cross-hybridization and background noise was $\sigma_{bc} = 268$. For the i-th gene in the list, a sample coefficient of variation α_i was then computed by dividing the standard deviation of intensities across the seven samples by the average intensity. A final, global estimate of the coefficient of variation was obtained by averaging α_i for all genes in the top quartile of the ranked list (524 samples, average intensity range $1{,}560 \leq x \leq 20{,}300$), resulting in $\alpha = 0.16$ (1 ± 0.023). Note that by restricting the intensities to the top quartile, we are insuring that $x > \sigma_{bc}/\alpha$ for nearly all genes and hence that the term ϵ_{cv} dominates in Equation (5). This justifies the estimation procedure, where we treat all variation as arising from only this term.

A second data set was used to investigate the effect of chip lot variability. Identical probes, generated from a single preparation of a biological sample (again, mouse C2C12 cell line) were hybridized to two Mu19KsubA chips from different chip lots. After scanning, the two resulting numerical data sets were normalized to each other, and all genes with at least one *present* call across the two scans (2,044 genes out of 7,045) were then ranked according to average intensity. For the i-the gene, an estimate of α^2 was obtained by using $\hat{\alpha}^2 \equiv ((x_2 - x_1)/(x_2 + x_1))^2/2$, where x_1 and x_2 refer to the intensities in scans 1 and 2, respectively.[7] A global estimate of α^2 was then obtained by averaging $\hat{\alpha}^2$ over the top quartile of genes (511 samples, average intensity range $759 \leq x \leq 19{,}127$, average $\sigma_{bc} = 149$). The resulting estimate of α, $\alpha = 0.23(1 \pm 0.037)$, is somewhat larger than the previous one: this suggests that the variation induced by using chips from different lots is greater than that induced by using different probes on chips from the same lot.

We have not extended the analysis of the previous paragraphs to a systematic study of all possible factors of variation. However, on the basis of these results, it appears that $\alpha \approx 0.25$ is a reasonable (and possibly slightly conservative) estimate of the maximum coefficient of variation expected when different probe preparations and different chip lots are simultaneously used in the hybridization process, which is typical when large data sets are generated.

Finally, note that when multiple replicate scans are available, it is preferable to estimate σ_{cv} directly from the available data, "on-the-fly," by the standard deviation of the replicates. The latter procedure, however, has not been used in the data analyses that are presented below, as replicates were not involved.

Typical magnitudes of the various noise terms discussed above are displayed in Table 1, specifically for Affymetrix Mu19KsubA chips. Note that because of the inherent variation of overall chip brightness, which makes rescaling necessary, the emphasis should be on the relative rather than absolute magnitude of these terms. Note also that, according to the model, the median signal-to-noise ratio is low, $x_m/\sigma_\epsilon^m \approx 2.2$.

2.5. *Combining the noise terms*

Because we do not seek to quantify absolute concentrations, but only expression ratios, for each gene we can redefine the concentration variable n_t in Equation (1) by defining the normalized variable $n = C_0 n_t$, which allows us to avoid writing C_0 in all the equations that follow. The noise model thus reduces to the slightly simpler equation

$$x = n + \epsilon, \tag{12}$$

where as in Equation (5)

$$\epsilon = \epsilon_{cv} + \epsilon_b + \epsilon_c, \tag{13}$$

but where now Equation (6) is written

$$\sigma_{cv} = \alpha n, \tag{14}$$

and where σ_{bc} is computed as before (Equation (11)).

[6] A chip lot refers to a set of chips derived from the same wafer during the fabrication process.

[7] Assuming Gaussian noise, the estimator $\hat{\alpha}^2$ for α^2 has bias $O(\alpha^4)$.

To compute the standard deviation σ_ϵ of the complete noise term ϵ, we use the fact that ϵ_{cv} and $\epsilon_{bc} = \epsilon_b + \epsilon_c$ are uncorrelated, so that

$$\sigma_\epsilon^2 = \mathrm{var}(\epsilon_{cv}) + \mathrm{var}(\epsilon_{bc}) = (\alpha n)^2 + \sigma_{bc}^2. \tag{15}$$

The coefficient of variation α used in Equation (15) is very similar to the coefficient of variation c defined by Chen *et al.* (1997). In the present model, however, other noise terms intervene, so that the total variance of the noise may be large even as $n \to 0$, in contrast to the assumptions of Chen *et al.*

Equation (15) requires a priori knowledge of n to determine σ_ϵ. However, if we write $n \approx x$ on the right-hand side of Equation (15) (the simplest approximation suggested by Equation (12)), we obtain an estimator $\hat{\sigma}_\epsilon$ for σ_ϵ,

$$\hat{\sigma}_\epsilon^2 = (\alpha x)^2 + \sigma_{bc}^2, \tag{16}$$

so that we do not need to know the underlying concentration beforehand to estimate the cv contribution to the total variance of the noise. In Appendix A, by deriving Taylor expansions in α for the bias and variance of $\hat{\sigma}_\epsilon$, we show that provided α is moderately small (say $\alpha \lesssim 0.25$), $\hat{\sigma}_\epsilon$ is always weakly biased, and that its standard deviation is never more than $\alpha\sigma_\epsilon$: thus, the fractional error inherent in using $\hat{\sigma}_\epsilon$ is given by α, a limitation we deem acceptable.

2.6. *Scan-to-scan noise correlation coefficient*

Equation (16) is an expression for the standard deviation of the noise for each gene across the entire population on a chip, but says nothing about correlations between different scans. For a given gene, consider two intensity measurements x_1 and x_2 obtained from two different samples, with noise terms labeled ϵ_1 and ϵ_2, respectively. The correlation coefficient of the two noise terms is defined as (assuming zero means) (Feller, 1966, p. 67)

$$\rho = \frac{< \epsilon_1 \epsilon_2 >}{\sigma_1 \sigma_2}, \tag{17}$$

where σ_1 and σ_2 are the standard deviations of ϵ_1 and ϵ_2, respectively. Note that in general we expect $\rho \neq 0$, because, for a given gene, the cross-hybridization component of the noise, ϵ_c, arises from effects which tend to persist in different contexts (Equation 8).

To evaluate the correlation coefficient ρ for the Affymetrix technology, we examined 50 pairs of scans sampled at random with no replacement from a population of 100 scans (Mu11KsubA chips) covering a large diversity of samples. Intensity pairs were accumulated for all genes and all scan pairs where the gene is signaled "absent" in both scans. The numerical evaluation of the correlation coefficient based on the resulting data yielded the estimate $\rho \approx 0.7$.

3. A POSTERIORI DISTRIBUTION OF CONCENTRATIONS

While Equation (12) gives the intensity measurement x in terms of the concentration n, it is exactly its inverse that we wish to obtain, namely, the concentration as a function of the measurement. We can formulate this in probabilistic terms by writing the Bayes Theorem (Drake, 1967, p. 26) (Van Trees, 1978) for the variables n and x,

$$\mathrm{P}(n|x) = \frac{\mathrm{P}(x|n)\mathrm{P}(n)}{\mathrm{P}(x)}. \tag{18}$$

In Equation (18), $\mathrm{P}(x|n)$ is the conditional probability distribution function (pdf) for x, conditional on n, $\mathrm{P}(n)$ is the a priori distribution of n (reflecting our state of knowledge of n before the measurement is taken), and $\mathrm{P}(x)$, the pdf for x, functions as a normalization term. From Equation (12) with the assumption of Gaussian noise, we can immediately write

$$\mathrm{P}(x|n) = \frac{1}{(2\pi\sigma_\epsilon^2)^{1/2}} \exp\left(-(x-n)^2/2\sigma_\epsilon^2\right), \tag{19}$$

where $\sigma_\epsilon = \sigma_\epsilon(n)$, is given by Equation (15).

For the distribution $P(n)$, as prior knowledge we use the fact that the concentration is necessarily nonnegative,

$$P(n) = \begin{cases} 0, & n < 0, \\ \mu e^{-\mu n}, & n \geq 0, \end{cases} \tag{20}$$

where we shall take the limit $\mu \to 0$ very shortly (this is just a device to get a step-function distribution in the limit $\mu \to 0$, while keeping $P(n)$ integrable at all times).

The step-function distribution which obtains in the $\mu \to 0$ limit of Equation (20) may seem a trivially simple choice for a prior distribution. In our minds, however, it has the extremely important feature of solving the "division by zero" problem, that is, enabling one to assign, by way of the Bayesian formulation, an expression ratio even when one or both intensities are zero or negative. Furthermore, we have been hesitant to assign a more detailed prior distribution because of uncertainties in the actual distribution of intensities of expressed genes, especially in the low-concentration limit. For instance, the distribution of *present* genes illustrated in Fig. 1a is approximately log-normal, and such a distribution, with fitted parameters, could be used in place of the step function prior. However, this distribution is based on using the GeneChip^{TM} decision algorithm for defining the *present* genes, and while we believe that the algorithm is adequate for approximating the gross characteristics of the distributions of *present* or *absent* genes, it is less clear that it accurately models the fine-scale features of these distributions, especially at low intensities. In view of these uncertainties, Equation (20) represents a conservative choice of prior.

In Equation (18), $P(x)$ is obtained by integration,

$$P(x) = \int_{-\infty}^{\infty} dn P(n) P(x|n). \tag{21}$$

Using the explicit formula in Equation (20), we can write Equation (18) as

$$P(n|x) = \frac{P(x|n)\mu e^{-\mu n}}{\int_0^{\infty} dn' P(x|n')\mu e^{-\mu n'}}, \qquad n \geq 0, \tag{22}$$

where $P(n|x) = 0$ for $n < 0$. In Equation (22), the constant coefficient μ factors out from both numerator and denominator, and in the limit $\mu \to 0$ the equation becomes

$$P(n|x) = \frac{P(x|n)}{\hat{P}(x)}, \qquad n \geq 0, \tag{23}$$

where $P(x|n)$ is given by Equation (19) and where the denominator is now

$$\hat{P}(x) = \int_0^{\infty} dn P(x|n). \tag{24}$$

Equation (24) can be readily evaluated using error functions. Rather than directly explore the consequences of Equation (23) on estimation of concentrations, we use it below to quantify the distribution of fold changes.

4. A POSTERIORI DISTRIBUTION OF FOLD CHANGES

For a given gene, let us assume we wish to evaluate the fold-change in expression between two experiments, 1 and 2. We assume that the mRNA concentrations in the experiments are n_1 and n_2, respectively, and write for the corresponding observed intensities, x_1 and x_2,

$$x_1 = n_1 + \epsilon_1, \tag{25}$$

$$x_2 = n_2 + \epsilon_2. \tag{26}$$

The fold-change R of the concentration in experiment 2 relative to experiment 1 is given by the ratio

$$R = \frac{n_2}{n_1}. \tag{27}$$

While in Equation (27) we do have direct access to n_1 and n_2, we can immediately formulate the estimation of R in Bayesian terms by writing the a posteriori distribution of R as

$$f_R(R|x_1, x_2) = \int_0^\infty dn_1 \int_0^\infty dn_2 \delta\left(\frac{n_2}{n_1} - R\right) \mathrm{P}(n_1|x_1)\mathrm{P}(n_2|x_2), \tag{28}$$

where x_1 and x_2 are the intensity measurements in experiments 1 and 2, respectively, where $\delta(\ldots)$ refers to the Dirac delta function, and where $\mathrm{P}(n|x)$ is given in Equation (23) above.

In formulating Equation (28) and in most of what follows, we make the simplifying assumption that the noise terms ϵ_1 and ϵ_2 are uncorrelated, so that $\rho = 0$ in Equation (17). In Appendix B, we lift this restriction and generalize the derivation to a correlated noise model, with $\rho \neq 0$, to be fully implemented in a future version of PFOLD. In fact, a correlated noise model is more realistic insofar as the cross-hybridization component of the noise, ϵ_c, arises from semi-deterministic effects which tend to persist in different contexts (thus $\rho \approx 0.7$ for Affymetrix chips, Section 2.6). However we wish to emphasize that the present implementation of PFOLD, with $\rho = 0$, is not unreasonable but merely results in a statistically more conservative estimate of fold-change.

Performing the integration indicated in Equation (28) is a very straightforward if slightly tedious task. We obtain the distribution function for R in the form (dropping the explicit dependence on x_1 and x_2 in $f_R(R|x_1, x_2)$),

$$f_R(R) = \frac{C(x_1)C(x_2)}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x_1^2(R - R_0)^2}{2(\sigma_2^2 + R^2\sigma_1^2)}\right) I(x_1, x_2), \tag{29}$$

where $\sigma_i^2 = \sigma_\epsilon^2(x_i)$, $i = 1, 2$, with $\sigma_\epsilon(x)$ now given by Equation (16), $R_0 \equiv x_2/x_1$, and with the normalization term

$$C(x) = \frac{2}{1 + \mathrm{erf}(x/\sqrt{2}\sigma_\epsilon(x))}, \tag{30}$$

where erf is the error function (Abramowitz and Stegun, 1972, p. 297). $I(x_1, x_2)$ is defined by

$$I = \sigma_{12}^2 \exp\left(-\frac{a_{12}^2}{2\sigma_{12}^2}\right) + a_{12}(2\pi\sigma_{12}^2)^{1/2}\frac{1}{2}\left(1 + \mathrm{erf}(a_{12}/\sqrt{2}\sigma_{12})\right), \tag{31}$$

where

$$\frac{1}{\sigma_{12}^2} = \frac{1}{\sigma_1^2} + \frac{R^2}{\sigma_2^2} \tag{32}$$

$$a_{12} = \left(\frac{x_1}{\sigma_1^2} + \frac{Rx_2}{\sigma_2^2}\right) \Big/ \left(\frac{1}{\sigma_1^2} + \frac{R^2}{\sigma_2^2}\right) \tag{33}$$

Though complex-looking, Eq. (29) has two simple limits.

Case 1: high concentrations. If in both experiments the RNA concentrations are large compared to the standard deviation of the noise, with consequence $x_i >> \sigma_\epsilon(x_i)$, $i = 1, 2$, we find that R has an approximately normal distribution, which for $x_2 \geq x_1$ has the form

$$f_R(R) \approx \frac{1}{(2\pi\sigma_R^2)^{1/2}} \exp\left(-\frac{(R - R_0)^2}{2\sigma_R^2}\right). \tag{34}$$

In this limit, the median of R is just the ratio of the measurements,

$$< R > = R_0 = \frac{x_2}{x_1}. \tag{35}$$

The variance σ_R^2 of R is given by

$$\sigma_R^2 = \frac{\sigma_2^2 + x_2^2\sigma_1^2/x_1^2}{x_1^2}. \tag{36}$$

Using Equation (16) in the limit $\alpha x \gg \sigma_{bc}$, in turn we find a simple approximation for the standard deviation of R,

$$\sigma_R = \sqrt{2}\alpha R_0. \tag{37}$$

Thus, in the high-concentration limit, the standard deviation of the fold-change relative to its median is given by a constant,

$$\frac{\sigma_R}{R_0} = \sqrt{2}\alpha. \tag{38}$$

Equation (37) indicates that no matter how large the signals, the fold-change will retain an irreducible coefficient of variation of order $\sqrt{2}\alpha$ ($\approx \pm 35\%$ for $\alpha = 0.25$).

Case 2: very low concentrations. If in both experiments the RNA concentrations are so low that $x_i << \sigma_\epsilon(x_i)$, $i = 1, 2$, then the distribution takes on the "universal" form,

$$f_R(R) \approx \frac{1}{\pi}\frac{1}{1+R^2}. \tag{39}$$

where for simplicity we assume $\sigma_1 = \sigma_2$. In this limit, the distribution of R is completely independent of the concentrations, the influence of which has been overwhelmed by the noise.

Equation (39) defines a so-called Cauchy distribution (Feller, 1966, p. 50), which does not have a finite mean because of its $1/R^2$ functional dependence for large R. In fact, the original distribution from which Equation (39) was derived, Equation (29), also has a $1/R^2$ dependence for large R, even in the quasi-normal limit of Equation (34), so that $f_R(R)$ does not have a well-defined mean under any circumstances. On the other hand, the median of Equations (29) or (39) is always well-defined, and we shall use it as an estimator for R in what follows.

The cumulative distribution function corresponding to Equation (39) is given by

$$P(R \le R') = \frac{2}{\pi}\tan^{-1} R'. \tag{40}$$

For instance, the 80% confidence limits for R are [0.16, 6.3], indicating that the distribution in Equation (39) is very broad. Note that these large bounds on R are obtained even when the intensities are equal, $x_2 = x_1$, provided $x_{1,2} \ll \sigma_{1,2}$.

The derivation of Equation (29) for the fold-change distribution $f_R(R)$ can be understood in much simpler geometrical terms, based on the following construction, illustrated in Fig. 2: for each pair of intensities (x_1, x_2), draw a box in the plane about the point (x_1, x_2) with extents $\pm\sigma_\epsilon$ in each dimension (Fig. 2a). This defines the range of concentrations (n_1, n_2) which are compatible with the observed data (x_1, x_2). One then draws lines from the origin to all the points in the box, creating the picture of a fan shown in Fig. 2a. The collection of slopes R of all the lines in the fan is the set of all fold-changes compatible with the observed intensities (x_1, x_2). Apart from the simplification of sharp confidence limits imposed by a rectangular box, the distribution of the slopes in the fan in Fig. 2a is thus $f_R(R)$, Equation (29). In the low signal-to-noise limit, Fig. 2b, a slight modification is imposed on the construction: the areas of the box that would correspond to negative concentrations are simply omitted from the construction of the fan (the white area in Fig. 2b). This procedure corresponds to imposing the nonnegative prior probability

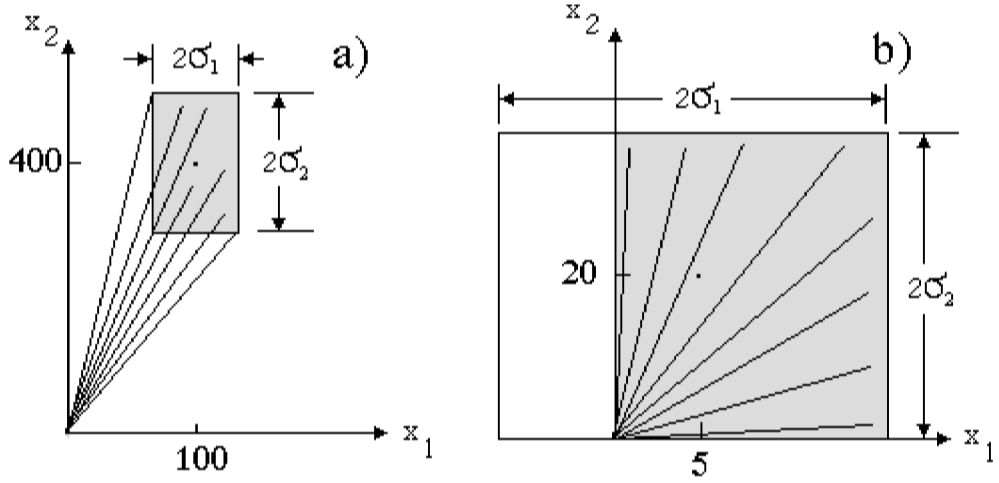

FIG. 2. Qualitative illustration of the derivation of Eq. (29), explaining the behavior of the distributions shown in Fig. 3. For each pair of signals (x_1, x_2), draw a box with extents $\pm\sigma_\epsilon$ about the point in the plane. Draw all lines from the origin to points in the box: the distribution of slopes of these lines is the a posteriori distribution of fold-changes, Eq. (29); a) construction for large signal-to-noise ratios, with the intensity pair (100, 400); b) construction for low signal-to-noise ratio, (5, 20); note that the part of the box which lies on the negative axis is excluded from the construction and this constraint is equivalent to stating the Bayes Theorem with the nonnegative prior for the concentrations that is used in the derivation of Eq. (29).

distribution on the concentrations (Equation (20)), and makes the geometrical construction described here Bayesian.

To clarify the behavior of Equation (29) and the transition from high to low signal-to-noise ratios, in Fig. 3 we display $f_R(R)$ for a series of intensities (x_1, x_2), for constant noise terms $\sigma_1 = \sigma_2 = 20$. In the figure the ratio of intensities x_2/x_1 is always 4 (except for the case where both signals are 0), but the

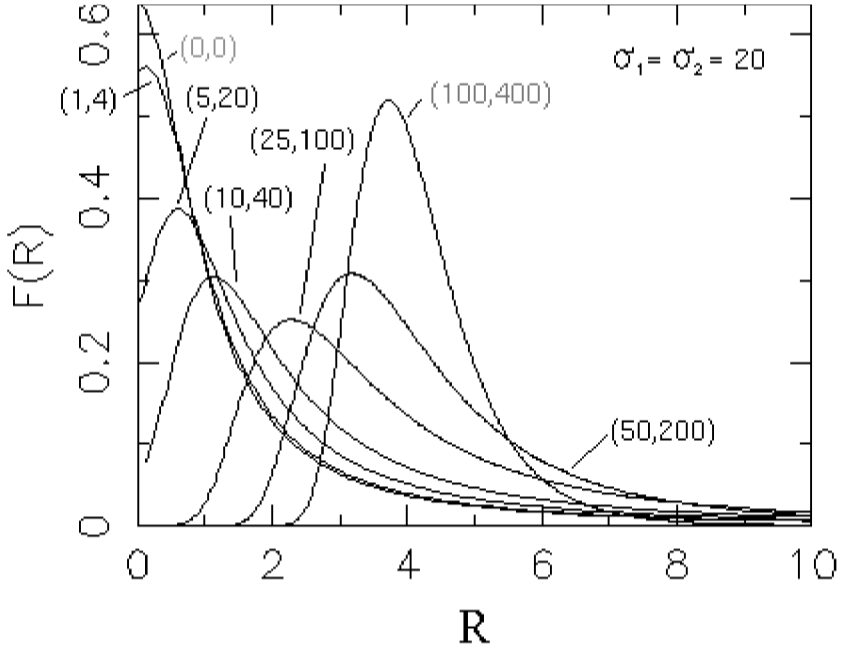

FIG. 3. A posteriori distribution of the fold change R, Eq. (29), for a series of intensity pairs (x_1, x_2) covering the range from high to low signal-to-noise ratios. In all cases but (0, 0), the ratio of intensities is 4. The standard deviation of both noise terms is kept constant at $\sigma_1 = \sigma_2 = 20$. The corresponding values of $\hat{R}$, P, and other statistics are given in Table 2.

signal-to-noise ratios are made to vary through a large range of values, from high to low. For the highest intensities, $(x_1, x_2) = (100, 400)$, as expected $f_R(R)$ is strongly peaked about $R = 4$. However, even in this limit, the 68% confidence interval for R (corresponding to a width of two standard deviations for a normal distribution) is [3, 5], so that even when the signal-to-noise ratios are relatively large, $(x_1/\sigma_1, x_2/\sigma_2) = (5, 20)$, the actual fold-change cannot be inferred to anything better than $3 \lesssim R \lesssim 5$.

With decreasing signal-to-noise ratio, the distribution $f_R(R)$ not only broadens, but its maximum (mode) shifts downwards. Thus, in Fig. 3 for the intensities (10, 40) the median of the distribution is about 2.2, with the mode of the distribution now occurring very close to 1. The broadening and shifting of the distribution function indicates how, for weakening signals, the simple ratio x_2/x_1 of the intensities becomes a less and less reliable indication of the actual fold-change. In the limit where both intensities are zero, (0,0), we recover the Cauchy distribution of Equation (39): the distribution is very broad, with median $R = 1$ and a peak at $R = 0$, and little can be inferred about the actual value of R.

4.1. Departures from Gaussianity

The model underlying the PFOLD algorithm assumes Gaussian noise, whereas measurements indicate that the actual noise is only roughly Gaussian, with tails joining a central, approximately normal distribution. For instance, for the distribution of *absent* genes shown in Fig. 1b, with standard deviation $\sigma_{bc} = 560$, for $x > 2\sigma_{bc}$ the cumulative distribution function $P(x \geq x_0)$ is roughly proportional to an exponential distribution of form $\exp(-x_0/a)$, with $a \approx \sigma_{bc}$. While we have not attempted to quantify this asymptotic behavior in any more detail, we believe that the analytic results we have obtained so far will not be qualitatively modified in any fundamental way by the inclusion of these non-Gaussian features. This is especially true of the important $\sim 1/R^2$ asymptotic dependency of the distribution $f_R(R)$ for large R, which can be shown to be obtained with exponentially distributed as well as with Gaussian noise.

5. BAYESIAN ESTIMATION OF FOLD CHANGES

Equation (29) is all we need to perform Bayesian estimation of the fold-change R, based on knowledge of the intensities x_1 and x_2 and of the corresponding noise terms σ_1 and σ_2. We define the cumulative distribution function

$$F(R') = \mathrm{P}(R \leq R') = \int_0^{R'} f_R(R)dR. \tag{41}$$

As we have not been able to find a closed-form analytic expression for Equation (41), we simply evaluate $F(R)$ using numerical integration, as described below. Based on the numerical values of $F(R)$, we can then obtain all of the following (Fig. 4):

1) *Fold-change estimator* $\hat{R}$: We choose as estimator $\hat{R}$ for the fold-change the *median estimator*

$$\hat{R} = \mathrm{Med}(R), \tag{42}$$

that is, the value of R for which $F(R) = 1/2$. Note that in general other estimators are possible, for instance the MAP (maximum a posteriori probability) estimator or the mean (Van Trees, 1978). The mean is not an option here, as $f_R(R)$ does not have a finite mean. The median estimator has the dual advantages of robustness and symmetry under the transformation $(R \to 1/R)$ and is the one adopted here. Formally, the median estimator is one that minimizes the absolute value of the (estimate–actual value) error term (Van Trees, 1978).

2) *Confidence limits* R_p *and* R_{1-p}: given a probability $p < 1$, we define the confidence limits R_p and R_{1-p} as the values of the corresponding quantiles,

$$F(R_p) = p, \tag{43}$$

$$F(R_{1-p}) = 1 - p. \tag{44}$$

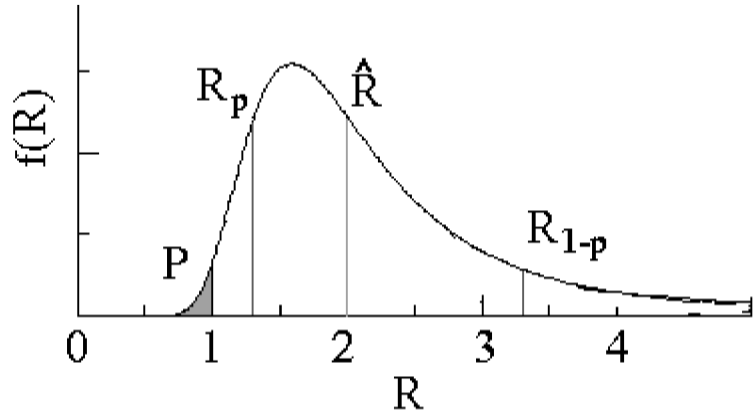

FIG. 4. Graphical definitions for the estimators formally defined in Section 5: $\hat{R}$ is the median of the distribution $f_R(R)$ (equal areas lie to the left and right of $R = \hat{R}$); R_p and R_{1-p} are the p-th and $1-p$-th quantiles, respectively (the areas to the left and right of R_p and R_{1-p}, respectively, are both equal to p, here with $p = 0.2$); the P-value P is the probability that the fold-change was less than 1 and provides a test of the significance of $\hat{R} > 1$. Note the asymmetry of the distribution.

3) *P-value for significance of change*: if $\hat{R} \geq 1$, we can test the hypothesis $R > 1$ ("a significant, positive fold-change occurred in experiment 2 relative to 1") by evaluating the probability of the complementary hypothesis, $R \leq 1$, and defining this as the P-value P of the test for significant change. A symmetrical expression is used if $\hat{R} < 1$. The resulting prescription is

$$P = \begin{cases} F(1), & \hat{R} \geq 1, \\ 1 - F(1), & \hat{R} < 1. \end{cases} \tag{45}$$

Note that by the one-sided nature of the test, the P-value is confined to the range $0 \leq P \leq 0.5$.

The quantity P as defined by Equation (45) is not, strictly speaking, a P-value; the latter is typically defined in the context of a given null hypothesis, which may or may not be rejected given the outcome of a test, whereas here P assesses the probability of an alternative outcome, given a single model which is assumed uniformly valid. However, because P provides a useful measure of significance, and furthermore is well approximated by an actual P-value (see Equation (46) below), we shall continue to describe it, rather loosely, as a "P-value."

Results for all the measurement pairs discussed in connection with Fig. 3 are shown in Table 2 below, with confidence limits determined by $p = 0.16$. Note that the P-value provides a useful selection criterion

TABLE 2. FOLD-CHANGE ESTIMATE $\hat{R}$ AND ASSOCIATED STATISTICS FOR EACH OF THE INTENSITY PAIRS (x_1, x_2) ILLUSTRATED IN FIG. 3[a]

x_1	x_2	R_p	$\hat{R}$	R_{1-p}	P	P_S
100	400	3.3	4.0	5.0	≈ 0.0	≈ 0.0
50	200	2.8	3.9	6.5	5.98×10^{-8}	5.70×10^{-8}
25	100	2.0	3.6	8.8	4.47×10^{-3}	4.00×10^{-3}
10	40	0.93	2.2	7.3	0.18	0.14
5	20	0.51	1.5	5.7	0.34	0.30
1	4	0.32	1.1	4.6	0.46	0.46
0	0	0.23	1.0	4.4	0.5	0.50

[a] R_p = lower confidence limit for R ($p = 0.16$); $\hat{R}$ = median estimator for R; R_{1-p} = upper confidence limit for E ($1 - p = 0.84$); P = P-value; and P_s = approximation to P, Eq. (46).

for retaining only significant measurements. Thus, while in Table 2 all intensity ratios are equal to 4 (except for (0,0)), only the first three entries ((100, 400), (50,200), (25,100)) are found to indicate change at the 0.05 confidence level. Furthermore, for each of the entries which are deemed significant, we can provide confidence limits for the fold-change. For instance, for the measurement pair (25, 100), with $P = 4.47 \times 10^{-3}$, the estimate of $\hat{R} = 3.6$ is bracketed by [2.0, 8.8], showing that in this case we cannot "nail" the fold-change to anything better than this interval (i.e., at the 68% confidence level fold-changes as small as 2 and as large as 8.8 are also consistent with the data).

An approximation P_S for P is based on a simple one-sided test of the significance of the difference $x_2 - x_1$,

$$P_S = \mathrm{erfc}\left(\frac{|x_2 - x_1|}{2^{1/2}(\sigma_1^2 + \sigma_2^2)^{1/2}}\right). \tag{46}$$

where erfc is the complementary error function (Abramowitz and Stegun, 1972). Values of P_S are listed in Table 2, where it can be seen that P_S is slightly smaller than P (and therefore slightly overestimates statistical significance), because the test in Equation (46) does not incorporate the positivity of the concentrations. Nonetheless, Equation (46) generally provides a very good approximation to P and can be used as a simpler formula than extraction from Equation (29). Note, however, that computation of $\hat{R}$ (Equation (42)) and of the confidence limits R_p and R_{1-p} (Equations (43) and (44)) explicitly require the complete formula given in Equation (29).

5.1. *Computer implementation: the PFOLD algorithm*

The estimation scheme described above and summarized by Equations (29) and (42–45) has been implemented in a C++ program called PFOLD. For a given set of input parameters $(x_1, x_2, \sigma_1, \sigma_2)$ specifying the two intensities and the corresponding standard deviations of the noise terms, PFOLD first numerically evaluates the distribution function $f_R(R)$ (Equation (29)) over a finite range $R_{min} \le R \le R_{max}$ at points on a regular mesh $R_i = R_{min} + i\,\Delta R$, $i = 0, 1, \ldots, N$, where R_{min}, R_{max} and N ($\Delta R = (R_{max} - R_{min})/N$) are automatically chosen to capture all of the variation of the function (Fig. 3). The cumulative distribution function $F(R)$ (Equation (41)) is then found by numerical integration of $f_R(R)$, following which all the estimators of Section 5, that is the fold change $\hat{R}$, the confidence limits (R_p, R_{1-p}), and the P-value P, can be readily evaluated by numerically solving for Equations (42), (43), (44) and (45), respectively. In finding the roots of these equations, a simple bisection method (Press *et al.*, 1997, p. 353) is used.

5.2. *Mapping intensity pairs* (x_1, x_2) *into the* $(\hat{R}, P)$ *plane*

Pairs of intensities (x_1, x_2) are mapped by Equations (42) and (45) into pairs of numbers $(\hat{R}, P)$, a mapping which results in a significance-weighted representation of the fold-changes. Figures 5a,b illustrate how mapping into the $(\hat{R}, P)$ plane provides a useful, alternative representation of experiments. Here, identical RNA samples were hybridized to two Affymetrix chips of the same type (Mu19KsubA, with features for 7,045 genes), with the data shown in Fig. 5a where the intensity pairs (x_1, x_2) are displayed in a scatter plot where each point represents one gene. In the figure, 4,761 genes are visible; the remaining 2,284 have a negative intensity in one or both of the scans, and are invisible in the log–log plot. Note also that the slight leveling off of the line of the scatter plot for the largest intensities, $x_1, x_2 > 10,000$, is due to saturation effects.

Based on the representation of Fig. 5a, a "traditional" way of selecting for genes with significant change is to require that the fold-change R or its inverse $1/R$ is above a given threshold R_c, with R directly computed from the ratio of intensities, $R = x_2/x_1$. In Fig. 5a, the decision boundaries for $R_c = 2$ are indicated, with the acceptance region outside the two parallel lines. Because this procedure also selects for a large number of genes with very low signal-to-noise ratio, one usually limits each of the intensity values entering into the ratio calculation x_2/x_1 to some lower bound, representative of the noise level. However, here we circumvent this device and instead directly use the PFOLD prediction for the fold-change, selecting genes with $\hat{R} \ge R_c$ or $\hat{R} \le 1/R_c$. The result is shown in Fig. 6a for $R_c = 2$, with 83 genes shown in the scatter plot.

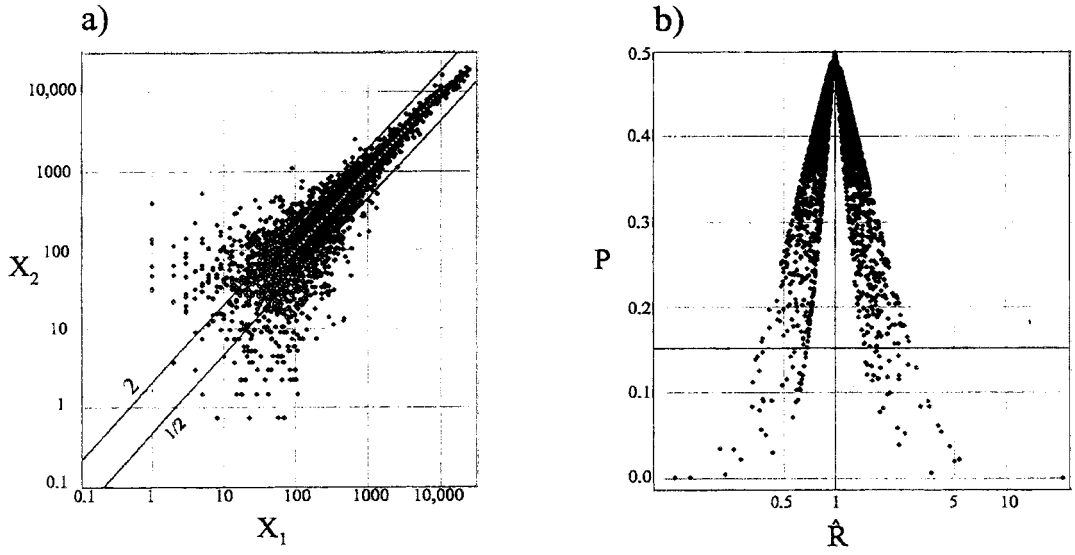

FIG. 5. Scatter plots showing two alternative views of a reproducibility experiment, in which identical samples derived from murine intestine tissue, here labeled 1 and 2, were hybridized to two Affymetrix chips of the same type (Mu19KsubA, features for 7,045 genes). **a)** Representation in the (x_1, x_2) plane: each point in the plot corresponds to one gene (scan 2 was rescaled by an overall factor of 0.73 to make its mean brightness equal to that of scan 1, so that the line of symmetry has slope 1). In a) 4,761 genes out of 7,045 are visible in the log-log plot, the remaining 2,284 having a negative intensity in one or both of the scans. **b)** Representation in the $(\hat{R}, P)$ plane. All 7,045 genes are present in this plot.

A drawback of the selection method outlined above is the risk of "throwing out the baby with the bathwater"; in the interest of controlling false positives, we may have to impose such a large value of R_c that high intensity genes with significant change will be rejected as well. An alternative representation of the data, which allows for more flexibility in operating selections, is the one shown in Fig. 5b, where each gene is now mapped into the $(\hat{R}, P)$ plane (all 7,045 genes are shown). The important property of this new

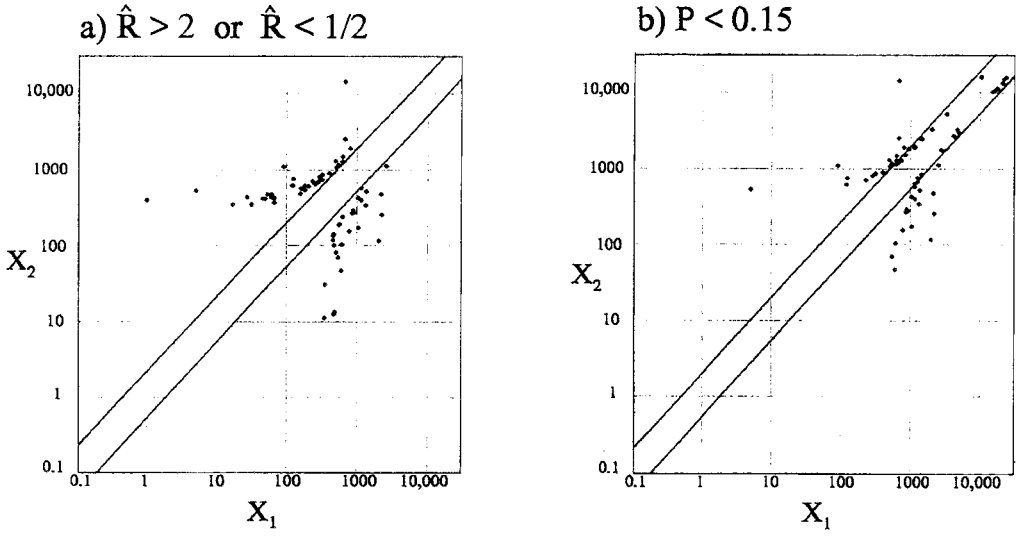

FIG. 6. Two subsets of genes selected from the data in in Figs. 5a,b, in ways suggested by the (x_1, x_2) or $(\hat{R}, P)$ representations, respectively. **a)** Select for genes with greater than 2-fold change in 2 relative to 1, using the PFOLD estimate of $\hat{R}$ (83 genes are obtained); **b)** select for genes with P-value P less than 0.15 (73 genes are obtained). Figs. a) and b) have 41 genes in common.

representation is that one can do selections for subsets of genes which are based on *both* the fold-change and P-value. Thus, in Fig. 6b we selected for genes with a P-value less than 0.15, with no constraint on the fold-change, resulting in the 73 genes shown. Note that while roughly the same number of genes are selected in Figs. 6a and b, the populations only partially overlap, with 41 genes in common. In particular, the use of the P-value for selections enables one to find high-intensity genes with significant fold-change (although less than 2), while simultaneously filtering out noisy, low-intensity data (compare Figs. 6a and 6b.) The expectation is thus that by making selections based on both P and $\hat{R}$, one should obtain detection which is more sensitive, at equal selectivity, than selections based on $\hat{R}$ alone. We examine this assumption in the next two sections.

6. THEORETICAL VALIDATION BY MONTE CARLO SIMULATIONS

In order to evaluate the usefulness of the PFOLD algorithm, we conducted Monte Carlo simulations (Cowan, 1998, p. 41; Press *et al*., 1997, p. 689) aiming to approximate actual experiments, with a focus on the ability of PFOLD to discriminate a class of genes with a given fold-change in expression from a "background" class of genes with unchanging expression levels. To approximate a physical distribution, concentration values n (normalized as in Equation (12)) were generated according to a log-normal distribution (Keeping, 1995, p. 89), by computing

$$n = \exp(y), \tag{47}$$

where y is a Gaussian random variable generated with mean and standard deviation $< y > = 6.56$ and $\sigma_y = 1.22$, respectively. To get an indication of the range of concentrations implied by these values, note that the 25th, 50th, and 75th percentiles of n are (300, 700, 1600), respectively, which are typical of experiments using Affymetrix chip technology.[8]

For each value of n generated by Equation (47), an actual fold-change of b, combined with noise, was simulated by computing the two intensity values

$$x_1 = n + \epsilon_1, \tag{48}$$

$$x_2 = bn + \epsilon_2, \tag{49}$$

where the noise terms ϵ_1 and ϵ_2 are uncorrelated Gaussian random variables with zero means and with equal standard deviation σ_ϵ given by Equation (15) with parameters $\sigma_{bc} = 300$, $\alpha = 0.25$, again chosen to be typical of Affymetrix chip experiments. The value $\sigma_{bc} = 300$ implies that in this model the lowest quartile of genes have intensities below or at most comparable to the noise level. The parameters chosen for the simulations are close to the "typical" values indicated in Table 1, but result in a slightly smaller median signal-to-noise ratio, $x_m/\sigma_{\epsilon^m} = 2$ versus of 2.2.

From the intensities (x_1, x_2) computed with Equations (48) and (49), the corresponding estimators $(\hat{R}, P)$ were then computed using Equations (42–45), with PFOLD parameters (σ_{bc}, α) identical to those used in generating the data. While in experiments values of b as large as 100 can be measured, the bulk of genes which change significantly are expected to vary with fold changes in the range $1 \lesssim b \lesssim 5$. In what follows, we initially focus on $b = 3$, then explore the effect of variation over a finite range of b.

6.1. Validation methodology

We conducted two simulations, with $b = 1$ and $b = 3$, defining the two classes of genes,

class 0: no change, $b = 1$,
class 1: change, $b = 3$.

[8]The numbers are representative of scans of Affymetrix chips where the fluorescence intensities are enhanced by antibody staining.

We then used PFOLD to classify genes by defining an acceptance region D in the $(\hat{R}, P)$ plane (Cowan, 1998, p. 47), and with prediction π for the class membership of a gene given by

$$\pi = \begin{cases} p, \text{ assign gene to class 1, if}(\hat{R}, P)\text{in}D, \\ a, \text{ assign gene to class 0, if}(\hat{R}, P)\text{not in}D, \end{cases} \tag{50}$$

where p and a stand for *present* and *absent* in the acceptance region, respectively. An example of an acceptance region D is one with a rectangular decision surface

$$D = \{\hat{R} \geq R_c, P \leq P_c\}, \tag{51}$$

but we considered other types of regions as well. Figures 7a and b display the $(\hat{R}, P)$ scatter plots generated by 1,000 genes from class 0 (the no-change class) and 1,000 genes in class 1 (the genes that changed 3-fold), respectively. In Fig. 7a, the boundaries of two acceptance regions, to be discussed in detail below, are indicated.

For a given D, one can estimate the misclassification probabilities

$P(p|0)$ = probability that a gene in class 0 gets assigned to class 1,
$P(a|1)$ = probability that a gene in class 1 gets assigned to class 0,

by directly counting the number of misclassification errors resulting from each simulation.

The model also uses as input the a priori probabilities

- P_0 = a priori probability that a gene is in class 0,
- P_1 = a priori probability that a gene is in class 1.

The values expected for P_1 are context-dependent, but will typically be small, as in many experiments only a small proportion of genes are actually changing in response to a perturbation or induction event. A biologically realistic range might be taken to be $P_1 \sim 0.01 - 0.2$ (1% to 20% of genes changing significantly).

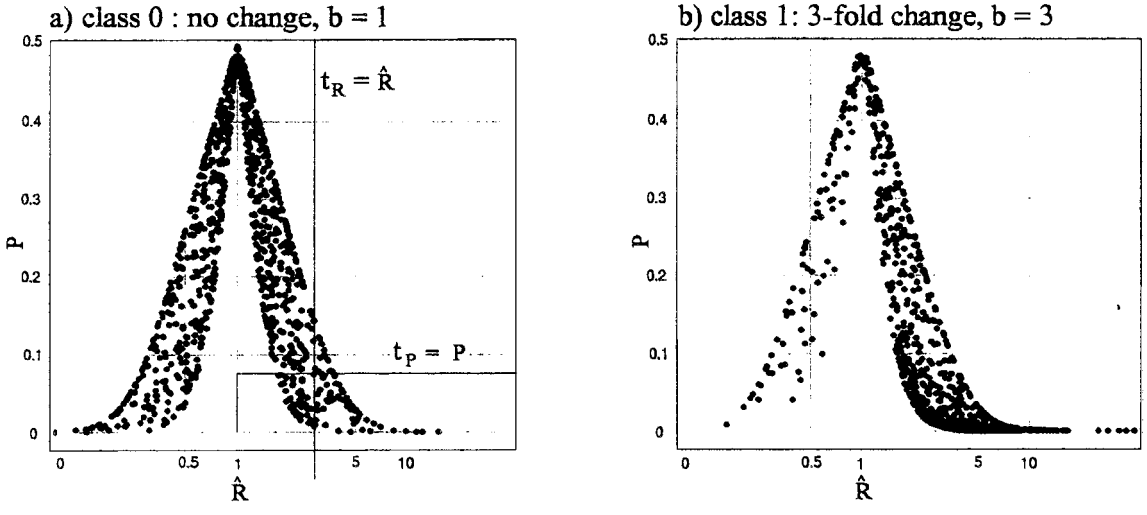

FIG. 7. Scatter plots in the $(\hat{R}, P)$ plane generated by the Monte Carlo methods described in Section 6, for the actual fold-changes $b = 1$ and $b = 3$. **a)** Class 0 (the no-change class, $b = 1$); **b)** class 1 ($b = 3$). The decision boundaries corresponding to $t_R = \hat{R}$ and $t_P = P$ are indicated in a).

From the Bayes Theorem we obtain the a posteriori probabilities for misclassification, defined as

- $P(0|p)$ = probability that a gene assigned a significant fold-change did not really change,
- $P(1|a)$ = probability that a gene assigned to the no-change category actually changed.

The result is (Cowan, 1998, p. 49)

$$P(0|p) = P_0 P(p|0)/P_p, \tag{52}$$

$$P(1|a) = P_1 P(a|1)/P_a, \tag{53}$$

where P_p and P_a, the total a posteriori probabilities of declaring a gene in class 1 or class 0, respectively, are given by

$$P_p = P(p|0)P_0 + P(p|1)P_1, \tag{54}$$

$$P_a = P(a|0)P_0 + P(a|1)P_1, \tag{55}$$

where all the conditional probabilities on the right-hand sides of Equations (52–55), $P(p|0)$, $P(a|1)$, etc., are estimated by direct counting of simulation events.

In order to evaluate the performance of PFOLD on the classification task, we focused on two metrics, a false-positive rate FP, and a sensitivity S, defined as follows:

$$FP = P(0|p) = \text{a posteriori false positive rate}, \tag{56}$$

$$S = P(p|1) = 1 - P(a|1) = \text{a priori true positive rate}. \tag{57}$$

The definitions of Equations (57) and (56) are not symmetrical because the calculation of $P(0|p)$ requires the value of the prior P_1, while the computation of $P(p|1)$ does not. The reason for this choice are the following: the a posteriori false-positive rate FP is a measure of the "contamination" (Cowan, 1998, p. 47) by spurious candidates of a gene list picked by the PFOLD classification. In the context of a search for drug targets, for instance, after picking, the gene list would be submitted to an experimental validation pipeline. The quantity FP is a direct measure of wasted effort (on spurious candidates) that would be expected downstream in the pipeline, and thus can be directly equated to a cost. The sensitivity S, on the other hand, is a measure of the efficiency (Cowan, 1998, p. 47) of the classification scheme in finding targets among all those actually in existence. As usual, the stringency of the decision process must be adjusted so as to balance FP and S in some optimal way, as discussed below.

6.2. *Simulation results*

A summary of the simulation results for classification is shown in Table 3, where we have based the evaluation of PFOLD performance on the so-called "receiver operating characteristic" (ROC) (Van Trees, 1978) for the various statistics used. The ROC enables one to visualize the tradeoffs in trying to simultaneously minimize false-positive and false-negative rates. Thus, in Fig. 8, the ROC based on the data of Figs. 7a,b is shown for the decision statistic $t_R = \hat{R}$, corresponding to the acceptance region (Fig. 7a)

$$D_R = \{\hat{R} \geq R_c\}. \tag{58}$$

The false-positive rate $P(0|p)$ and false-negative rate $P(a|1)$ are plotted as functions of the cutoff R_c for $P_1 = 0.2$ (20% a priori probability that a gene actually changed). At the very lowest stringency, $R_c = 0$, all genes are indiscriminately accepted, so that $P(a|1) = 0$, $P(0|p) = P_0 = 0.8$. As R_c is increased, the false-positive rate decreases, but there is a concomittant rise in the false-negative rate. Thus, the choice of the "optimal" cutoff R_c depends on deciding on an acceptable compromise between the error rates. For instance, if a false-positive rate $FP = P(0|p) = 0.3$ is given as the largest acceptable (30% of accepted

TABLE 3. PERFORMANCE OF THREE STATISTICS USED FOR THE DETECTION OF GENES WITH TRUE FOLD-CHANGE OF $b = 3$ (CLASS 1) AGAINST A BACKGROUND OF UNCHANGING GENES (CLASS 0), AS SIMULATED BY THE MONTE-CARLO METHOD DESCRIBED IN SECTION 6[a]

Decision parameters		*Results for constant F.P. rate* $P(0\|p) = 0.3$		
Main statistic t	*Region*	t_c	*Sensitivity* $S = P(p\|1)$	*Median fold-change* $Med(\hat{R})$
Fold-change $t_R = \hat{R}$	$\hat{R} \geq t_c$ $P \leq 0.5$	4.37	0.24	6.1
P-value $t_P = P$	$P \leq t_c$, $\hat{R} \geq 1$	0.037	0.52	4.2
Discriminant $t_F = \log(\hat{R}) - 5.48P$	$t \geq t_c$	0.95	0.50	4.4

[a]The fraction of changing genes is $P_1 = 0.2$, and performance is given for a "clamped" false-positive rate of $FP = P(0|p) = 0.3$. For each case, we specify: the acceptance region used in connection with the statistic; the value t_c of the statistic for $FP = 0.3$; the corresponding sensitivity S obtained for $FP = 0.3$; and the median of the predicted fold-changes for all the accepted genes.

genes spurious hits), then $R_c = 4.37$, at which point the sensitivity is $S = P(p|1) = 1 - P(a|1) = 0.24$ (only 24% of all positives are actually detected; see Table 3). Note also that the false-positive rate levels off as a function of R_c for R_c large, so that it cannot be made arbitrarily small, even when reduced sensitivity is acceptable; this is because in Equations (52, 54), defining $P(0|p)$, both $P(p|0)$ and $P(p|1)$ have the same asymptotic trend as $R_c \to \infty$.

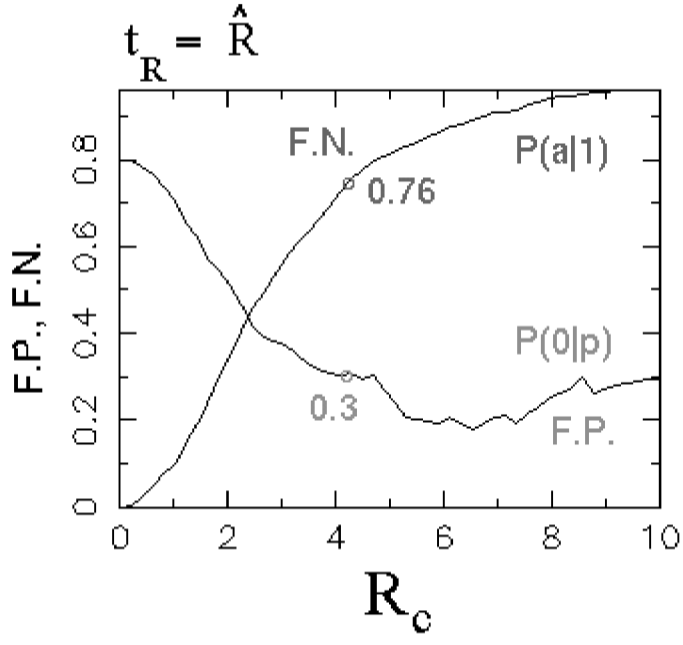

FIG. 8. Receiver operating characteristic (ROC) for the statistic $t_R = \hat{R}$, with decision surface given by Eq. (58), using the data generated by the Monte Carlo simulations of Section 6 for an actual fold-change of 3 against a background of unchanging genes, and with a fixed prior probability of change $P_1 = 0.2$. The decision boundary is at $R = R_c$; increasing R_c increases the stringency of the acceptance process by selecting for larger-and-larger fold-changes. The curve marked F.P. is the false-positive rate FP as a function of R_c (Eq. (56)), the curve labeled F.N. is the false-negative rate ($1 - S$, with S given by Eq. (57)). The point at which $FP = 0.3$ ($S = 1 - 0.76 = 0.24$) is indicated in the figure.

In Fig. 9, we show the ROC for a second decision statistic, $t_P = P$, used in conjunction with the fixed filter $\hat{R} \geq 1$. The acceptance region is the rectangle (Fig. 7a)

$$D_P = \{P \leq P_c, \hat{R} \geq 1\}. \tag{59}$$

The false-positive rate $P(0|p) = 0.3$ is now achieved with $P_c = 0.037$, with the sensitivity $S = P(p|1) = 0.52$ (Table 3), over twice the sensitivity achieved by the t_R statistic. The t_P statistic also has the advantage over the t_R statistic that there is no leveling off of the false-positive rate $P(0|p)$, which here can be indefinitely reduced by taking $P_c \to 0$.

For each statistic, the median values of $\hat{R}$ for the population of detected genes are given in Table 3, giving an indication of the accuracy of quantitation of the fold-change (the exact value should be 3). While in both cases there is an upward bias resulting from the acceptance process, it can be seen that the effect is weaker for the t_P statistic, yielding a more accurate estimate of the fold-change.

In Table 3, we also indicate the results obtained using a Fisher linear discriminant (Duda and Hart, 1973) based on an analysis in $(\log(\hat{R}), P)$ space. Computing scatter matrices while restricting the data to $\hat{R} \geq 1$, the Fisher discriminant is found to be

$$t_F = \log(\hat{R}) - 5.48P, \tag{60}$$

with results shown in Table 3. Because the sensitivity obtained with t_F is not better than that found with the simpler t_P, we did not pursue the use of this statistic, although it might prove useful in other contexts.

In Fig. 10, the dependence of the sensitivities on P_1, the fraction of the total gene population that has actually changed, is explored for both t_R and t_P statistics. In obtaining these results, the false-positive rate is "clamped" to the constant value $P(0|p) = 0.3$. For $P_1 > 0.5$ (more than half the genes changing, not shown in the figure), the sensitivities are almost equal, and on the basis of sensitivity, there is no reason to prefer one statistic over the other. On the other hand, for the range $P_1 \leq 0.5$ shown in the figure, which is biologically much more relevant, the sensitivity obtained with t_P is markedly superior. Furthermore,

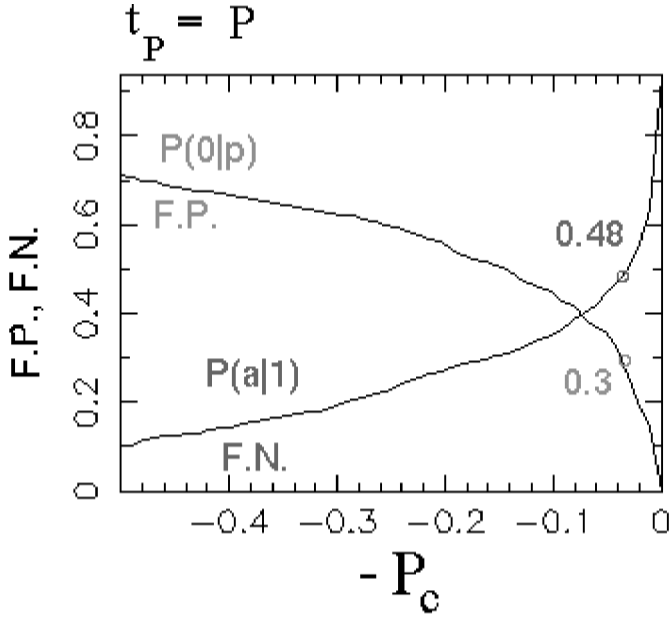

FIG. 9. Receiver operating characteristic (ROC) for the statistic $t_P = P$ with decision surface given by Eq. (59), using the data generated by the Monte Carlo simulations of Section 6 for an actual fold-change of 3 against a background of unchanging genes, and with a fixed prior probability of change $P_1 = 0.2$ (compare with Fig. 8). The decision boundary is at $P = P_c$; *decreasing* P_c increases the stringency of the acceptance process by selecting for changes with greater significance. The curves marked F.P. and F.N. are as in Fig. 8, with the point at which $FP = 0.3$ ($S = 1 - 0.48 = 0.52$) indicated in the figure.

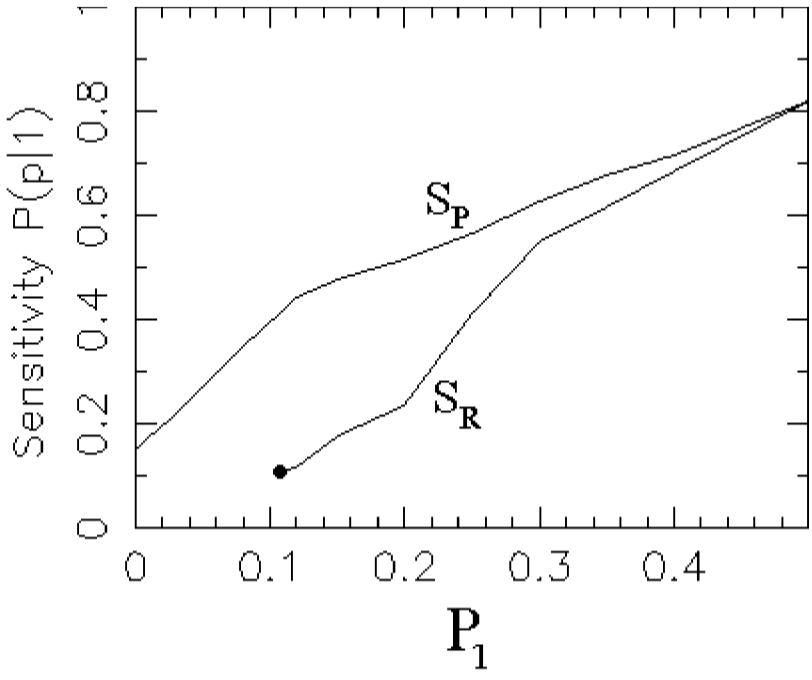

FIG. 10. Sensitivity as a function of P_1, the fraction of genes that underwent a 3-fold change against a background of unchanging genes, for the t_R and t_P statistics (denoted by S_R and S_P, respectively) using the data generated by the Monte Carlo simulations of Section 6. The false positive rate $FP = P(0|p)$ is "clamped" to the value 0.3 for all values of P_1. The black dot indicates the breakdown of the decision process based on the t_R statistic: below that point the false positive rate cannot be kept at 0.3.

for $P_1 \leq 0.11$, the selection based on t_R cannot even maintain the required minimum false-positive rate of $P(0|p) = 0.3$, while the t_P statistic can be used to arbitrarily small values of P_1. At the point of breakdown of the t_R statistic, $P_1 = 0.11$, the sensitivity for t_P is about four times greater, with $S = 0.42$ versus $S = 0.11$ (42% of all changing genes detected by t_P, versus only 11% by t_R).

The dependence of the sensitivities on the fold-change b, for a constant fraction of changing genes $P_1 = 0.2$, is explored in Figs. 11a,b. As in Fig. 10, the false positive rate is kept fixed at $P(0|p) = 0.3$. Figure 11a shows that that the sensitivities are almost equal for $b \gtrsim 5$, so that detection of genes with large fold-changes is about equal with both t_R and t_P statistics. On the other hand, the magnification to the biologically important range $1 \leq b \leq 5$, Fig. 11b, shows that for moderate to small fold-changes, the t_P statistic is definitely superior in sensitivity. In particular, detection with the t_R statistic breaks down at $b = 2.3$, as a false positive rate $P(0|p) = 0.3$ cannot be maintained below that value of the fold-change.

In summary, a decision statistic based on the PFOLD P-value appears markedly superior in sensitivity to one based only on the fold-change estimator $\hat{R}$, whenever one wishes to detect changing genes in the biologically relevant range of $P_1 \lesssim 0.2$ and $1 \lesssim b \lesssim 5$.

6.3. Dependence of detector performance on the choice of PFOLD parameters

The analysis presented above used PFOLD parameters identical to those chosen in the first place to generate the Monte Carlo simulation data. Because in the analysis of actual experiments the PFOLD parameters will be at best only approximations to the actual parameters, we systematically investigated the sensitivity of the results to variations in the choice of the PFOLD parameters.

In Fig. 12a, we show the dependence of the sensitivity S on the ratio of parameters $\sigma_{bc}^{pfold}/\sigma_{bc}$, where σ_{bc}^{pfold} is the value used in PFOLD and $\sigma_{bc} = 300$, the value used in generating the simulation data. In this study, the coefficient of variation used in PFOLD is the same as in the simulations, $\alpha^{pfold} = \alpha = 0.25$, and S is calculated for fixed false-positive rate $FP = 0.3$ and fixed proportion of changing genes, $P_1 = 0.1$. As expected, the sensitivity is greatest when the model parameter coincides with that of the simulation

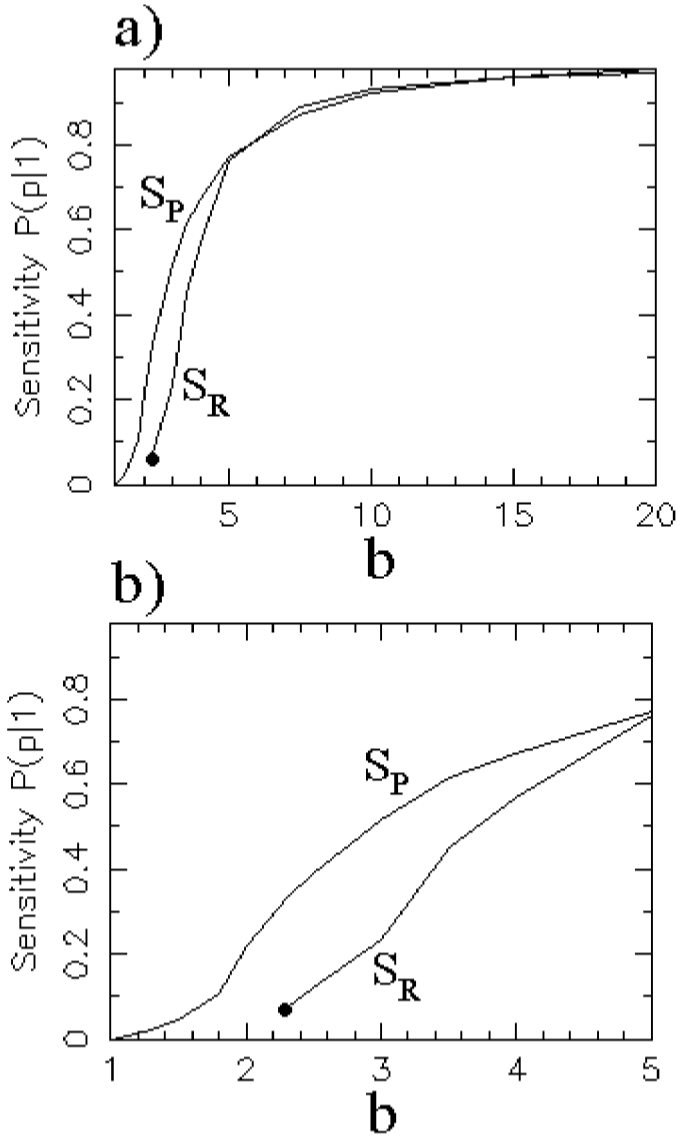

FIG. 11. Sensitivity as a function of the true fold-change b, for a constant fraction of changing genes $P_1 = 0.2$, using the data generated by the Monte Carlo simulations of Section 6 and for decision based on the $t_R = \hat{R}$ or $t_P = P$ statistics (denoted by S_R and S_P, respectively). The false positive rate $FP = P(0|p)$ is "clamped" to a fixed value of 0.3; **a)** dependence in the range $1 \leq b \leq 20$; **b)** magnification to the range $1 \leq b \leq 5$. The black dot indicates the breakdown of the t_R statistic as a function of b: below that point the false positive rate cannot be kept at 0.3.

data, $\sigma_{bc}^{pfold} = \sigma_{bc}$, with performance falling off when $\sigma_{bc}^{pfold} \neq \sigma_{bc}$. However, with overestimation of the noise, $\sigma_{bc}^{pfold} > \sigma_{bc}$, the degradation in detector performance is "graceful": even for $\sigma_{bc}^{pfold}/\sigma_{bc} = 2$, the sensitivity is still 85% of its maximum. On the other hand, underestimation the noise has more serious consequences and leads to a rapid degradation in performance.

The dependence of the sensitivity S on the asssumed coefficient of variation α^{pfold} is shown in Fig. 12b, now for fixed $\sigma_{bc}^{pfold} = \sigma_{bc}$ and otherwise under the same conditions as in Fig. 12a. As before, sensitivity is greatest when the model parameter coincides with the actual simulation parameter, $\alpha^{pfold} = \alpha = 0.25$ (arrow). For $\alpha^{pfold} \neq \alpha$, there is little degradation in performance in the range $0.1 \leq \alpha^{pfold} \leq 0.3$, with a sharper dropoff outside of this range.

In a final set of numerical experiments, we investigated the effect of introducing a nonzero correlation coefficient ρ between the noise terms used in the simulations (Equation (17)), thereby changing the

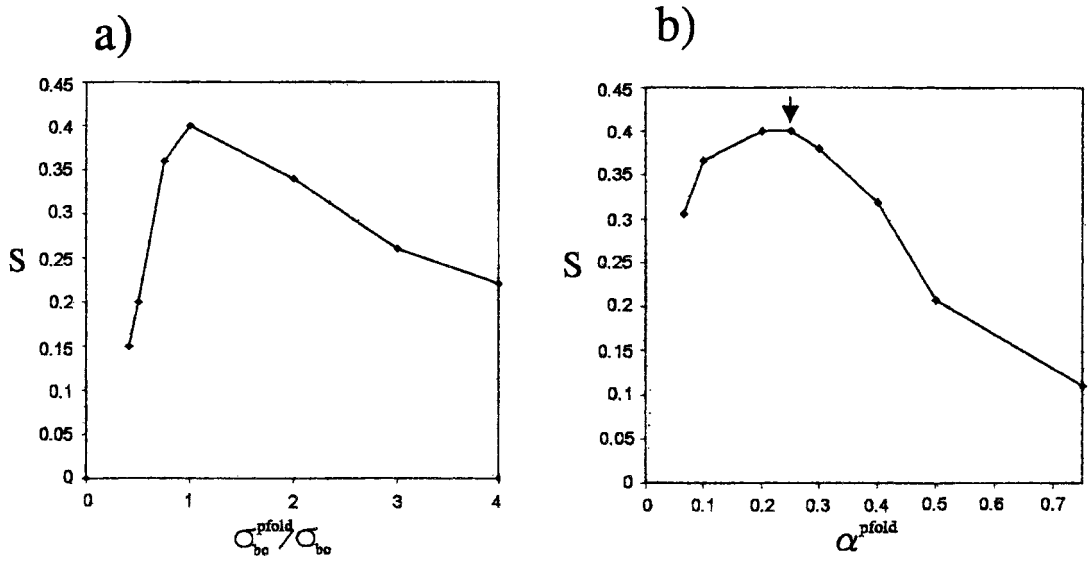

FIG. 12. Dependence of detector performance on the choice of PFOLD parameters. Throughout, the sensitivity S is calculated for a fixed false-positive rate $FP = 0.3$ and fixed proportion of changing genes, $P_1 = 0.1$, for the simulation data of Section 6 ($\sigma_{bc} = 300$, $\alpha = 0.25$). **a)** Dependence of S on $\sigma_{bc}^{pfold}/\sigma_{bc}$ where σ_{bc}^{pfold} is the value used by the PFOLD noise model, and with $\alpha^{pfold} = \alpha$. Note the breakdown of the scheme at the limiting value $\sigma_{bc}^{pfold}/\sigma_{bc} = 0.41$, below which the false-positive rate of 0.3 cannot be maintained. **b)** Dependence of S on α^{pfold} for $\sigma_{bc}^{pfold} = \sigma_{bc}$; the point where $\alpha = 0.25$ is indicated by an arrow.

statistical nature of the data, while maintaining $\sigma_{bc}^{pfold} = \sigma_{bc}$ and $\alpha^{pfold} = \alpha$, and with PFOLD itself not incorporating a correlated noise model. Under the same conditions of detection as above ($FP = 0.3$, $P_1 = 0.1$), the sensitivities computed for $\rho = 0$, 0.5, and 0.8 were $S = 0.4$, 0.59, and 0.63, respectively. These results show that with correlated noise, detection actually becomes "easier," even for for the current version of PFOLD, which does have a correlated noise model.

In summary, PFOLD does not display extreme sensitivity to the choice of model parameters and is robust to changes in the statistical nature of the noise used in generating the simulation data. Based on the results shown in Figs. 12, in choosing parameters for PFOLD estimation, it is preferable to overestimate σ_{bc} rather than underestimate it and, inversely, to slightly underestimate α rather than overestimate it.

7. EXPERIMENTAL VALIDATION BY cRNA SPIKING EXPERIMENTS

While Monte Carlo simulations enable one to explore the parameter space relevant to the PFOLD algorithm, they are no substitute for actual experiments. Thus, a set of cRNA[9] spiking experiments was designed to apply the decision methodology used in the previous sections to a realistic setting.

Thirteen cRNA probe samples targeting a total of 13 genes featured on the Affymetrix Mu11KsubA chip were separately produced by in vitro transcription of the corresponding cDNAs and spiked together at known concentrations into a complex, biological background of cRNAs derived from MC3T3 cell lines. Spiking concentrations of 0 pM (no spike, background only) and 5, 15, and 50 pM were chosen to approximate a range of naturally occurring concentrations. The resulting spiked samples were then hybridized to Mu11KsubA chips. For the 0 pM, background-only sample, all 13 genes were signaled as "absent" by the Affymetrix GeneChip decision algorithm.[10] We took this result as sufficient indication that

[9] See footnote 2 on page 587.

[10] See footnotes 3 on page 587 and 5 on page 589.

the corresponding mRNA transcripts were physically absent (or present in only negligible concentrations) in the complex background, so that the total, in-background concentrations for the 13 genes could be assumed equal to the known concentrations of the spikes alone.

Based on the intensity data obtained from the chip scans, the $(\hat{R}, P)$ values were then computed using the PFOLD algorithm for all 13 genes and for the sample pairs (5pM, 15pM) and (15pM, 50pM), corresponding to actual fold-changes of 3 and 3.33 respectively. The 26 resulting values of $(\hat{R}, P)$ were then pooled together, the resulting composite data set simulating a total of 26 distinct genes changing expression by about 3-fold, against a complex background of many unchanging genes, and with a distribution of initial concentrations equally split between 5 and 15 pM. This composite data set can be regarded as a crude approximation to the actual situation of a much larger number of genes changing 3-fold, against a large background of unchanging genes, and starting from a continuous and broad distribution of initial concentrations (rather than just the two values considered here).

The average intensities for the 13 spiked genes at the 5 and 15 pM concentrations were 700 and 1,300, respectively, corresponding to the median and 75th percentiles of the 2,900 genes signaled "present" on the chip (out of a total of 7,045); thus, the spiking experiments simulated what would be the up-regulation, in a biological setting, of the genes which are among the moderate to high expressors, but not the concomittant up-regulation of the low expressor genes, which would signal at or below the 25th percentile in intensity.

Figs. 13a and b display the scatter plots in the $(\hat{R}, P)$ plane based on either a repeat experiment of the background-only sample (Fig. 13a, defining class 0, of no-change genes) or on the 26 values of $(\hat{R}, P)$ obtained from the spiking experiments (Fig. 13b, defining class 1, of changing genes). In Fig. 13b, the 13 points corresponding to the 15pM:50pM concentration ratios are indicated by black dots and the 13 points for the 5pM:15pM concentration ratios by open circles; not surprisingly, most of the P-values for the changes at the higher concentrations (higher signal-to-noise ratios) are markedly smaller than those for the changes at lower concentrations.

The same sensitivity analysis based on receiver operating characteristics was performed on this data set as had been performed for the Monte Carlo simulations (Section 6.2). Note that although only 26 data points are provided to define class 1, whereas 7,045 are provided for class 0, the data can be used to estimate sensitivity for any a priori probability of change P_1 through Equations (52–55). In Fig. 14 the

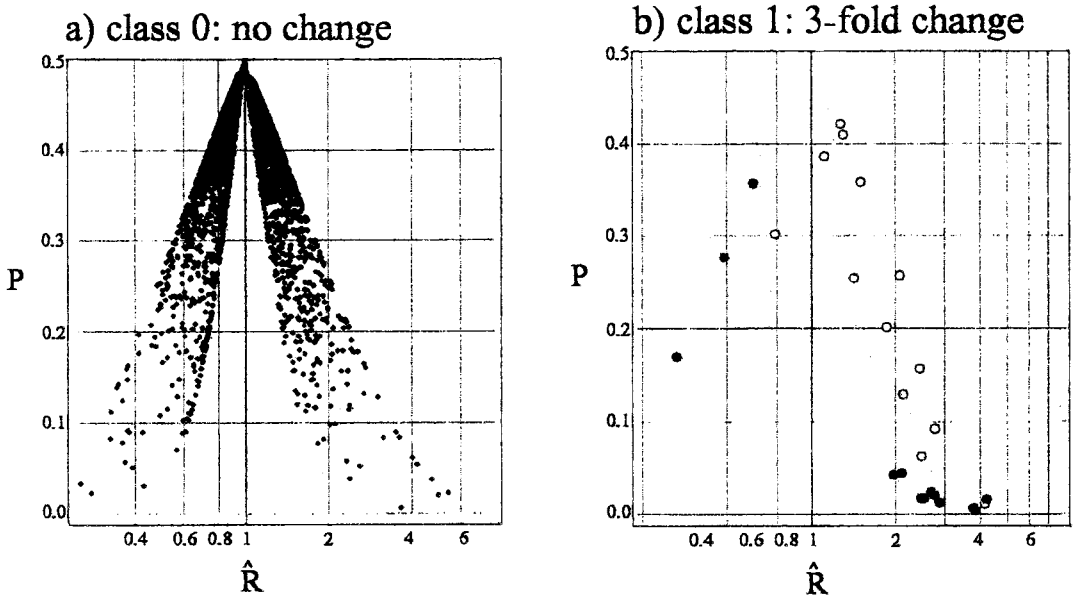

FIG. 13. Scatter plots in the $(\hat{R}, P)$ plane for the validation spiking experiments discussed in Section 7; **a)** class 0, no change: plot for the 6,584 genes on the Mu11KsubA chip, derived from replicates of the background-only hybridization; **b)** class 1, 3-fold change: plot for the 26 $(\hat{R}, P)$ pairs obtained from the spikes at finite concentrations; black dots: points for the 15pM:50pM concentration ratios; open circles: points for the 5pM:15pM concentration ratios.

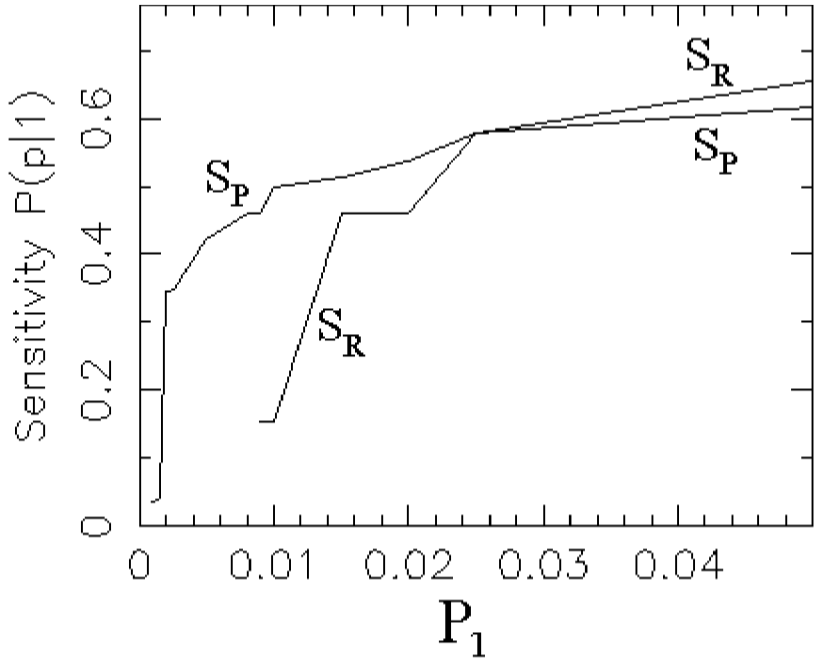

FIG. 14. Sensitivity as a function of the fraction P_1 of changing genes, using the data generated by the spiking experiments, and shown in Fig. 13 (Section 7). The false positive rate $FP = P(0|p)$ is "clamped" to the value 0.3. The dependency of the sensitivity on P_1 is shown for both $t_R = \hat{R}$ and $t_P = P$ decision statistics (denoted by S_R and S_P, respectively).

sensitivity S is plotted as a function of P_1, for both the $t_R = \hat{R}$ (Equation (58)) and $t_P = P$ (Equation (59)) statistics, for a "clamped" value of the false-positive rate $P(0|p) = 0.3$.

The experimental results for the sensitivity shown in Fig. 14 confirm the results of the Monte Carlo simulations (Fig. 10) in that for small P_1 ($P_1 \lesssim 0.03$) the t_P decision statistic becomes markedly superior to the t_R statistic. A salient difference however is that, based on the experimental data, sensitivity for t_P is superior when $P_1 \lesssim 0.03$ ($\lesssim 3\%$ of changing genes), whereas the the Monte Carlo simulations suggest a larger range, $P_1 \lesssim 0.3$ ($\lesssim 30\%$ of changing genes). We believe that this dicrepancy arises from an overestimation of the noise in the Monte Carlo simulations relative to the experimental setup: the Monte Carlo simulations assume strictly uncorrelated noise, $\rho = 0$, whereas in the spiking experiments the artificially unchanging nature of the background impose a correlation coefficient $\rho \sim 1$, which results in significantly less variance in the fold-change R, as is shown in Appendix B. In actual experiments, tracking expression levels across a changing background, an intermediate value $\rho \approx 0.7$ is expected (Table 1), resulting in an intermediate value for the P_1 sensitivity divergence point.

8. CONCLUSIONS

A general noise model for the measurement of expression levels by microarrays was presented and then systematically used to derive a Bayesian scheme for estimating expression ratios. This scheme is currently implemented as the "PFOLD" algorithm. The PFOLD algorithm not only provides an estimate of the fold-change in expression and its confidence limits, but also assigns a P-value which gauges the significance of the change; in this respect, it is analogous to the BLAST algorithm for sequence matching, which also returns a P-value quantifying the significance of a result. The PFOLD output in turn generates a new, two-dimensional representation of the data in which one axis can be thought of as gauging quantity (the fold-change) and the other quality (significance through the P-value).

Monte Carlo simulations and cRNA spiking experiments indicate that the two-dimensional representation afforded by PFOLD is quite useful at the fundamental task of discriminating in a given experimental context

the genes with significant change from a large background of unchanging genes. Thus, at equal selectivity, use of the P-value as a decision statistic allows for markedly greater sensitivity than is obtained with the fold-change alone.

One consequence of using the PFOLD noise model is that noise parameters are carried alongside intensities on an equal footing, a state of affairs which will influence the choice of data structures and database schema in any set of expression analysis tools (GATC, 1998).

Current work-in-progress on PFOLD includes the full implementation of the correlated noise model presented in this paper but not yet fully integrated in the algorithm.

In summary, by providing a general and explicit noise model for microarray measurements, we have allowed for the systematic development of estimators of change and significance in gene expression, with the PFOLD algorithm as a specific, tested implementation.

APPENDIX A: BIAS AND VARIANCE OF THE ESTIMATOR FOR σ_ϵ

In Equation (16) we write $x = n + \epsilon$ and obtain the explicit dependence of the estimator $\hat{\sigma}_\epsilon$ on the random variable ϵ,

$$\hat{\sigma}_\epsilon^2 = \alpha^2(n^2 + 2\epsilon n + \epsilon^2) + \sigma_{bc}^2. \tag{61}$$

Taking the mean of Equation (61) and using $< \epsilon >= 0$, $< \epsilon^2 >= \sigma_\epsilon^2$, we obtain the second moment of $\hat{\sigma}_\epsilon$,

$$< \hat{\sigma}_\epsilon^2 >= (1 + \alpha^2)\sigma_\epsilon^2, \tag{62}$$

where σ_ϵ^2 is given by Equation (15). By taking the square root of Equation (61) and systematically expanding in a Taylor series in α^2, initially keeping terms up to and including $O(\alpha^6)$, and then taking the average, one obtains an expansion for the mean $< \hat{\sigma}_\epsilon >$. By combining this result with Equation (62), an expansion for the variance is obtained as well. The final results for bias and variance can be written relative to σ_ϵ and σ_ϵ^2, respectively, and take the form

$$\frac{< \hat{\sigma}_\epsilon > - \sigma_\epsilon}{\sigma_\epsilon} = \frac{\alpha^2}{2} - \frac{\alpha^4}{8}\left(\frac{4n^2}{\sigma_\epsilon^2} + 3\right) + O(\alpha^6), \tag{63}$$

$$\frac{< \hat{\sigma}_\epsilon^2 > - \sigma_\epsilon^2}{\sigma_\epsilon^2} = \alpha^4\left(\frac{n^2}{\sigma_\epsilon^2} + \frac{1}{2}\right) - \alpha^6\left(\frac{4n^2}{\sigma_\epsilon^2} + \frac{27}{8}\right) + O(\alpha^8). \tag{64}$$

These equations have two simple limits: 1) when $n \ll \sigma_{bc}$ (low concentration limit), $\sigma_\epsilon \to \sigma_{bc}$, and we have

$$\frac{< \hat{\sigma}_\epsilon > - \sigma_\epsilon}{\sigma_\epsilon} = \frac{\alpha^2}{2} + O(\alpha^4), \tag{65}$$

$$\frac{< \hat{\sigma}_\epsilon^2 > - \sigma_\epsilon^2}{\sigma_\epsilon^2} = \frac{\alpha^4}{2} + O(\alpha^6). \tag{66}$$

It can be seen that with values of α that are moderately small (say, $\alpha \lesssim 0.25$), the relative bias and variance are both small in this limit. 2) When $\alpha n \gg \sigma_{bc}$ (high concentration limit), $\sigma_\epsilon \to \alpha n$, and we have in turn

$$\frac{< \hat{\sigma}_\epsilon > - \sigma_\epsilon}{\sigma_\epsilon} = \frac{15}{8}\alpha^4, \tag{67}$$

$$\frac{< \hat{\sigma}_\epsilon^2 > - \sigma_\epsilon^2}{\sigma_\epsilon^2} = \alpha^2(1 - \alpha^2) + O(\alpha^6). \tag{68}$$

In this limit, the bias is small, and the largest uncertainty arises from the variance. Neglecting all terms above α^2, Equations (67) and (68) can be rewritten in terms of confidence limits

$$\sigma_\epsilon = \hat{\sigma}_\epsilon \pm \alpha\sigma_\epsilon. \tag{69}$$

Equation (69) shows that α is simply the measure of the fractional error in estimating σ_ϵ using $\hat{\sigma}_\epsilon$.

APPENDIX B: CORRELATED NOISE MODEL

The derivation of the a posteriori distribution of fold changes can be readily extended to include correlations between the noise terms. Once again we assume two intensity measurements

$$x_1 = n_1 + \epsilon_1, \tag{70}$$

$$x_2 = n_2 + \epsilon_2, \tag{71}$$

but in which the noise terms ϵ_1 and ϵ_2 are now jointly Gaussian, with zero means and variances σ_1^2 and σ_2^2 respectively, and a nonzero correlation coefficient ρ,

$$\rho = \frac{<\epsilon_1\epsilon_2>}{\sigma_1\sigma_2}. \tag{72}$$

Equation (28) now reads

$$f_R(R|x_1, x_2) = \int_0^\infty dn_1 \int_0^\infty dn_2 \delta\left(\frac{n_2}{n_1} - R\right) \mathrm{P}(n_1, n_2|x_1, x_2), \tag{73}$$

where $\mathrm{P}(n_1, n_2|x_1, x_2)$ is the joint probability distribution of n_1 and n_2,

$$\mathrm{P}(n_1, n_2|x_1, x_2) = \frac{\mathrm{P}(x_1, x_2|n_1, n_2)\mathrm{P}(n_1)\mathrm{P}(n_2)}{\mathrm{P}(x_1, x_2)}. \tag{74}$$

where $\mathrm{P}(n_1)$ and $\mathrm{P}(n_2)$ are the (independent) priors for the concentration, Equation (20), and where $\mathrm{P}(x_1, x_2|n_1, n_2)$ is the joint conditional probability distribution (Cowan, 1998, p. 34)

$$\mathrm{P}(x_1, x_2|n_1, n_2) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-n_1)^2}{\sigma_1^2} + \frac{(x_2-n_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-n_1)(x_2-n_2)}{\sigma_1\sigma_2}\right)\right). \tag{75}$$

The normalization term $\mathrm{P}(x_1, x_2)$ is obtained by integrating the numerator over n_1 and n_2, in analogy with the derivation followed in Equations (21–24).

The integration indicated in Equation (73) results in the distribution function for R in the form (dropping the explicit dependence on x_1 and x_2 in $f_R(R|x_1, x_2)$),

$$f_R(R) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}D(x_1, x_2)} \exp\left(-\frac{x_1^2(R-R_0)^2}{2(\sigma_2^2(1-\rho^2) + (R-\rho\sigma_2/\sigma_1)^2\sigma_1^2)}\right) J(x_1, x_2), \tag{76}$$

where $R_0 \equiv x_2/x_1$, and where $J(x_1, x_2)$ is defined by

$$J = \sigma_{12}'^2 \exp\left(-\frac{a_{12}'^2}{2\sigma_{12}'^2}\right) + a_{12}'(2\pi\sigma_{12}'^2)^{1/2}\frac{1}{2}\left(1 + \mathrm{erf}(a_{12}'/\sqrt{2}\sigma_{12}')\right), \tag{77}$$

where

$$\frac{1}{\sigma_{12}'^2} = \frac{1}{\sigma_1^2} + \frac{(R - \rho\sigma_2/\sigma_1)^2}{(1-\rho^2)\sigma_2^2}, \tag{78}$$

$$a_{12}' = \left(\frac{x_1}{\sigma_1^2} + \frac{(R - \rho\sigma_2/\sigma_1)(x_2 - \rho\sigma_2 x_1/\sigma_1)}{(1-\rho^2)\sigma_2^2}\right) \Bigg/ \left(\frac{1}{\sigma_1^2} + \frac{(R - \rho\sigma_2/\sigma_1)^2}{(1-\rho^2)\sigma_2^2}\right). \tag{79}$$

The normalization constant $D(x_1, x_2)$ is given by

$$D(x_1, x_2) = \frac{1}{4}\left(1 + \mathrm{erf}(x_1/\sqrt{2}\sigma_1)\right) + \frac{1}{2}\int_{-x_1/\sigma_1}^{\infty} \frac{dt}{(2\pi)^{1/2}} \exp(-t^2/2)\mathrm{erf}\left(\frac{x_2/\sigma_2 + \rho t}{\sqrt{2}(1-\rho^2)^{1/2}}\right). \tag{80}$$

Because $D(x_1, x_2)$ is a normalization constant, in a numerical computation the explicit evaluation of Eq. (80) can be replaced by a normalization step, following the evaluation of $f_R(R)$ in Equation (76) *without* the factor $1/D(x_1, x_2)$. Thus, the explicit calculation of the right-hand side of Equation (80) is not necessary in a computer program implementation of the estimation scheme.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Steven Perrin, who in the course of his own research gave extensive feedback on the use of PFOLD, and Dr. Michael Rosenberg and Dr. Anatoly Ulyanov for many scientific and technical comments regarding this work.

REFERENCES

Abramowitz, M., Stegun, I.A. 1972. *Handbook of Mathematical Functions*, Dover, New York.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403.

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.

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., and Davis, R.W. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. 1999. The transcriptional program of sporulation in budding yeast. *Science* 282, 699–705.

Cowan, G. 1998. *Statistical Data Analysis*, Clarendon Press, Oxford.

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.

Drake, A.W. 1967. *Fundamentals of Applied Probability Theory*, McGraw–Hill, New York.

Duda, R.O., and Hart, P.E. 1973. *Pattern Classification and Scene Analysis*, John Wiley, New York.

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* 95, 14863–14868.

Feller, W. 1966. *An Introduction to Probability Theory and its Applications, Vol. II*, John Wiley, New York.

Fodor, S.P.A., Rava, R.P., Huang, X.C., Pease, A.C., Holmes, C.P., and Adams, C.L. 1993. Multiplexed biochemical assays with biological chips. *Nature* 364, 555–556.

GATC, 1998. The current "GATC" schema for expression data does not provide for explicit noise terms to be carried in the database, on an equal footing with intensities (Genetic Analysis Technology Consortium, *GATC Software Specifications*, Version 1.0, May 15, 1998; available at http://www.gatconsortium.org/specifications.html).

Iyer, V.R., Eisen, M., Ross, D.T., Schuler, G., Moore, T., Lee J.C.F., Trent, J.M., Staudt, L.M., Hudson Jr., J., Boguski, M.S., Lashkari, D., Shalon, D., Botsetin, D., and Brown, P. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87.

Keeping, E.S. 1995. *Introduction to Statistical Inference*, Dover, New York.

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., and Brown, E.L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675–1680.

Nature Genetics, 1999. The chipping forecast. *Nature Genet.* 21, supplement.

Press, W., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 1997. *Numerical Recipes in C*, 2nd ed., Cambridge University Press, Cambridge.

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.

Van Trees, H.L. 1978. *Detection, Estimation, and Modulation Theory, Part I*, John Wiley, New York.

Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H., and Lockhart, D.J. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15, 1359.

Address correspondence to:
Joachim Theilhaber
Aventis Pharmaceuticals
Cambridge Genomics Center
26 Landsdowne Street
Cambridge, MA 02139

E-mail: joachimtheilhaber@aventis.com