JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 4, 2001
Mary Ann Liebert, Inc.
Pp. 443–461

Unfolding of Microarray Data

ANDREW B. GORYACHEV,[1] PASCALE F. MACGREGOR,[2] and ALED M. EDWARDS[2]

ABSTRACT

The use of DNA microarrays for the analysis of complex biological samples is becoming a mainstream part of biomedical research. One of the most commonly used methods compares the relative abundance of mRNA in two different samples by probing a single DNA microarray simultaneously. The simplicity of this concept sometimes masks the complexity of capturing and processing microarray data. On the basis of the analysis of many of our microarray experiments, we identified the major causes of distortion of the microarray data and the sources of noise. In this study, we provide a systematic statistical approach for extraction of true expression ratios from raw microarray data, which we describe as an unfolding process. The results of this analysis are presented in the form of a model describing the relationship between the measured fluorescent intensities and the concentrations of mRNA transcripts. We developed and tested several algorithms for inference of the model parameters for the microarray data. Special emphasis is given to the statistical robustness of these algorithms, in particular resistance to outliers. We also provide methods for measurement of noise and reproducibility of the microarray experiments.

Key words: cDNA microarrays, gene expression, statistical data analysis.

1. INTRODUCTION

DNA MICROARRAYS ARE USED TO EXPLORE gene expression on a genome-wide scale (Brown and Botstein, 1999; Duggan *et al.*, 1999). By quantifying the relative abundance of thousands of mRNA transcripts simultaneously, researchers can discover new functional relationships among genes (Wen *et al.*, 1998) and observe the response of whole genomes to various experimental perturbations (Iyer *et al.*, 1999; Spellman *et al.*, 1998). Microarrays have also found extensive application in medicine and pharmacology (Debouck and Goodfellow, 1999). For example, microarrays can be used to identify genes whose pattern of expression distinguishes various types of cancer (Alizadeh *et al.*, 2000; Golub *et al.*, 1999).

There are two major types of DNA microarray technologies that can be broadly termed as one-channel and two-channel. The one-channel measures the absolute concentrations of mRNA transcripts. The two-channel estimates the relative abundance between a sample and a control specimen. In this second method, which is the focus of this paper, fluorescently-labeled cDNA is prepared from the two biological sources that are to be compared. The two cDNA populations are labeled with different fluorescent dyes, pooled, and simultaneously cohybridized to thousands of different DNA molecules, which are arrayed on a modified

[1] Ontario Cancer Institute, Princess Margaret Hospital, Toronto, Canada.
[2] C. H. Best Institute, University of Toronto, Toronto, Canada.

glass slide. The fluorescence from the glass slide is measured at the two wavelengths and the individual microarray spots are identified using image-processing algorithms. The background fluorescent intensity can then be subtracted from the intensity of each spot. Finally, the two data sets, one for each fluor, are scaled to each other and the normalized intensities at each spot are compared to generate a list of genes that are differentially expressed.

The success of DNA microarray analysis is critically dependent on the quality and precision of the data, but the miniaturization and massive parallelization involved pose challenging physical and mathematical problems for data acquisition and subsequent analysis (Chen *et al.*, 1997; Claverie, 1999; Schuchhardt *et al.*, 2000). Because gene expression levels in cells can vary by several orders of magnitude, it is necessary to measure signals over a wide dynamic range. In microarray experiments, it is common to observe both saturated signals and signals that are lost in the background noise. For low intensity spots, distinguishing the targets from the background poses significant challenges and inevitably results in large measurement error. The integration and scaling of the two data sets are also difficult for two reasons. First, the two commonly used flors, Cy3 and Cy5, have different physical properties. They are differentially incorporated into the cDNA, and the quantum fluorescent yields are different. Second, the microarray spots are widely separated with respect to their individual size and thus spots may be hybridized, washed, and scanned in different conditions depending on their location on the array.

There are various algorithms that accomplish the target separation, background correction, and normalization steps. For the DNA microarray applications, these methods are commonly disseminated by means of web sites or through private communication. A number of recently developed software packages also provide numerous options for image quantification, background correction, and normalization. The difficulty in selecting the optimal suite of programs for data reduction stems from the fact that the underlying algorithms are rarely known and their relative performance is difficult to assess. In addition, many of the methods have not been sufficiently substantiated by statistical analysis. In particular, the statistical robustness of background correction and normalization methods has not received adequate attention since the pioneering paper by Chen *et al.* (1997).

In this paper we provide a systematic statistical approach to dealing with raw DNA microarray data. We introduce a model that describes a relationship between the measured fluorescent signals and the mRNA concentrations. This model stems from the analysis of many of our microarray experiments, including several that were designed specifically to assess contribution of factors that complicate microarray analysis. We show that one can extract true expression ratios from the raw microarray data by statistically estimating the parameters of the model. In broad terms, our technique is analogous to the method known in statistical data analysis as unfolding (Cowan, 1998), and our work presents a unified framework for the unfolding of microarray data.

2. MODEL FOR MEASURED EXPRESSION RATIOS

The goal of this section is to introduce a model that describes the relationship between the intensities of fluorescent signals and the concentrations of mRNA transcripts. To achieve this, we must account for the sequence of chemical and physical processes that take place between the RNA isolation and the quantification of the microarray image. In our notations, let the RNA messages for the i-th gene be present at concentrations c_i^s and c_i^c in the sample and control specimens, respectively. For the sake of clarity, we temporarily assume that the sample is labeled with Cy3 ("green") and the control with Cy5 ("red"). The other labeling order is possible and often necessary to determine the systematic effects introduced by the type of flor (see Section 3.3). Therefore, it is important to distinguish sample and control probes from "red" and "green" channels.

If the reverse transcriptase incorporates on average γ_i molecules of dye into the cDNA copy of the i-th transcript, then after labeling the i-th gene is represented by $\gamma_i^G c_i^s$ optically detectable molecules in the sample and $\gamma_i^R c_i^c$ in the control. In practice, the efficiency with which the dye is incorporated γ_i^G, γ_i^R depends on the conditions of the reverse transcription (RT) reaction, the ternary structure of mRNA, and the gene nucleotide sequence. The quantity γ_i^G is not generally equal to γ_i^R due to the difference in molecular properties of the dyes, and the preferential incorporation of the dyes into specific cDNAs can be reproducibly shown in experiment.

After the hybridization of the labeled cDNA to the DNA on the microarray surface, the resulting fluorescence is quantified using fluorescence microscopy. The intensity $I_i^G(\boldsymbol{r})$ of the green fluorescent signal from the i-th gene spotted at location $\boldsymbol{r}$ on the array is proportional to $\gamma_i^G c_i^s$ with a coefficient A^G. The scaling factors A^G, A^R incorporate the quantum yield of dyes (i.e., efficiency of emission) and the sensitivity of the scanner. Since the excitation and detection of fluorescence are performed separately for red and green channels, the fluorescent properties of dyes are different, A^R and A^G are not identical. In fact, they can vary widely depending on the scanner parameters (e.g., laser emission power and photomultiplier voltage).

In some experimental layouts, A^R and A^G depend on $\boldsymbol{r}$. This dependence can be explained in part by the spatial microheterogeneity of the hybridization conditions and also the warped shape of the glass slide. The spatial heterogeneity of A^R and A^G may also arise as a consequence of the optical design of the scanner if it is based on confocal microscopy. The typical focal depth of a confocal scanner is 25 μm. Given the dimensions (25 × 75 mm) of a typical microarray glass slide, even infinitesimally small angles of 10^{-3} in the position of the slide during scanning could result in the gradual loss of focus as the scanning beam sweeps across the array. This imperfection manifests itself in $A(\boldsymbol{r})$ gradients observed in top–bottom and left–right directions on the array (see Section 3.2 for discussion).

The description of the microarray signal intensities is complete when stochastic terms $\xi_i^R(\boldsymbol{r})$, $\xi_i^G(\boldsymbol{r})$ representing all noisy contributions to $I_i(\boldsymbol{r})$ are added. They take into account the cumulative effects of chemical, optical, and computational factors that are introduced by the microarray technology. One of the major challenges in the analysis of the microarray data stems from the fact that assessing and controlling the relative contributions of the many sources of noise is difficult. Chemical factors, such as those introduced during RNA preparation, labeling, hybridization, and washing, contribute to the ξ terms in two ways. Most of the variability derives from incomplete hybridization or nonspecific binding to the array. The nonspecific binding appears to be more of a problem for genes with a higher content of low-complexity sequence motifs (hence subscript "i" in the $\xi_i^R(\boldsymbol{r})$, $\xi_i^G(\boldsymbol{r})$ terms). The issues and complexities of correction for the background signal, whose nature is primarily defined by the chemical factors, require special attention and are considered in a greater detail in Section 3.1. Optical noise derives from factors intrinsic to fluorescence microscopy (Inoue and Spring, 1997). For example, the background current from the photomultiplier tube (PMT), the "dark current," complicates the analysis of low-intensity spots. Computational variability arises from errors introduced by the quantification algorithms. For example, distinguishing the pixels in a target from those in the background is a challenging image recognition problem since the microarray spots often have low intensity and imperfect morphology.

Finally, assembling all introduced terms as shown in Fig. 1, we can present the observed green/red signal ratio, $T_i(\boldsymbol{r})$, as

$$T_i(\boldsymbol{r}) = \frac{I_i^G(\boldsymbol{r})}{I_i^R(\boldsymbol{r})} = \frac{A^G(\boldsymbol{r})\gamma_i^G c_i^s + \xi_i^G(\boldsymbol{r})}{A^R(\boldsymbol{r})\gamma_i^R c_i^c + \xi_i^R(\boldsymbol{r})}. \tag{1}$$

This equation demonstrates that the observed value $T_i(\boldsymbol{r})$ can be a fairly distorted representation of the actual ratio of sample and control mRNA concentration $T_i^a = c_i^s/c_i^c$, which is the final goal of the microarray experiment.

3. THE MAIN STEPS OF THE UNFOLDING

The model introduced in the previous section provides a theoretical framework for the unfolding of microarray data. Our model explicitly takes into account the most important factors that are thought to distort the true expression ratios T_i^a. For the practical analysis of microarray data, it is important to distinguish distorting factors that can be inferred from a single microarray from those factors that require multiple repetitions of the microarray experiment. The first category is mainly constituted by factors that are independent of the gene and vary continuously with location on the array (e.g., scaling factors $A^{G,R}(\boldsymbol{r})$, whose dependence on $\boldsymbol{r}$ is defined by conditions of hybridization and distance to the scanner lens). The second category consists of factors that vary markedly with both gene and location, such as nonspecific binding and quantification errors.

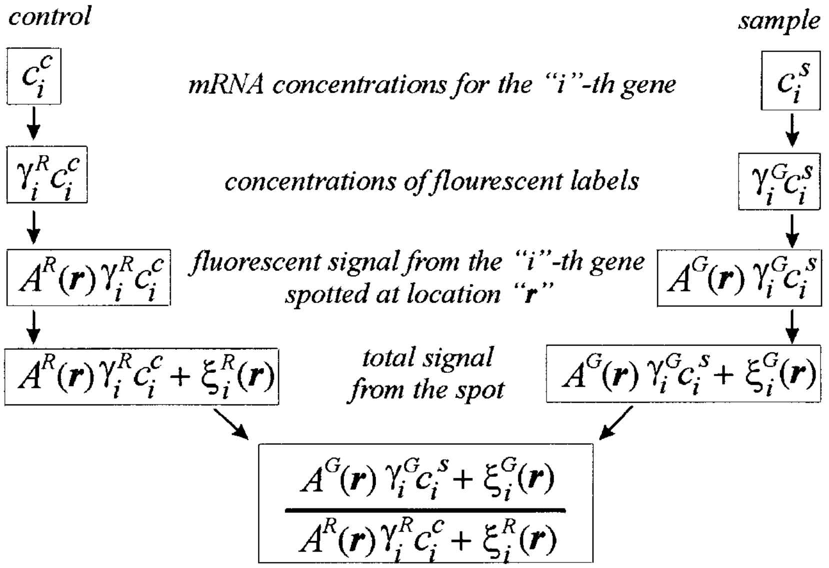
control
sample
c_i^c
mRNA concentrations for the "i"-th gene
c_i^s
$\gamma_i^R c_i^c$
concentrations of flourescent labels
$\gamma_i^G c_i^s$
$A^R(\boldsymbol{r})\gamma_i^R c_i^c$
fluorescent signal from the "i"-th gene spotted at location "r"
$A^G(\boldsymbol{r})\gamma_i^G c_i^s$
$A^R(\boldsymbol{r})\gamma_i^R c_i^c + \xi_i^R(\boldsymbol{r})$
total signal from the spot
$A^G(\boldsymbol{r})\gamma_i^G c_i^s + \xi_i^G(\boldsymbol{r})$
$\frac{A^G(\boldsymbol{r})\gamma_i^G c_i^s + \xi_i^G(\boldsymbol{r})}{A^R(\boldsymbol{r})\gamma_i^R c_i^c + \xi_i^R(\boldsymbol{r})}$

FIG. 1. Diagram illustrating the derivation of the model (1) for the measured ratio of fluorescent intensities.

Before going into the detailed discussion of individual steps, we give an overview of the entire unfolding procedure and stipulate the limits of the unfolding set forth by the noise factors. From the form of the model (1), it follows that one should remove the additive terms $\xi_i^{R,G}(\boldsymbol{r})$ as the first step towards extracting T_i^a from the experimental value $T_i(\boldsymbol{r})$. However, it is generally impossible to find exact estimates for the $\xi_i^R(\boldsymbol{r})$, $\xi_i^G(\boldsymbol{r})$ terms for each spot due to the stochastic nature of factors contributing to the noise terms. A closer look at the raw microarray data shows that the $\xi_i^{R,G}(\boldsymbol{r})$ terms possess a significant systematic component from the background fluorescence. For example, in Fig. 2A, we show that additive background terms can profoundly affect the measured ratios. As the signal intensity diminishes, one observes deviation of the expression data from the straight line, which reflects the growing influence of the $\xi_i^G(\boldsymbol{r})$ terms. At low intensities, the numerator of (1) is dominated by $\xi_i^G(\boldsymbol{r})$, whose average value saturates while the denominator diminishes.

We can calculate an estimate for the systematic components of the background signal, e.g., the most probable values per channel, $B^R(\boldsymbol{r})$, $B^G(\boldsymbol{r})$, from the raw data. The "truly random" zero-centered noise can be obtained by subtracting the systematic components $B^R(\boldsymbol{r})$, $B^G(\boldsymbol{r})$ from the terms $\xi^R(\boldsymbol{r})$, $\xi^G(\boldsymbol{r})$.

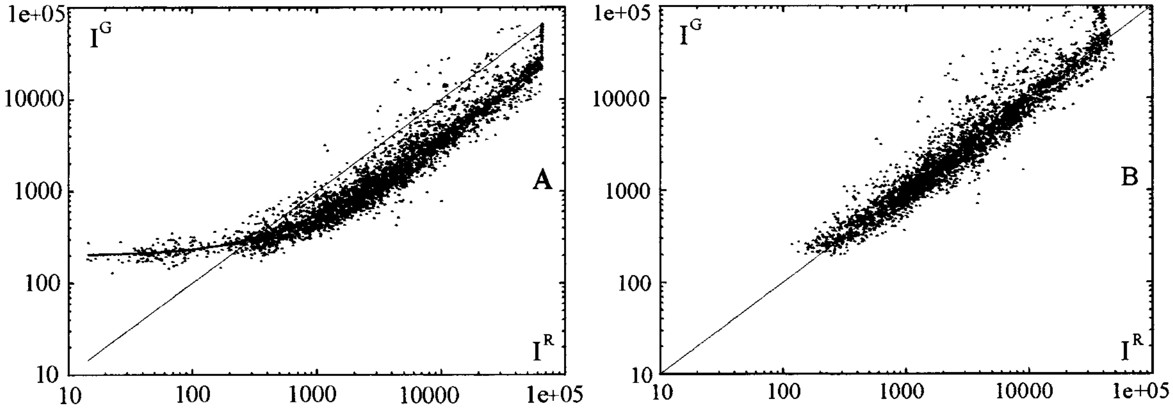

FIG. 2. Unfolding of the microarray data. (**A**) Generic raw data, fitted to the data equation $I^G = a + bI^R$, is shown by the solid line. (**B**) Same data after correction for the background and normalization.

This residual noise contributes to the scatter of the expression points but does not lead to their systematic deviation from a straight line. We now can rewrite model (1) in the form

$$T_i(\boldsymbol{r}) = \frac{A^G(\boldsymbol{r})\gamma_i^G c_i^s(1+\varepsilon_i^G)}{A^R(\boldsymbol{r})\gamma_i^R c_i^c(1+\varepsilon_i^R)},$$

where the noise terms are moved inside the product. Here we assume that the statistical properties of the residual noise do not depend on location on the microarray. Since ε_i^G, ε_i^R are purely stochastic, their contribution cannot be filtered out. Therefore, as the result of the unfolding of a single microarray, we find the effective expression ratios T_i^e, which are equal to $\tilde{c}_i^s/\tilde{c}_i^c$, $\tilde{c}_i = c_i(1 = \varepsilon_i)$. By simplifying the previous equation further, we finally get

$$T_i(\boldsymbol{r}) = \frac{A^G(\boldsymbol{r})\gamma_i^G}{A^R(\boldsymbol{r})\gamma_i^R} \cdot T_i^e.$$

After background correction, the entire set of expression data is stretched on the plane (I^G, I^R) along a certain line which is parallel to the main diagonal and offset by the ratio of scale factors $A^G(\boldsymbol{r})$ and $A^R(\boldsymbol{r})$. We can remove systematic distortion of expression ratios due to the scale factors $A^G(\boldsymbol{r})$, $A^R(\boldsymbol{r})$ and center the data points around the line $I_0^R = I_0^G$ (Fig. 2B) by applying one of the normalization methods described below.

The last step of the unfolding is the correction for the differential dye incorporation represented by the γ_i^G, γ_i^R terms. Since incorporation coefficients do not depend on the location and vary widely from gene to gene, they cannot be inferred from a a single microarray experiment. A special experimental procedure involving the two possible variants of labeling and known as "dye switch" or "dye flip" is required to provide the data necessary to estimate the incorporation coefficients.

Finally, the unfolding transformation, which allows us to calculate effective expression ratios T_i^e as the best approximation for the actual ratio T_i^a, can be symbolically written as

$$T_i^e = \frac{\gamma_i^R A^R(\boldsymbol{r})(I_i^G(\boldsymbol{r}) - B^G(\boldsymbol{r}))}{\gamma_i^G A^G(\boldsymbol{r})(I_i^R(\boldsymbol{r}) - B^R(\boldsymbol{r}))}. \qquad (2)$$

In practice, the transformation (2) is calculated using the numeric algorithms presented in the following subsections. Our method accounts for a number of practical problems associated with background correction, normalization, and correction for the differential labeling. We describe these problems in the following sections.

3.1. Background correction

The treatment of background is possibly one of the most controversial and the least explored problems of microarray data analysis. Before discussing strategies for background correction, it is important to define how the target and the background signals are determined in the course of the image quantification. In this study, we assume that the quantification software performs image segmentation (i.e., separation of different targets) and integrates *all* pixels covered by a mask of regular shape and constant size. The size of the mask is chosen sufficiently large to encompass the entire spot and possibly some of the surrounding area. The intensity of the target is then defined as the integral pixel count averaged over the mask. The intensity of the local background is calculated in the area surrounding the target mask so that target masks of the neighboring spots are excluded. This quantification method ensures that pixels belonging to the spots are not attributed to the background and the background level is not overestimated.

Many image quantification packages provide an option to automatically subtract the local estimate of the background from the intensity of the corresponding spot. This method of background correction is not generally recommended, for two reasons. Firstly, typical background signal demonstrates significant variation with location on the array. Local background estimates may be affected by small-scale fluctuations with high amplitude and their subtraction might result in additional noise in calculated expression ratios.

Secondly, substantial systematic error can be incurred in regions of the microarray which are obviously defective and characterized by unusually high level of background (e.g., regions with fluorescent smears). In such regions, it is not uncommon to observe spots that have intensities lower than the surrounding background. It is clearly inappropriate to subtract the background in such areas. Instead, they must be identified and processed separately from the rest of the array.

Thus, the treatment of background can be split conceptually into two problems: to locate areas of the microarray occupied by defects with abnormal background properties and to find the optimal estimates for the background correction factors $B^R(\boldsymbol{r})$, $B^G(\boldsymbol{r})$ for spots which are not affected by defects. The defects, which commonly result from slide coating failures, chemical stains, smudged spots, and dust particles, are easily detectable by a human eye. However, identification of defects presents a significant challenge for image quantification software and is not even attempted by many programs. We developed statistical means to separate regions of homogeneous background from those that harbor significant defects. Our method is based on the properties of the probability density function for the local background counts on the microarray.

Figure 3A shows probability density functions for the logarithm of background intensity (PDFL) calculated for three typical background samples (see Section 5.2 for sample definition). The logarithmic transformation (see Section 5.1) was used to facilitate subsequent application of numeric algorithms relying on quasi-symmetry and quasi-normality as the original probability density functions (PDF) for the intensities were found strongly asymmetric and skewed to the right. The three samples represent three consecutive degrees of contamination by the defects. Case ***a*** is an "ideal" background without inhomogeneities and defects, while ***c*** contains significant defects, such as extended bright blotches. The presence of the defects can be inferred from the form of the right tail of the respective PDFLs. Thus, in case ***a***, the PDFL vanishes faster than that for ***b***. In case ***c***, the defects are so prominent that they form a second maximum at very high intensities. Further insight into the statistical properties of the background can be gleaned from the normal quantile–quantile plots, which are shown for the same three background samples in Fig. 3B. Remarkably, the PDFL for the normal homogeneous background (***a***) shows only minor deviation from the normality. Samples ***b*** and ***c*** clearly demonstrate an earlier departure from the normality, which can be attributed to the presence of the defects. Therefore, we assume that the "ideal" background is described by a normal PDFL (log-normal PDF) with possibly slightly different standard normal deviations for the right (σ^+) and left (σ^-) tails.

These observations can be employed to resolve areas with homogeneous normal background from defective areas. The aim is to seek an upper estimate for confidence limits $\overline{I}_B^G$, $\overline{I}_B^R$ such that if the local background intensity exceeds it, then, with a certain confidence (e.g., 95%), the corresponding area is deemed defective. To find upper confidence levels $\overline{I}_B^G$, $\overline{I}_B^R$, one needs to estimate the parameters (μ, σ^+) for the right-side normal fit of the respective empirical PDFLs. Since experimental PDFLs are approximated

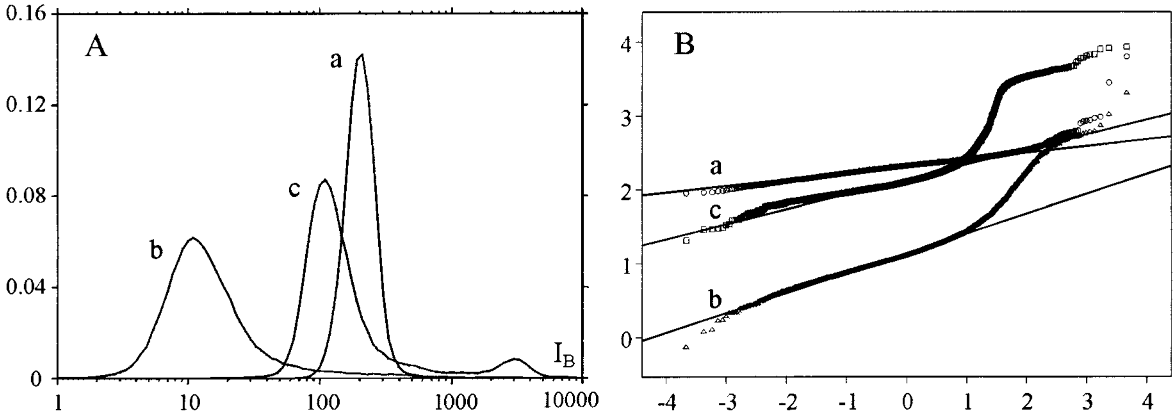

FIG. 3. Statistical properties of the background. (**A**) Probability density functions for the logarithm of the background intensity (base 10). (**B**) Normal quantile–quantile plots for the PDFLs shown on the left panel.

by a normal function only near their mode, it is important to use robust methods to estimate the parameters. Thus, we found the least trimmed squares (LTS) estimator m_{LTS} (Rousseeuw and Leroy, 1987) useful for calculation of the mode μ. In the LTS method, the squared regression residuals are computed as in the usual least squares method. However, only a fraction α of the smallest residuals is retained and used for minimization. In this treatment, m_{LTS} is resistant to the presence of up to $(1 - \alpha) \cdot 100\%$ of outliers. Given robust estimates for the parameters (μ, σ^+) and applying the "rule of two sigma" which gives 95% confidence for a normal distribution (Snedecor and Cochran, 1973), for the upper confidence limit $\overline{I}_B$ we finally get

$$\overline{I}_B = \exp(\mu + 2\sigma^+).$$

After setting aside the defective areas, the background correction factors $B^{G,R}(\boldsymbol{r})$ are calculated for the spots in regions of homogeneous background. We argued above that the local background found by the image quantification software can exhibit undesirably high variability due to the abrupt spatial fluctuations in background intensity. To avoid this problem one can calculate $B^{G,R}(\boldsymbol{r})$ as the median background intensity computed for some neighborhood centered at $\boldsymbol{r}$ with size δ. The larger the δ, the higher is the confidence that $B^{G,R}(\boldsymbol{r})$ will not be affected by the outliers. At the same time, the choice of excessively large δ can result in the loss of important information on the spatial variation of the background. Therefore, it is necessary to find the optimal size of the sampling neighborhood so that $B^{G,R}(\boldsymbol{r})$ is both outlier resistant and sufficiently sensitive.

To achieve this, we investigated the spatial correlation of the background using Fourier analysis (see Section 5.2). Figure 4 shows typical Fourier power spectra in which the ordinate value P_k represents the contribution of the spatial frequency k/L ($k = 0, \ldots L/2$; $L = 200$). The sharp maxima at low k indicate the presence of significant long-range background gradients. For $k > k^* \approx 20$, the spectrum is essentially flat and the equal contributions of high frequencies create noise on top of the long-range spatial modes. Therefore, it is reasonable to accept $\delta = L/k^*$ as the optimal size of the sampling neighborhood. For the example shown in Fig. 4, we find that δ approximately corresponds to six distances between spots. Thus, neighborhoods of 5 × 5 or 7 × 7 spots are optimal for the background sampling.

3.2. *Normalization*

In practice, the scale factors $A^G(\boldsymbol{r})$ and $A^R(\boldsymbol{r})$ described by our model cannot be directly extracted from the experimental data. Therefore, to obtain the normalization factor $A^R(\boldsymbol{r})/A^G(\boldsymbol{r})$, one needs additional information or a biological hypothesis. This hypothesis is often formulated as follows: assuming that certain genes should not change the expression level, find the normalization factor such that the expression ratios of these genes are indeed close to one. Then the expression ratios for the rest of the genes are calculated relative to the baseline established by the "constant" genes. These assumptions hold for some experimental

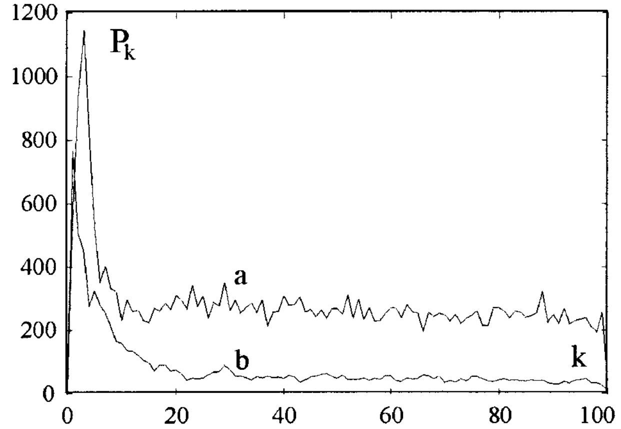

FIG. 4. Fourier power spectra computed for the background samples a and b whose PDFLs are shown in Fig. 3.

conditions but fail in the others. The full set of conditions for which the method holds defines its domain of biological validity. In this section, we consider four widely used normalization methods and discuss their applicability. The fitness of a normalization method is defined by the breadth of experimental conditions for which it is valid and the extensibility of the method for possible spatial dependence of the normalization factor.

It is a common practice to assume that the normalization factor is independent of the location. This assumption is not always true and needs to be validated for each experimental setup. An example of a microarray experiment that displays spatial dependence of the normalization factor is shown in Fig. 5. In this experiment, an RNA sample was divided into two equal aliquots. One was reverse-transcribed with the Cy3 label and the other with Cy5. The two labeled cDNA samples were cohybridized to a microarray. Since the sample and the control are biologically identical, each ratio in this experiment is expected to be equal to or close to one. The observed ratios were averaged over subarrays, or grids, comprising 400 spots and the results were plotted as a 2D function of the corresponding grid indexes. One sees the systematic dependence of average ratios on the location that takes the form of the pronounced gradient such that the normalization factor varies from 0.7 on the left side of the array to 1.4 on the right. This example demonstrates that the spatial dependence of the normalization factor cannot be simply dismissed. None of the existing normalization methods that we describe in the following subsections addresses these issues. Where possible, we provide an extension of the existing normalization methods to include the case of spatial dependence.

The method of housekeeping genes. The concept of housekeeping genes emerged from the observation that some genes involved in basic cell metabolism seem to be insensitive to many experimental perturbations. In this method, one seeks the normalization factor that minimizes the distance between the ratios of housekeeping genes and unity. A detailed mathematical treatment for normalization using 78 putative housekeeping genes was developed by Chen and coworkers (Chen *et al.*, 1997). Unfortunately, in any individual experiment, identifying such genes is a nontrivial problem and different experimental groups often arrive at different lists of candidate housekeeping genes. Recent advances in high-throughput expression profiling have resulted in erosion of the concept of “housekeeping genes,” as many genes earlier thought to be constant were shown to change their expression under some experimental conditions (Schuchhardt *et al.*, 2000). It should also be noted that the approach of housekeeping genes is valid only in experiments in which cells reach a stable steady state. In extreme conditions that lead to massive irreversible changes in cellular homeostasis and often to cell death, the concept of housekeeping genes is inappropriate. Practical application of the housekeeping gene method is also complicated by the fact that it cannot be easily extended for the case of spatial dependence of the scale factors $A^G(\boldsymbol{r})$, $A^R(\boldsymbol{r})$.

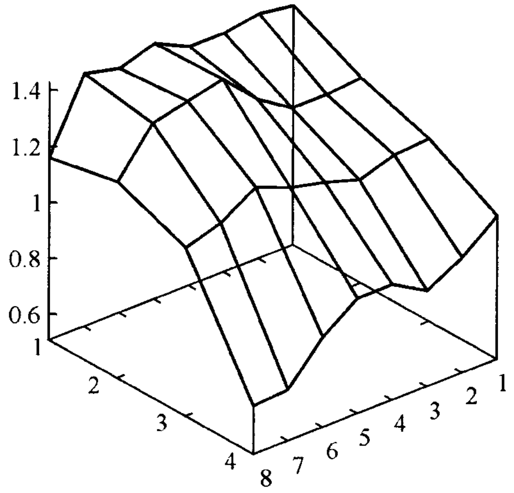

FIG. 5. Spatial dependence of the scaling factors A^G, A^R. Averaged per grid expression ratios (solid line) obtained for a yeast microarray are shown as a function of the grid indexes (4 × 8).

The method of control spots. Another method of normalization, in which one arrays an unrelated DNA and adds exogenous RNA to the sample, overcomes some of the problems of the method of housekeeping genes. Typically, a DNA species (gene) Q or a group of genes Q_i whose sequence is sufficiently dissimilar from all those under study is spotted on the microarray at specific locations. Before labeling, equal amounts of Q mRNA are added to both sample and control. In this treatment, Q essentially becomes a "housekeeping gene" with a guaranteed actual "expression" ratio of $T_Q^a \equiv 1$. In practice, the observed ratios of the control spots, T_Q^e, are scattered around 1 and the presence of several control spots is necessary to ensure statistical accuracy.

By using spatially compact groups of control spots, one is able to remove the exponential spatial dependence of the scaling factors. To demonstrate this, consider a ratio $T_i(\mathbf{r}) = A^G(\mathbf{r})\gamma_i^G / A^R(\mathbf{r})\gamma_i^R \cdot T_i^e$ that represents gene i spotted at location $\mathbf{r}$ and the average ratio $T_Q(\mathbf{r}') = A^G(\mathbf{r}')\gamma_Q^G / A^R(\mathbf{r}')\gamma_Q^R$ observed for a nearby group of control spots. Assuming the $\mathbf{r}$ and $\mathbf{r}'$ are in close proximity and $A(\mathbf{r}) \approx A(\mathbf{r}')$, we get

$$\frac{T_i(\mathbf{r})}{T_Q(\mathbf{r}')} = \frac{\gamma_i^G \gamma_Q^R}{\gamma_i^R \gamma_Q^G} T_i^e.$$

This relationship demonstrates that normalization using control spots can be employed to resolve the spatial dependence; however, the ratio baseline may be systematically offset because of two factors. First, all expression ratios are divided by an unknown factor $\Gamma_Q = \gamma_Q^G / \gamma_Q^R$ which may be arbitrarily far from 1. Additional offset can arise if there is a systematic preference for incorporation of Cy5 or Cy3, for example, if under some conditions of the labeling reaction $\gamma_i^G > \gamma_i^R$ for most genes.

The method of constant majority. The method of constant majority assumes that the majority of genes do not change their expression level in response to the experimental perturbation. This, however, does not imply that the differentially expressed genes must constitute only a small fraction of all genes. As we show below, under appropriate formalization, this method is valid even if up to 50% of the genes are differentially expressed. The use of the method appears to be appropriate under the same biological restrictions as apply to the method of housekeeping genes. Both methods presume that the cell responds to the experimental treatment without a total disruption of gene expression. However, the method of constant majority is more flexible because it does not require that a certain subset of genes remains unchanged under all possible experimental conditions. In addition, one does not need to know in advance *which* genes have not changed their expression level. In this section, we first discuss the method of constant majority in a simplifying assumption of the independent-of-space normalization factor. Then we demonstrate how it can be extended for the case of spatial dependence.

The probability density function for the distribution of ratios can be used to formalize the underlying assumptions of the method of constant majority. Intuitively, it is obvious that since the majority of genes do not change their expression level, the ratio baseline lies somewhere in the vicinity of the PDF mode. The rigorous treatment was developed by Chen *et al.* (1997). Chen and colleagues developed an analytic expression for the PDF of ratios exhibited by the constant genes. It was assumed that the intensity of the fluorescent signals on both channels $I_i^G = \overline{I}_i^G + \varepsilon_i^G$ $(I_i^R = \overline{I}_i^R + \varepsilon_i^R)$ is measured with normally distributed errors ε_i^G, ε_i^R. It was also postulated that the corresponding coefficient of variation C is independent of the dye type and is constant for the entire set of genes. Despite the fact that $\overline{I}_i^G = \overline{I}_i^R$, due to the measurement noise $\varepsilon_i^{G,R}$, the expression ratio of a "constant" gene is a stochastic variable and is described by an asymmetric, skewed-to-the-right PDF. The surprising conclusion of their study was that the mode μ of the PDF satisfied the inequality $\mu < 1$ for all $C > 0$. Therefore, contrary to the intuitive expectation, the correct normalization factor m (in this case $m \equiv 1$) is not equal to μ. Moreover, the error incurred by normalizing the PDF mode to unity may reach 10–15% for realistic values of C. On the basis of their model, Chen and colleagues also developed an iterative normalization method that requires the parameter C to be estimated from the data. The estimation procedure explicitly relies on the subset of genes which are *known* to be constant (i.e., the method requires "housekeeping" genes).

Here we demonstrate that, with some alteration, the model of Chen and colleagues can be used to devise a normalization method that does not require estimation of C and does not rely on the housekeeping genes.

We transform the PDF defined by Chen *et al.* (1997) into the corresponding PDFL using logarithmic transformation of ratio $z = \ln t$ (see Section 5.1) and obtain

$$f(z) = \frac{e^Z(1+e^Z)\sqrt{1+e^{2Z}}}{\sqrt{2\pi}C(1+e^{2Z})^2} \cdot \exp\left(-\frac{(e^Z-1)^2}{2C^2(1+e^{2Z})}\right). \tag{3}$$

Both the pre-exponential factor and the exponential itself are even functions of z and, therefore, $f(z)$ is a symmetric function with both mean and mode equal to zero, independently of the value of C. For small z ($|z| \ll 1$), the Taylor expansion $e^Z \approx 1 + z$ can be applied. This gives

$$f(z) \approx f^*(z) = \frac{1}{\sqrt{2\pi} \cdot \sqrt{2}C} \cdot \exp\left(-\frac{z^2}{4C^2}\right), \tag{4}$$

a normal distribution with $\sigma = \sqrt{2}C$. Although this estimate is strictly valid only in a small neighborhood of the origin, numerical calculations show that both functions remain close in L_1 norm on the entire z axis (i.e., $\int_{-\infty}^{+\infty} |f(z) - f^*(z)| dz < \delta$).

These properties of the PDFL can be exploited to great advantage. First, for a set of genes whose expression is constant, it implies that the value for the normalization factor given by the exponent of the PDFL mode (same as the mean) is equal to the geometric mean m_g of the expression ratios. Second, for practically relevant experiments involving differentially expressed genes, it allows one to apply a number of well-developed computational methods that explicitly rely on symmetry and normality of the probability density function in question. Since in the presence of differentially expressed genes the actual PDFL is no longer strictly symmetric, the mode should be located by one of the robust techniques, e.g., the method of least trimmed squares (LTS). For the symmetric probability density function, the mode is resistant to up to 50% of outliers (Rousseeuw and Leroy, 1987). Therefore, in the best case in which the numbers of up- and down-regulated genes are approximately equal and the quasi-symmetry of the ratio PDFL is preserved, the breakdown point for the method of constant majority can be estimated as 50%.

The method of constant majority can be extended to accommodate the spatial dependence of $A(\boldsymbol{r})$ in the following way. The underlying assumption of the method, if true for the entire set of genes (population), should also hold for sufficiently large subsets (samples). Suppose that the entire microarray is partitioned into subdomains with size L and number of spots N. Then, if $N \gg 1$ and the gradients of $A(\boldsymbol{r})$ are small on the scale of the domain ($\nabla A(\boldsymbol{r}) \cdot L < 1$), we can perform the normalization independently inside each domain. Here we also assume that genes are not spotted on the array in any particular order, e.g., according to functional categories, and any spatial partition would represent a statistically independent sample of the entire population.

As in the case of background correction, the optimal size of the partition is a tradeoff between spatial specificity and statistical significance. An estimate for the size of the partition can be derived from the following consideration. According to (4), let the logarithm of ratio $z = \ln t$ be normally distributed with population mean $\mu = \mu_0$ and standard deviation σ. Then a partition (sample) mean $m = \sum z_i/N$ can be applied to test the hypothesis H_0, whether the sample belongs to the population. If H_0 is true, the observed difference between m and μ is statistically insignificant and no correction for spatial dependence is necessary. The acceptance interval for H_0 with confidence $(1 - \alpha) \cdot 100\%$ is given by Walpole *et al.* (1998):

$$|m - \mu_0| < Z_{\alpha/2}\frac{\sigma}{\sqrt{N}},$$

where $Z_{\alpha/2}$ is the standard normal distribution z-value for $\alpha/2$ ($Z_{\alpha/2} = 1.96$ for $\alpha = 0.05$). The optimal number of spots per partition N can now be estimated as a function of maximum allowable tolerance $\delta = |m - \mu_0|$ as

$$N \approx \left(\frac{Z_{\alpha/2}\sigma}{\delta}\right)^2.$$

If we support that $\sigma = 0.3$, $\delta = 0.05$, and $\alpha = 0.05$ (95% confidence), then the desired N is 144. Recalling that a significant proportion of spots (e.g., 60%) may have poorly defined ratios due to their low intensity and are not suitable for normalization, we arrive at an estimate of ≈ 400. This number corresponds to a typical microarray grid consisting of 20×20 spots. Thus, this simple estimate corroborates the use of microarray grids as normalization domains.

The method of integral balance. Another normalization method is based on the assumption that the total levels of gene expression in the sample and the control are the same. Prior to the labeling reaction, the total amounts of RNA in sample and control are usually equalized. Therefore, one expects that after correct normalization the integral intensity of all Cy3 signals should be equal to that of Cy5 signals. The normalization factor is thus defined as the ratio of the total sums of signal intensities on the two channels. In this treatment, differentially expressed genes correspond to those that change their relative contribution to the integral signal. An important caveat to this method is that, in a microarray experiment, the total amount of RNA isolated from the cells is equalized, but only mRNA levels are measured. All experimental conditions under which cells are expected to significantly increase or decrease the total level of mRNA transcription should be considered as potentially problematic. The extension of the method of the integral balance for the case of a spatially dependent normalization factor can be achieved in the same way as for the method of constant majority. The normalization factors are computed independently for the subdomains of the array in question.

It is often argued that the relative contribution of high intensity spots into the normalization sum would significantly outweigh that of the low intensity ones and a handful of high intensity spots with outlying ratios might negatively affect the normalization. However, our analysis shows that this is not the case for a typical microarray experiment. The integral signal for a channel can be presented as

$$I^{R,G} = \int_0^\infty x f^{R,G}(x)\, dx,$$

where x is the signal intensity and $f(x)$ is the PDF for the signal intensities on the array. The contribution of a certain range of intensities is therefore proportional to $xf(x)$. Figure 6 shows a typical behavior of $xf(x)$ together with the underlying PDF. One sees that those spots that contribute to the normalization lie in a sufficiently broad range of magnitude, in this case, between 1,000 and 40,000 counts. Within this range $xf(x)$ has a relatively flat shape; e.g., the contribution of spots with intensity 2,000 is roughly equal to that of spots with count 20,000. Neither low nor very high intensity signals contribute significantly to the integral signal. This type of behavior was observed for the majority of our human and yeast samples.

The method of integral balance inherently possesses some degree of resistance to outliers since summing the intensities prior to calculation of the ratio effectively smoothes out the contribution of individual spots with potentially outlying ratios. However, some heuristics can be implemented to further improve the

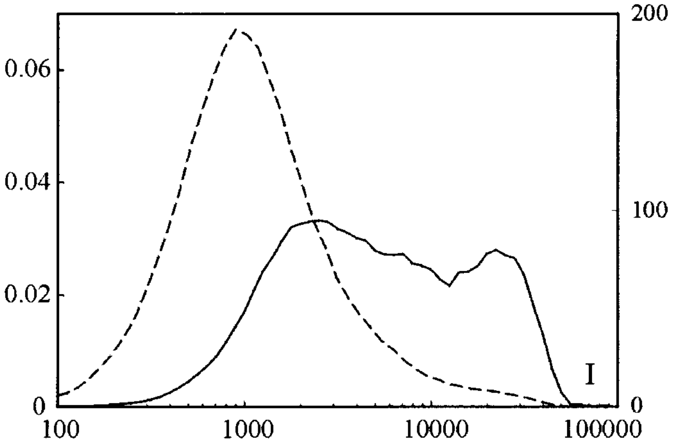

FIG. 6. The contribution of spots with different intensities to the normalization sum for the method of integral balance. The profile of $xf(x)$ is shown by the solid line (right scale); underlying PDFL $f(x)$ is shown by the dashed line (left scale).

robustness of the method. We tested the following iterative procedure, which can be seen as an extension of trimming (Barnett and Lewis, 1994). The initial normalization factor ν_0 was computed as a ratio of the total signal intensities of all spots within selected domain. Then all data points were normalized by ν_0 and those spots whose ratio falls into the range $[1/\eta, \eta]$ (with an empirically chosen value of trimming factor $\eta = 2$) were selected to calculate the integral intensities in the next step to give ν_1. After a few iterations, this procedure converges to some asymptotic factor ν, provided that the distribution of the ratios is unimodal. For experiments with a moderate number of differentially expressed genes, the methods of constant majority and integral balance show predictable convergence in estimates for normalization factors as illustrated by Fig. 7.

3.3. *Differential labeling*

The information extracted from a single microarray is not sufficient to correct the distortion of the ratios caused by the differential incorporation of the dyes ($\gamma_i^G \neq \gamma_i^R$). It is common to deal with this problem by performing two experiments with the same sample RNA and control RNA but labeled with alternate dyes. In the notations used throughout this paper, it was assumed that the sample was labeled with Cy3 and the control with Cy5 (direct order). Thus, for the inverse experiment in which the sample is labeled with Cy5 and the control with Cy3, model (1) should be modified by alternating the superscripts R and G. By performing background correction and normalization separately for both experiments, we arrive at two vectors of ratios $T_i^D = \Gamma_i T_i^{e(D)}$ and $T_i^I = \Gamma_i^{-1} T_i^{e(I)}$, where $\Gamma_i = \gamma_i^G / \gamma_i^R$. The superscripts D and I refer to the quantities found in the direct and the inverse experiments respectively. Ideally, the effective expression ratios $T_i^{e(D)}$ and $T_i^{e(I)}$ obtained in the direct and the inverse experiments should be equal. However, in practice, $T_i^{e(D)}$ and $T_i^{e(I)}$ are not identical as they incorporate different experiment-dependent noise terms $\varepsilon_i^{G,R}$ (see definition of effective ratios in Section 3).

Finally, the geometric average,

$$\overline{T_i^e} = \sqrt{T_i^D T_i^I} = \sqrt{T_i^{e(D)} T_i^{e(I)}},$$

gives an unfolded value which provides an unbiased estimate for the true expression ratio T_i^a. Figure 8 presents microarray data obtained for the stimulation of human cells with interferon. Both direct and inverse experiments were performed in duplicate. Four arrays (direct: numbers 1, 2; and inverse: 3, 4) were corrected for background and normalized. The log-transformed ratios (base 2) were plotted 1 versus 3

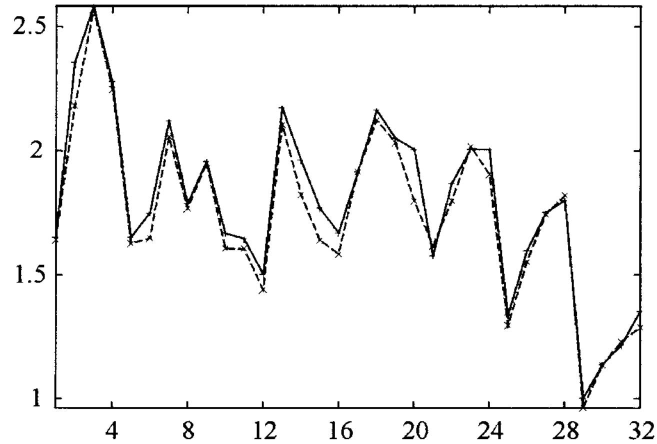

FIG. 7. Comparison of normalization factors calculated per grid (24 × 25 spots) for an array with 32 grids computed with the methods of constant majority (solid) and integral balance (dash). Grids, arranged as 4 × 8 array, are numbered in the order from left to right and top to bottom.

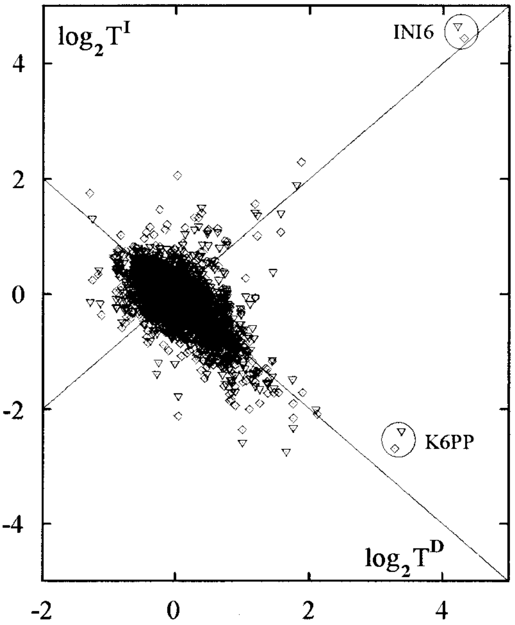

FIG. 8. Reproducible differential labeling. Log-transformed (base 2) ratios are plotted vs. ratios obtained in experiment with inverse labeling. Two replications are shown by diamonds and open triangles. Differentially labeled genes display clear anticorrelation and, therefore, align along the line $y = -x$.

(diamonds) and 2 versus 4 (triangles). This representation clearly reveals that we can distinguish differentially expressed from differentially labeled genes. Indeed, since

$$\log_2 T_i^D = \log_2 T_i^{e(D)} + \log_2 \Gamma_i,$$

$$\log_2 T_i^I = \log_2 T_i^{e(I)} - \log_2 \Gamma_i,$$

differentially expressed genes demonstrate a significant correlation between the direct and inverse experiments and, hence, a significant projection onto the main diagonal $y = x$. On the contrary, if cDNAs are differentially labeled, they exhibit anticorrelated ratios and align along line $y = -x$. A considerable number of genes show some degree of preference for incorporation of either dye (see Fig. 8). However, a few ratios fall significantly out of the range [0.5, 2.0], for example, the ratio demonstrated by phosphofructokinase K6PP. It is critical that these genes are eliminated from the final data set.

By comparing several direct-inverse pairs of experiments performed at different times, we found that the incorporation efficiencies, γ_i, depended sensitively on the conditions of the labeling reaction, but not on the source of RNA. Therefore, the genes that exhibit differential labeling may change from experiment to experiment. It may be advisable that both direct and inverse experiments are performed simultaneously, with the same preparation of RNA and fixed conditions of the labeling reaction.

4. DISCUSSION

The motivation for this paper was to provide a unified framework for the processing of DNA microarray data. On the basis of the analysis of hundreds of microarrays, we identified plausible causes for the systematically observed distortions and developed a model relating the abundance of mRNA transcripts to the measured signal intensities. The model strives to capture the major factors that affect labeling, scanning, and image quantification techniques. By applying its inverse transformation to the raw data, we were able to unfold it, i.e., recover a somewhat noisy image of the actual mRNA ratios. In some detail, we discussed our methods for the statistical estimation of the model parameters. Two features of these algorithms were emphasized as indispensable: statistical robustness, resistance to outliers in particular, and extensibility for accommodation of spatial dependence.

For practical application, it is important to estimate the success of the unfolding process. A number of experimental techniques are available for validation of the expression ratios found in a microarray experiment. These methods, however, can be applied only to a very limited number of important genes. A different approach is necessary to estimate validity of expression ratios for the entire set of genes spotted on a microarray. Our strategy to address this problem followed from the very idea of unfolding. The method is designed to remove systematic errors introduced by the microarray technology and provide close approximation for the true expression ratios. Such systematic errors, e.g., those due to background or arbitrary scaling factors, vary widely from one array to another. Therefore, the reproducibility of correctly unfolded replications of the same experiment (same sample and control hybridized to different arrays) should be distinctively higher than that of the corresponding raw data. The interchip variability coefficient ϑ, which is introduced in Section 5.4, can be applied for such a comparison.

We consider the unfolding to be successful if, upon removal of the systematic errors, the variability between replicate experiments does not exceed the variability inherent in a single array (e.g., as quantified by the intrachip variability coefficient σ_ω introduced in Section 5.3). Hence, the target value for ϑ is 1. To verify the performance of the method, we computed the variability coefficient for several pairs of microarrays (each pair being two replicates of the biologically identical experiment). First, ϑ_u was obtained for the unfolded expression ratios. For comparison, ϑ_g was calculated for the partially processed raw data. The preprocessing, in the form of the global normalization (regardless of possible spatial dependence) without background correction, was applied since the direct comparison of raw data by means of ϑ is not strictly correct. Indeed, just by varying the scaling factors A^G, A^R for otherwise identical data sets, we could obtain arbitrarily large values of ϑ. For control, ϑ_{ru} and ϑ_{rg} were also computed for randomized unfolded and randomized globally normalized raw data sets, respectively. The randomization was achieved by calculating the statistics ω for randomly permuted pairs of genes.

The results of these tests are summarized in Table 1. The performance of the unfolding appeared to be nearly optimal because all ϑ_u were close to 1 but slightly higher. In all cases, significant improvement was gained over global normalization without background correction. The last two columns of Table 1 compare relative improvement of the variability coefficient for actual and randomized data. One observes negligible improvement for the random data. This might be attributed to nonspecific factors. For example, global normalization, when applied to the raw data, does not remove spatial gradients, and therefore the ratio PDF for the randomized data is wider than that for the properly unfolded data.

For most users, the relatively high cost of microarray experiments prohibits the multiple replications that are required for statistical assurance. Often, with only a few replicates, the poorer experiments should be eliminated. Methods that can estimate the overall quality of an individual microarray are therefore quite important. Several quantities introduced in this study can be utilized to this end. For example, the measure of intrachip variability σ_ω can be used to single out experiments with unusually high levels of noise. The complementary approach compares the intensity of a spot to the intensity of the background noise. The necessary statistic here is a signal-to-noise ratio (SNR) defined as a difference between intensities of signal and local background normalized by the measure of background dispersion. The distribution of the SNR on a microarray provides a clear and easily comprehensible picture of the overall quality of an experiment.

Table 1. Test of Performance of the Unfolding for Four Pairs of Replicate Microarray Experiments

	ϑ_u	ϑ_g	ϑ_{ru}	ϑ_{rg}	$\dfrac{\vartheta_g - \vartheta_u}{\vartheta_g + \vartheta_u}$	$\dfrac{\vartheta_{rg} - \vartheta_{ru}}{\vartheta_{rg} + \vartheta_{ru}}$
1	1.18	2.15	6.07	6.21	0.291	0.011
2	1.27	2.53	7.50	7.75	0.331	0.016
3	1.22	2.43	5.45	5.61	0.332	0.014
4	1.15	2.12	5.32	5.52	0.296	0.018

The analysis presented in this paper also helps to elucidate areas of the microarray technology that contribute most to variability and uncertainty of the results. For example, the classical labeling procedure appears to be one of the major contributing factors. The coefficients of incorporation $\gamma_i^{R,G}$, depending on type of dye and varying from gene to gene and experiment to experiment, introduce significant variability. To overcome this problem, a number of labeling methods which ensure that every cDNA molecule carries the same number of fluorescent labels are being currently developed (e.g., the 3DNA$^{\text{TM}}$ technique, Genisphere). With the labeling problem solved, a universal normalization method could be devised with the help of artificially prepared controls carrying exactly equal numbers of dyes of both types. Thus, constant development of experimental techniques complemented by ongoing improvement of analysis methods ensures that microarray technology will live up to the high expectations of modern genomics.

5. APPENDIX

5.1. Transformation $y = \ln x$

Logarithmic transformation has been proved useful in dealing with quantities whose standard deviation is proportional to the mean (Snedecor and Cochran, 1973) and therefore can be applied to expression ratios. Suppose that random variable x is distributed according to the PDF $f(x)$. Then the log-transformed variable $y = \ln x$ is described by the PDF for the logarithm (PDFL) $g(y)$ given by

$$g(y) = e^y f(e^y).$$

The properties of PDF and PDFL may differ significantly. Consider PDFL $g(y)$ represented by a Gaussian normal function $N(0, \sigma)$; then the corresponding PDF is given by the log-normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\ln x)^2}{2\sigma^2}\right).$$

While $g(y)$ is symmetric with zero mean, mode, and median, $f(x)$ is a strongly asymmetric, skewed to the right function whose mean, mode, and median depend on σ. For example, solving the equation $f'(x) = 0$, we find the position of the PDF mode

$$\mu = e^{-\sigma^2}.$$

Consider now a finite-sized sample x_i, $i = 1, N$ log-transformed into set y_i. By calculating sample mean $m_y = \sum y_i/N$ and transforming it back, one finds the *geometric mean* of the original sample

$$m_g = \sqrt[N]{x_1 \cdot \ldots \cdot x_N}.$$

This quantity is often more useful (e.g., when dealing with ratios) than the simple arithmetic mean and is extensively utilized throughout this paper.

5.2. Preparation of background samples

To obtain the upper confidence limit $\overline{I}_B$ and background correction factors $B^R(\boldsymbol{r})$, $B^G(\boldsymbol{r})$, we needed to study the global statistical properties of the background as well as its spatial correlation. Towards this aim, we performed several experiments using total yeast RNA and complete yeast genome microarrays. The intensity of background on both channels was integrated on long strips partitioned into rectangular arrays of 20×200 square elements with side $l = 12$ pixels (corresponding to 120 μm in actual array dimension). This size of the element was selected to equalize its area with that of an average spot. The strips were positioned on images of 10 randomly chosen microarrays in regions free of spotted DNA to ensure that we were measuring a background signal uncontaminated by fractured, irregular, or smeared spots. To investigate spatial correlation of the background, we performed Fourier analysis along each column of 200 elements. The resulting power spectra were averaged over 20 such columns for every sample.

5.3. Intrachip variability

The microarrays used in our study have duplicate spots for every cDNA clone (gene) which are printed next to each other. Therefore, in a single microarray experiment, we obtain two ratios t_1 and t_2 for every gene. This feature can be exploited to evaluate the level of the ratio noise for a given microarray experiment. Consider the statistics

$$\omega = \sqrt{2}\frac{t_2 - t_1}{t_1 + t_2},$$

which describes the relative variability of the ratios. The factor $\sqrt{2}$ is introduced so that $|\omega|$ is equal to the coefficient of variation for a sample of size two. (This makes definition of ω easily extensible for cases with more than two replicate spots per gene.)

The physical proximity of the two spots on the microarray ensures that all distorting factors that systematically vary with $\boldsymbol{r}$ are essentially equal for both spots. In addition, ω does not change if both t_1 and t_2 are multiplied by the same factor. Therefore ω effectively measures purely random, spatially uncorrelated ratio noise emerging due to inherent variability of hybridization, fluorescence detection shot noise, or quantification errors. A probability density function of ω calculated for a typical microarray is shown by the thick solid line in Fig. 9. The Lorenzain shape of the PDF can be explained by the fact that the statistical properties of ω depend on the spot intensities. To demonstrate this, the entire population of spots was divided into three equal parts according to the intensity, and corresponding distributions of ω were computed. The resulting PDF's are shown in Fig. 9 by thin dotted lines. Note that their shapes are closer to the expected normal curves. Informative estimates for ratio noise can be obtained by averaging $|\omega|$ over groups of spots with similar intensities. In our experiments, we found the variability of ratios to be on average 5–7% at the higher end of intensity distribution and 20–25% at the lower end. The standard deviation, σ_ω, of the statistic ω can be employed to characterize intrachip variability of a given experiment with a single value.

5.4. Interchip variability

In practice, it is necessary to estimate the reproducibility of replicate microarray experiments. Consider the case of two replicates in which every gene is represented by ratio t_1 on the first array and t_2 on the second. After logarithmic transformation of ratios $x = \ln t_1$, $y = \ln t_2$, the results of both arrays can be represented on place (x, y) by a bivariate distribution, which under certain assumptions, is also normal (see Section 3.2). It is common to characterize the strength of linear association between two random variables with correlation coefficient

$$\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_x \sigma_y}},$$

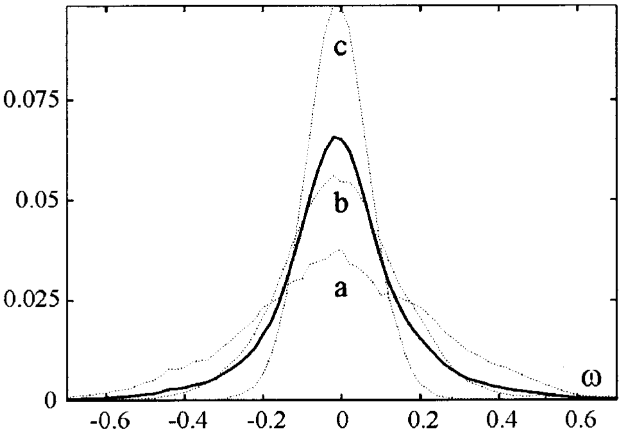

FIG. 9. Probability density function for ω statistic (solid line). Corresponding PDFs for the three intensity groups are shown by dotted lines: **a** shows the lower third of the intensities, **b** the middle, **c** the upper.

where σ_x and σ_y are respective standard deviations of x and y, and σ_{xy} is their covariance. However, ρ is not an adequate measure of reproducibility for replicate experiments, and its value can be misleading. To demonstrate this, we applied further transformation of the variables

$$u = \frac{x+y}{\sqrt{2}} = \sqrt{2}\ln\sqrt{t_1 \cdot t_2},$$

$$v = \frac{y-x}{\sqrt{2}} = \frac{1}{\sqrt{2}}\ln\frac{t_2}{t_1},$$

which results in the rotation of the axes by $\pi/4$ (see Fig. 10). Here, u is the logarithm of the geometric mean of t_1 and t_2, and contains no information on their difference. On the contrary, the value of v only shows the difference between t_1 and t_2. (In fact, u and v are the principal components for the considered bivariate ratio PDFL.) Expressing x, y through u, v after some algebra we find

$$\rho^2 = \frac{\sigma_u^2 - \sigma_v^2}{\sigma_u^2 + \sigma_v^2},$$

in which σ_u, σ_v are respective standard deviations of the PDFL in the rotated axes u, v. From the above discussion, it follows that σ_u effectively represents the variability of the expression ratios within the experiment, whereas σ_v is indicative of the variation between the replicates. Therefore, the ρ value for two replicates of an experiment with a large number of differentially expressed genes (large σ_u; case ***a*** in Fig. 10), is always higher than that for two replicates of an experiment with no changers but the same variability σ_v (case ***b***). The correlation coefficient is thus unsuitable for comparing reproducibility of different experiments. However, σ_v can be employed as a measure of interchip variability.

The approach, which we developed to measure intrachip variability, can be extended to estimate variation between replicates. Consider again the statistic ω but with t_1 and t_2 now being ratios on the first and second arrays, respectively. By the same argument, the standard deviation σ_ω is an integral measure of the discrepancy between the two replicates. Despite the obvious difference in definition, ω and v are closely related. Indeed, making use of the log-transformed variables x and y, we find

$$\omega = \sqrt{2}\frac{e^y - e^x}{e^y + e^x} = \sqrt{2}\tanh\left(\frac{y-x}{2}\right) = \sqrt{2}\tanh\left(\frac{v}{\sqrt{2}}\right).$$

In most cases, t_1 and t_2 are not significantly different and $|v| < 1$ ($|\ln t_2/t_1| < \sqrt{2}$). This allows us to develop a hyperbolic tangent in Taylor series to obtain an approximation $\omega \approx v$.

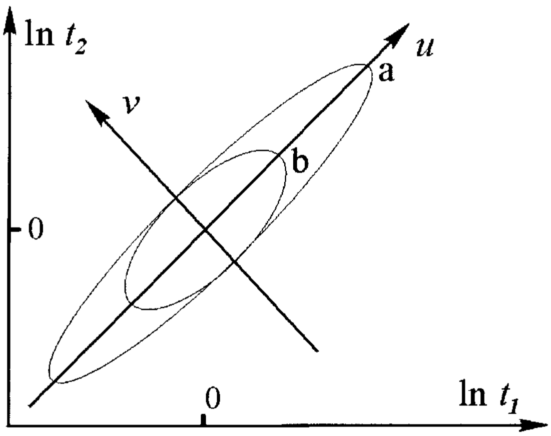

FIG. 10. Diagram illustrating transition to the principal components (u, v). Two hypothetical bivariate PDFs with equal measure of variability σ_v and different σ_u ($\sigma_u^{(a)} > \sigma_u^{(b)}$) are schematically drawn as ellipses.

Note that both σ_ω and σ_v measure interchip variability regardless of the level of noise present in individual arrays. In practice (see Section 4), one often wants to know how variability between two replicates relates to the variability inherent in each of them. To address this issue, a normalized interchip variability coefficient can be defined as

$$\vartheta = \frac{\sigma_{12}}{\sqrt{\sigma_{11} \cdot \sigma_{22}}},$$

where σ_{11}, σ_{22} are the intrachip variabilities of the two replicates.

5.5. *Preparation of microarrays*

The microarrays that we used in our experiments (Microarray Centre, Ontario Cancer Institute), were printed on standard modified-glass slides (Corning) by a 32-pin contact arrayer (SDDC II, Engineering Services Incorporated). The full genome yeast array comprised 6200 ORF's. The large human array consisted of two slides bearing approximately 19,000 human EST clones (Genetic Systems). A subset of 1,718 clones for which protein products are characterized in SwissProt was singled out for a smaller (1.7K) human array. Detailed information on the layout of microarrays can be found on the site of the Microarray Centre (*http://www.oci.utoronto.ca/services/microarray*). The protocols used for preparation of RNA, hybridization, and washing of microarrays can be downloaded from our website (*http://january.med.utoronto.ca*).

ACKNOWLEDGMENTS

This work was supported by the National Research Council, the National Science and Engineering Research Council (NSERC), and the Medical Research Council of Canada (MRC). A.G. was supported by the NSERC postdoctoral fellowship. A.M.E. is a Scientist of the MRC. The authors gratefully acknowledge the help and advice of Bryan MacNeil and the staff of the Microarray Centre at the Ontario Cancer Institute. A.G. acknowledges many helpful discussions of statistical methods with David Tritchler. The authors thank Yuval Kluger for comments on the manuscript.

REFERENCES

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T.D., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., and Staudt, L.M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. 1993. *Current Protocols in Molecular Biology*, Wiley, New York.

Barnett, V., and Lewis, T. 1994. *Outliers in Statistical Data*, Wiley, New York.

Brown, P.O., and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* suppl. 21, 33–37.

Chen, Y., Dougherty, E.R., and Bittner, M.L. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2, 364–374.

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

Cowan, G. 1998. *Statistical Data Analysis*, Oxford University Press, Oxford, U.K.

Debouck, C., and Goodfellow, P.N. 1999. DNA microarrays in drug discovery and development. *Nature Genet.* suppl. 21, 48–50.

Duggan, D.J., Bittner, M.L., Chen, Y., Meltzer, P., and Trent, J.M. 1999. Expression profiling using cDNA microarrays. *Nature Genet.* suppl. 21, 33–37.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfeld, C.D., Lander, E.S. 1999. Molecular classification of cancer: Class discover and class prediction by gene expression monitoring. *Science* 286, 531–537.

Inoue, S., and Spring, K.R. 1997. *Video Microscopy*, Plenum Press, New York.

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

Rousseeuw, P.J., and Leroy A.M. 1987. *Robust Regression and Outlier Detection*, Wiley, New York.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. 2000. Normalization strategies for cDNA microarrays. *Nucl. Acids Res.* 28, e47.

Snedecor, G., and Cochran, W. 1973. *Statistical Methods*, Iowa State University Press, Ames, IA.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* 9, 3273–3297.

Walpole, R.E., Myers, R.H., and Myers, S.L. 1998. *Probability and Statistics for Engineers and Scientists*, Prentice Hall, Upper Saddle River, NJ.

Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., and Somogyi, R. 1998. Large-scale temporal gene expression mapping of central nervous system development. *PNAS* 95, 334-339.

Address correspondence to:
Aled M. Edwards
C. H. Best Institute, Rm. 402
University of Toronto
112 College Street
Toronto, ON M5G 1L6, Canada

E-mail: aled.edwards@utoronto.ca