



Ratio statistics of gene expression levels and applications to microarray data analysis

Yidong Chen¹, Vishnu Kamat², Edward R. Dougherty^{2,*}, Michael L. Bittner¹, Paul S. Meltzer¹ and Jeffery M. Trent¹

¹Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Building 50, Room 5154, 50 South Drive, MSC 8000, Bethesda, MD 20892, USA and ²Department of Electrical Engineering, Texas A&M University, 3128 TAMU, College Station, TX 77843-3128, USA

Received on November 19, 2001; revised on March 4, 2002; accepted on March 19, 2002

ABSTRACT

Motivation: Expression-based analysis for large families of genes has recently become possible owing to the development of cDNA microarrays, which allow simultaneous measurement of transcript levels for thousands of genes. For each spot on a microarray, signals in two channels must be extracted from their backgrounds. This requires algorithms to extract signals arising from tagged mRNA hybridized to arrayed cDNA locations and algorithms to determine the significance of signal ratios.

Results: This paper focuses on estimation of signal ratios from the two channels, and the significance of those ratios. The key issue is the determination of whether a ratio is significantly high or low in order to conclude whether the gene is upregulated or downregulated. The paper builds on an earlier study that involved a hypothesis test based on a ratio statistic under the supposition that the measured fluorescent intensities subsequent to image processing can be assumed to reflect the signal intensities. Here, a refined hypothesis test is considered in which the measured intensities forming the ratio are assumed to be combinations of signal and background. The new method involves a signal-to-noise ratio, and for a high signal-to-noise ratio the new test reduces (with close approximation) to the original test. The effect of low signal-to-noise ratio on the ratio statistics constitutes the main theme of the paper. Finally, and in this vein, a quality metric is formulated for spots. This measure can be used to decide whether or not a spot ratio should be deleted, or to adjust various measurements to reflect confidence in the quality of the measurement.

Contact: e-dougherty@tamu.edu

INTRODUCTION

Expression-based analysis pertaining to large families of genes has become possible in recent years owing to the

development of cDNA microarrays, in which transcript levels can be simultaneously determined for thousands of genes (Schena *et al.*, 1995). Microarray data have been used to cluster genes based on expression profiles, characterize and classify disease based on the expression levels of gene sets, and to design discrete nonlinear predictors of gene expression levels based on both the expression levels of other genes and quantifiers for external stimuli. When using cDNA microarrays, the signal at each spot must be extracted from the background. This inevitably involves image-analysis algorithms capable of extracting signals arising from tagged mRNA hybridized to arrayed cDNA locations (Chen *et al.*, 1997; Schadt *et al.*, 2000; Kim *et al.*, 2001) and variability analysis and measurement quality control assessment (Chen *et al.*, 1998; Newton *et al.*, 2001; Brown *et al.*, 2001; Wang *et al.*, 2001).

One of the first image processing algorithms having basic components specially developed for microarrays (Chen *et al.*, 1997) applies image processing techniques to measure the signals and ratio statistics to determine whether a ratio is significantly high or low. This paper builds on and extends the algorithm by replacing the original hypothesis test for determining ratios that significantly deviate from unity. The original test involves a ratio statistic under the supposition that the measured fluorescent intensities subsequent to image processing can be assumed to reflect the signal intensities. Under this supposition, the hypothesis test has proven to work satisfactorily when signals are strong; however, it has not performed consistently satisfactorily for weak signals. This paper introduces and studies a refined hypothesis test in which the test statistic involves a ratio of measured intensities that are combinations of signal and background noise. The extension involves a signal-to-noise ratio and it is seen that for a high signal-to-noise ratio the new test reduces (with close approximation) to the original test.

The paper also proposes a quality metric for spots. This

*To whom correspondence should be addressed.

metric can be used to decide whether or not a spot ratio should be accepted, or to adjust various measurements to reflect confidence in the quality of the measurement. The overall quality measure is defined as the minimum of four individual quality metrics. The individual metrics relate to fluorescent intensity, target area, background flatness, and signal intensity consistency.

Since measurement via image processing is essential to the ratio statistics discussed in this paper, we provide a web page (http://arrayanalysis.nih.gov/resources/pub_download/bio1_supplement.htm) detailing the changes in our image processing methods that have been adopted since an earlier paper describing our system (Chen *et al.*, 1997). A block diagram of the image analysis system is shown in Figure 1 (which is on the web page, along with all other figures referred to in this paper). The description of image processing on the web page treats the following particulars: target segmentation, clone information assignment, background detection (Figure 2), target detection, intensity measurement, and ratio calculation.

RATIO STATISTICS

In typical biological samples, the number of genes that express at a similar level is approximately exponentially distributed (Bishop *et al.*, 1974). Most genes are weakly expressed, while a limited number are abundantly expressed. The detection of weakly expressed genes is limited by the fluorescent background, which is typically a combination of nonspecific bonding of the tagged mRNA samples to the glass slide and background fluorescence of the confocal microscope. In many practical microarray analyses, the ratio statistic has been chosen to quantitate the relative expression levels differentially expressed in two biological samples. This choice is based on certain assumptions: (1) the level of a transcript depends roughly on the concentration of the related factors which, in turn, govern the rate of production and degeneration of the transcript; (2) the random fluctuation for any particular transcript is normally distributed; and (3) as a fraction of abundance, the variation of any transcript is constant relative to most of the other transcripts in the genome, which means that the coefficient of variation can be taken as constant across the genome. In most practical microarray applications, the evidence of equally distributed genes, regardless of their expression levels, around the 45° diagonal line in a log–log scatter plot shows that the constant-coefficient-of-variation assumption is not overly violated. Nonetheless, even if the assumption intrinsically holds for expression levels, a detected expression level may not satisfy the assumption, particularly when the gene expresses weakly. Expression-level variation increases when the levels approach the background fluorescence, even when image processing techniques reliably detect

the cDNA target. Therefore, it is necessary to understand the properties of the ratio statistic when the expression level is near the background.

In the following subsections, we first briefly review our previous results for gene expression ratio statistics under the ideal situation. A new ratio model is then proposed in which the background fluorescence is included. We will demonstrate that the weaker expression level causes a larger coefficient of variation, thereby violating the previous assumption of constant coefficient of variation and causing a larger confidence interval for the null hypothesis. We will propose a simulation method to compensate the confidence interval for expression ratio analysis.

Ratio statistics assuming a constant coefficient of variation

Consider a microarray having n genes, with red and green fluorescent expression values labeled by R_1, R_2, \dots, R_n and G_1, G_2, \dots, G_n , respectively. Previously, we have proposed a ratio-based hypothesis test for determining whether R_k is over- or underexpressed relative to G_k , assuming a constant coefficient of variation across the microarray (Chen *et al.*, 1997). This assumption facilitates pooling statistics on gene expression ratios across the microarray. Letting μ_{R_k} and σ_{R_k} denote the mean and standard deviation of R_k (similarly for G_k), the assumption means that

$$\begin{aligned}\sigma_{R_k} &= c\mu_{R_k} \\ \sigma_{G_k} &= c\mu_{G_k}\end{aligned}\quad (1)$$

where c denotes the common coefficient of variation (cv). The desired hypothesis test is

$$\begin{aligned}H_0 &: \mu_{R_k} = \mu_{G_k} \\ H_1 &: \mu_{R_k} \neq \mu_{G_k}\end{aligned}\quad (2)$$

using the ratio test statistic $T_k = R_k/G_k$. Under the null hypothesis H_0 , Equation (1) implies that $\sigma_{R_k} = \sigma_{G_k}$. Assuming R_k and G_k to be normally and identically distributed, T_k has the density function

$$f_{T_k}(t; c) = \frac{(1+t)\sqrt{1+t^2}}{c(1+t^2)^2\sqrt{2\pi}} \exp\left[\frac{-(t-1)^2}{2c^2(1+t)}\right]\quad (3)$$

The subscript k does not appear on the right-hand side of Equation (3). Hence, the density function holds for all genes, and all ratios satisfying the null hypothesis can be pooled to estimate the parameter of Equation (3). The estimate is given by

$$\hat{c} = \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{(t_i - 1)^2}{(t_i^2 + 1)}}\quad (4)$$

where t_1, t_2, \dots, t_n are ratio samples taken from a family of housekeeping genes on a microarray.

For a microarray derived from two identical mRNA samples co-hybridized on one slide (self-self experiment), the parameter c (cv of the fluorescent intensity) provides the *variation of assay*. However, for any experiment, to guarantee the null hypothesis condition in Equation (2) is not always possible. One alternative is to duplicate some or all clones where the same expression ratio is expected. For the ratio of expression ratios, $T = t/t'$, and its natural log-transform,

$$\log T_k = \log t_k - \log t'_k = (\log R_k - \log R'_k) - (\log G_k - \log G'_k) \quad (5)$$

it can be shown that (see Appendix on the web page)

$$c \approx \sqrt{\frac{1}{n} \sum_{i=1}^n (\Delta \log R_k)^2} = \sigma_{\log R_k} \quad (6)$$

where $\Delta \log R = (\log R - \log \mu_R)$. Therefore,

$$\sigma_{\log T}^2 \approx (\sigma_{\log R}^2 + \sigma_{\log R'}^2) + (\sigma_{\log G}^2 + \sigma_{\log G'}^2) = 4c^2 \quad (7)$$

when the measurement of the log-transformed expression level is approximately normally distributed (see web page). For any given experiment with some duplicated clones, $\sigma_{\log T}^2$ is easily calculated, and along with it the coefficient of variation of the assay.

The constant coefficient of variation condition imposed by Equation (1) implies a constant standard deviation of the log-transformed ratio (approximation) given by Equation (6), where the log-normal distribution is assumed. This condition is behind many higher-level analyses such as clustering, classification, and other differential expression significance assessments (Tusher *et al.*, 2001). On the other hand, some studies indicate that the coefficient of variation varies with the expression intensities (Hughes *et al.*, 2000; Nadon *et al.*, 2001). We will extend our study in later sections.

In practical application, a constant amplification gain m may apply to one signal channel, in which case the null hypothesis in Equation (2) may become $\mu_{R_k} = m\mu_{G_k}$. Under this uncalibrated signal detection setting, the ratio density can be modified by

$$f_T(t; c, m) = \frac{1}{m} f_T(t/m; c, 1) \quad (8)$$

where $f_T(c, 1)$ is given by Equation (3). In (Chen *et al.*, 1997), an estimation procedure for the parameter m is proposed. It has proven efficient during four years of application. This procedure is called *normalization*. To satisfy the null hypothesis given by Equation (2), a set

of pre-selected housekeeping genes (about 100 different clones) or the entire gene set available in each array may be used for the estimation procedure. A set of commonly used housekeeping genes has been preferred in the procedure since they are believed to carry minimum changes in many biological systems. Since no set of genes is unchanged in all conditions, the normalization procedure automatically eliminates housekeeping genes if their expression ratios vary by more than 2-fold. The same requirement is used when the entire gene set is used in the normalization procedure.

To decide whether expression levels of a gene from two biological samples are significantly different, we would like to find a confidence interval such that within the confidence interval, the null hypothesis given in Equation (2) cannot be rejected: the expression ratio, $T_k = R_k/G_k$, of the gene under consideration is not significantly deviated from 1.0 if the ratio is within the confidence interval. The confidence interval can be evaluated by integrating the ratio density function given by Equation (3). Examples of confidence intervals are given in Table 2 of (Chen *et al.*, 1997). Since the confidence interval is determined by the parameter c , one can either use the parameter derived from pre-selected housekeeping genes (Equation (4)), or a set of duplicated genes (Equation (7)) if they are available in the array. The former confidence intervals contain some levels of variation from the fluctuation of the biological system that also affect the housekeeping genes, while the latter contains no variation of the biological fluctuation, but contains possible spot-to-spot variation. Spot-to-spot variation is not avoidable if one wishes to repeat the experiment. The confidence interval derived from the duplicated genes is termed as the *confidence interval of the assay*.

Ratio statistics for low signal-to-noise ratio

The condition of constant cv is made under the assumption that R_k and G_k are the expression levels detected from two fluorescent channels. The assumption is based on image processing that suppresses the background noise relative to the true signal. It proves to be quite accurate for strong signals, but problems arise when the signal is weak in comparison to the background. Even with image processing, the actual expression intensity measurement is of the form

$$R_k = (SR_k + BR_k) - \mu_{BR_k} \quad (9)$$

where SR_k is the expression intensity measurement of gene k , BR_k is the fluorescent background level, and μ_{BR_k} is the mean background level. The measurable quantities are (1) signal with background, $SR_k + BR_k$, and (2) the surrounding background. The null hypothesis of interest

is $\mu_{SR_k} = \mu_{SG_k}$. Taking the expectation in Equation (9) yields

$$\mu_{R_k} = E[R_k] = E[(SR_k + BR_k) - \mu_{BR_k}] = \mu_{SR_k} \quad (10)$$

Since $\mu_{G_k} = \mu_{SG_k}$, the hypothesis test of Equation (2) is still the one with which we are concerned, and we still apply the test statistic T_k .

There is, however, a major difference. The assumption of a constant cv applies to SR_k and SG_k , not to R_k and G_k , and the density of Equation (3) is not applicable. The matter can be quantified by defining an appropriate signal-to-noise ratio. Assuming that SR_k and BR_k are independent,

$$\sigma_{R_k}^2 = \sigma_{SR_k}^2 + \sigma_{BR_k}^2 = (c\mu_{SR_k})^2 + \sigma_{BR_k}^2, \quad (11)$$

where c is the (assumed constant) cv of SR_k . Re-arranging Equation (9) yields $SR_k = R_k - (BR_k - \mu_{BR_k})$, so that the strength of the signal is determined by the extent to which R_k exceeds $BR_k - \mu_{BR_k}$. Taking into account the variation in the background (which tends to obscure the signal), we define the *signal-to-noise ratio* (SNR) for the red channel for gene k to be

$$SNR_{R_k} = \frac{E[SR_k]}{E[BR_k - \mu_{BR_k}] + \sigma_{BR_k}} = \frac{\mu_{SR_k}}{\sigma_{BR_k}} \quad (12)$$

Let c_{R_k} be the cv of the observed expression level of gene k from the red channel. Then

$$\begin{aligned} c_{R_k}^2 &= \left(\frac{\sigma_{R_k}}{\mu_{R_k}} \right)^2 = \frac{(c\mu_{SR_k})^2 + \sigma_{BR_k}^2}{\mu_{SR_k}^2} \\ &= c^2 + \frac{\sigma_{BR_k}^2}{\mu_{SR_k}^2} \\ &= c^2 + \left(\frac{1}{SNR_{R_k}} \right)^2 \end{aligned} \quad (13)$$

If $SNR \gg 1$ for gene k , meaning the expression signal is strong, then the measured cv is close to a constant, namely $c_{R_k} \approx c$, but if the signal is weak, the constant cv condition is violated and Equation (3) no longer holds. Figure 3 shows how in most practical applications, the weaker expression signals (at the lower left corner of the scatter plot) produce a larger spread of gene placement.

The observation of larger variation for weaker signals has been previously reported (Hughes *et al.*, 2000 (Supplemental); Nadon *et al.*, 2001). In the error model studied in (Hughes *et al.*, 2000), the variation of the expression signal consists of two parts: a constant part and a function of the average signal. Equation (13) differs from that model in assuming that it is the fluorescent background variation (instead of a function of the signal) and its interaction with the signal that causes the large deviation when the signal is weak.

Confidence interval for the test statistic

To take into account the lack of a constant cv for the observed levels, we need to examine the hypothesis test of Equation (2) relative to critical regions for the test statistics, where now

$$T_k = \frac{R_k}{G_k} = \frac{(SR_k + BR_k) - \mu_{BR_k}}{(SG_k + BG_k) - \mu_{BG_k}} \quad (14)$$

We maintain the following assumptions: SR_k , SG_k , BR_k , and BG_k are normally distributed and independent, there is a constant c for SR_k and SG_k (for all k), and there is a collection of housekeeping genes that can be used as an internal control set.

Due to the complexity of Equation (14), we use the Monte Carlo method and simulate various conditions to obtain properties of the ratio statistic. Dropping the subscript k to ease notation, the simulation is set up according to

$$T \leftarrow \frac{N(p, \sigma_p) + N(\mu_{BR}, \sigma_{BR}) - \mu_{BR}}{N(p, \sigma_p) + N(\mu_{BG}, \sigma_{BG}) - \mu_{BG}} \quad (15)$$

where the notation ' \leftarrow ' is used to express the distributions in the simulation and the distribution of T . We assume a given expression intensity has a normal distribution $N(\mu, \sigma)$ with mean expression intensity of $p = \mu_{SR_k} = \mu_{SG_k}$ ($= \mu_{R_k} = \mu_{G_k}$) under the null hypothesis. This normality assumption says that when multiple measurements are taken for one expression level, they follow a normal distribution. The mean expression level, p , may possess some other distribution according to biological properties, such as an exponential distribution or log-normal distribution. Under the assumption of constant cv for the signal (without the background), $\sigma_p = cp$. There is no assumption that either the means or the variances of the background levels from the two channels are the same. We introduce the following parameters: a variance parameter $\sigma_B = \max\{\sigma_{BR}, \sigma_{BG}\}$, a signal-to-noise-ratio $s = p/\sigma_B$, and the background standard-deviation ratio $\kappa = \sigma_{BR}/\sigma_{BG}$. Then

$$T \leftarrow \frac{N(s\sigma_B, cs\sigma_B) + N(0, \kappa\sigma_{BG})}{N(s\sigma_B, cs\sigma_B) + N(0, \sigma_{BG})} \quad (16)$$

Figure 4(a) shows the simulation for $c = 0.2$ and $\sigma_{BR} = \sigma_{BG} = 100$ (or $\kappa = 1$). The characteristics of the adaptive confidence interval match practical observation. First, the interval diverges when the SNR is small. Thus, if the expression signal is close to the background variation, it is harder to reject the null hypothesis. Second, as the SNR increases, the interval converges to the result given by Equation (3), meaning that the background-noise influence is becoming negligible. Third, if $SNR \rightarrow 0$, then the interval is limited by the ratio distribution

of background levels from the red and green channels, or $T = N(0, \kappa\sigma_{BG})/N(0, \sigma_{BG})$, which has a Cauchy distribution, $f_T(t) = \kappa\pi^{-1}(1 + \kappa^2 t^2)^{-1}$. For $\kappa = 1$, the 99% confidence interval of the Cauchy distribution is $[-63.7, 63.7]$, and it is the limit of Equation (16) when the signal is zero. In practical application, the estimate of the fluorescent intensity (both R_k and G_k) is always positive. Thus, the ratio will never be negative. To reflect this limitation, a negative ratio bound is always set to 0. Finally, when the two background variations are not the same ($\kappa \neq 1$), the upper and lower bounds are not symmetric, as illustrated in Figure 4(b). Figure 5 shows a surface of confidence intervals for $\kappa = 0.1$ to 10.

If the background is relatively flat, we can use the average of the local background intensity and its empirical standard deviation as estimates of μ_{BR} and σ_{BR} , respectively, and similarly for μ_{BG} and σ_{BG} . The coefficient of variation, c , of the expression intensity is derived from ratio analysis given by Equation (4) (or the improved version in Equation (13)) using a set of housekeeping genes. The simulation of Equation (15) provides an adaptive confidence interval for every gene, dependent on p , which is estimated as the average of the red and green expression intensities R_k and G_k , which share common mean p_k under the null hypothesis (p being dependent on gene k). A sequence of confidence intervals derived from the null hypothesis estimator of p_k is shown in Figure 6.

Based on the image analysis, the background may depend on location. If local background is extracted instead of global, then the simulation may have to be performed at each gene location. Simulation is necessary only when the SNR is less than 6; otherwise, the previous method dependent on analytic formulation of the distribution of T is sufficiently accurate to be used.

Correction of background estimation

Owing to interaction between the fluorescent signal and background, local-background estimation is often biased. Generally, the bias of the estimation increases when the background level is higher. To demonstrate the effect of the bias, we again use the model of Equation (15). The simulation is performed in the following manner:

- (i) generate 10 000 data points from an exponential distribution with mean 2000 to simulate 10 000 gene expression levels,
- (ii) assuming a constant coefficient of variation $c = 0.2$, simulate actual fluorescent intensity measurements given the null hypothesis by taking each of the 10 000 points as a mean for a Gaussian model. For each point, two simulated intensities are generated from the model to represent measurements from different channels,
- (iii) simulate background level by a normal distribution.

If no bias is assumed, then add a random quantity generated from $N(0, 100)$ to all fluorescent intensities in both channels. If some bias is assumed, then add a random quantity generated from $N(b, 100)$ to one channel or both.

When no background bias is added, the log-scaled scatter plot of simulated intensities shown in Figure 7a exhibits expected characteristics. Due to the constant c , the spread of expression data around the 45° diagonal is consistent when the SNR is high. The spread of expression data diverges when the SNR is low (lower-left corner).

If we add a constant bias to the background estimation, as in Figure 7b, then the centroid of the expression data turns to produce the commonly observed ‘dog-leg’ effect (which can also result from other factors, such as the possible nonlinearity of amplification between the Cy3 and Cy5 channels). Here we introduce a method that may partly fix the dog-leg problem, although it is up to individual researchers to determine the exact causes of the dog-leg effect.

A large-scale simulation reveals that there is no difference if both channel background levels are either underestimated or overestimated. Moreover, there is no clear way to identify which channel background is under or overestimated. The best we can do is to find the difference of the biases from two background-level estimations. To compensate for this difference, we can add the bias difference to one channel, depending on the sign of the bias difference. The risk of this procedure is possible ratio distortion, since a positive constant is added to the background-subtracted intensity measurement. We recommend data visualization before using the procedure.

To estimate the bias difference, we find the relationship between the red and green intensities under the null hypothesis by assuming a linear relation, $G = aR + b$. After the calibration given by Equation (8), the slope a is usually close to 1, and measurement bias is represented by b . Thus, only one parameter needs to be estimated from a set of intensity measurements, $\{R_k, G_k\}$, $k = 1, 2, \dots, N$. Typically, linear regression by least-squares fitting is used; however, it is not appropriate here because the variances of the intensity measurements R_k and G_k increase when the intensities increase. Therefore, we employ a chi-square fitting method (Bevington, 1969) that minimizes

$$\chi^2 = \sum_{k=1}^N \frac{(G_k - (aR_k + b))^2}{\sigma_{R_k}^2 + \sigma_{G_k}^2} \quad (17)$$

which yields a solution for the bias difference (see Appendix on web page),

$$b = \frac{\sum_{k=1}^N (c^2(R_k + G_k) + 2\hat{\sigma}_{BR}^2 + 2\hat{\sigma}_{BG}^2)^{-1}(G_k - R_k)}{\sum_{k=1}^N (c^2(R_k + G_k) + 2\hat{\sigma}_{BR}^2 + 2\hat{\sigma}_{BG}^2)^{-1}} \quad (18)$$

The result from chi-square fitting is more complicated if we do not assume a slope parameter $a \approx 1$, but $a \approx 1$ is appropriate for practical application and can be validated before applying the χ^2 -square test. When $b > 0$, b is added to all intensity measurements from the green channel; otherwise, it is added to intensities from the red channel. In practical application, the entire estimation process may have to be iterated since some large outliers should be removed. The appendix gives an example to demonstrate the precision of this estimation procedure.

QUALITY METRIC FOR RATIO STATISTICS

It is advantageous to attach a quality metric to each ratio prior to subsequent analysis. A metric must be designed at an early stage because information is lost when ratios are extracted from the image and forwarded to higher level processing. Moreover, in practical application, filters may be used, or there may be human intervention during image analysis. We propose a quality metric for cDNA expression ratio measurement that combines local and global statistics to describe the measurement quality at each ratio measurement. The single quality metric enables unified and universally applicable data filtering, and it can be used directly in higher-level data analysis.

For a given cDNA target, the following factors affect ratio measurement quality: (1) Weak fluorescent intensities from both channels result in less stable ratios. This quality problem has been addressed in the last section through the confidence interval if over or underexpression is the only concern; however, when comparing one experiment to another via a common reference sample, low intensities provide less reliable ratios. (2) A smaller than normal detected target area indicates possible poor quality in clone preparation, printing, or hybridization. (3) A very high local background level may suggest a problematic region in which any intensity measurement may be inaccurate. (4) A high standard deviation of target intensity is usually caused by the contamination of strong fluorescent particles within the target region. Our image processing package extracts all of this information, and for each factor, a quality metric w is defined, w taking a value from 1 (highest measurement quality) to 0 (lowest measurement quality).

Fluorescent intensity measurement quality

For gene k , a red-channel SNR exceeding 6 means that

$$\mu_{R_k} \geq E[BR_k - \mu_{BR_k}] + 6\sigma_{BR_k} \quad (19)$$

The signal is very strong relative to background variation, and we judge the intensity quality to be 1. A decline in quality can be measured relative to the quotient SNR_{R_k} . Similar statements apply to the green channel. For a conservative quality metric, we take the minimum of the SNRs to define the quality metric. Under the null

hypothesis, the signal means are equal, so that

$$\min\{SNR_R, SNR_G\} = \frac{\mu_R}{\max\{\sigma_{BR}, \sigma_{BG}\}} = \frac{\mu_R}{\sigma_B} \quad (20)$$

(k dropped to ease notation). For the intensity-measurement quality relative to the background variation, we replace μ_R and σ_B by their null-hypothesis estimators, $(R+G)/2$ and $\hat{\sigma}_B$, to obtain

$$w_I = \begin{cases} 0, & \frac{R+G}{2\hat{\sigma}_B} \leq 3 \\ \frac{R+G}{6\hat{\sigma}_B}, & 3 < \frac{R+G}{2\hat{\sigma}_B} \leq 6 \\ 1, & \text{otherwise} \end{cases} \quad (21)$$

Target area measurement quality

In a typical printing process, each printing tip produces a relatively consistent spot area. We use the number of pixels to describe the target area quality. Since each print-tip produces a unique spot shape and spot area (total number of pixels), we use the target mask derived in the image analysis (see web page) to judge each target. Let A_M be the area of mask of the cDNA target for a particular print-tip, and let A_{T_k} be the area of the two largest connected components of the target k . The proportional area of each target is $a_k = A_{T_k}/A_M$. We define the area measurement quality by

$$w_a = \begin{cases} 0, & a < s_{\min} = \max\{10/A_M, 0.05\} \\ \frac{a - s_{\min}}{s_{\min} - s_b}, & s_{\min} < a \leq s_b = 0.20 \\ 1, & \text{otherwise} \end{cases} \quad (22)$$

(subscript k dropped). The first branch in Equation (22) assigns metric 0 to any target with a size less than $0.05 \times A_M$, or less than 10 pixels. For sizes exceeding 20% of the mask size A_M , the quality metric is 1; otherwise, the quality factor varies from 0 to 1, depending on where the target size is in the interval $[s_{\min}, s_b]$. We choose the largest two connected components for target-area quality measurement (not to be confused with the union area used for target intensity measurement) since the smaller fragments of the target are usually due to heavy noise interaction.

Background flatness quality

Background abnormality can cause signal detection problems or incorrect measurements of the local background level. One way to detect this abnormality is to compare the local background intensity BR_k to a global mean intensity μ_{BR} using the corresponding global standard deviation σ_{BR} . The global background level is taken as an average of local background levels within a relatively large area, say,

the sub-array produced by one printing tip. We assume the abnormality is brighter. It is possible to have a local background area darker than the global background owing to scratches or some other mishandling steps; however, image processing usually deals with this problem correctly. If the local background BR_k is less than $\mu_{BR} + 4\sigma_{BR}$, then it is within the flatness requirement of the background; if not, then we linearly rate the quality from 1 to 0 until it reaches $\mu_{BR} + 6\sigma_{BR}$. The green channel background flatness is defined similarly. We obtain the background flatness definition $w_b = \min\{w_{BR}, w_{BG}\}$, where

$$w_{BR} = \begin{cases} 1, & BR_k < \mu_{BR} + 4\sigma_{BR} \\ \frac{(\mu_{BR} + 6\sigma_{BR}) - BR_k}{3\sigma_{BR}}, & \mu_{BR} + 4\sigma_{BR} \leq BR_k < \mu_{BR} + 6\sigma_{BR} \\ 0, & BR_k \geq \mu_{BR} + 6\sigma_{BR} \end{cases} \quad (23)$$

and w_{BG} is defined similarly.

Signal intensity consistency quality

In cases where contamination crosses the cDNA target or strong speckle spots sit atop the target area, the reported signal intensity may not truly reflect the actual signal. Another problem arises when one channel is very strong, the other is weak, and targets tend to be larger in the stronger channel. This causes the weak target region to contain too many background pixels, thereby increasing signal variation. In both cases, the signal standard deviation will be unusually high. To establish a normal range of signal intensity relative to its standard deviation, or coefficient of variation, we have simulated three cases. Figure 8 shows: (a) a bell-shaped intensity profile from a Gaussian distribution; (b) a stretched bell-shaped profile with a flat-top; and (c) a bimodal profile corresponding to a donut-shaped target. Simulated cv s for these three cases are 0.48, 0.45, and 0.31, respectively. The cv can be significantly perturbed by intensity inconsistencies. If more background area is included, corresponding to Figure 8d, then the cv is 0.81. If some strong pixels are included, corresponding to Figure 8e, the cv becomes 0.98. Adding noise to the target shape will not dramatically change the cv , as shown in Figure 8f, in which strong noise has been added to Figure 8a. In this case, the cv is 0.59, an increase of only 0.11 from Figure 8a. Based on these considerations, we define a signal-intensity-consistency quality factor. Letting $cv_{min,k}$ denote the minimum between the intensity coefficients of variation for the red and green channels,

$$w_s = \begin{cases} 0 & 1.1 < cv_{min,k} \\ \frac{cv_{min,k} - 0.9}{0.2} & 0.9 < cv_{min,k} \leq 1.1 \\ 1 & cv_{min,k} \leq 0.9 \end{cases} \quad (24)$$

Total measurement quality

Overall, we would like to report only one quality factor to the user, so that a reference quality can be recommended for any experiment. Motivated by the desire that spots rated to be of high quality are good with respect to all the criteria, we define the total quality metric as the minimum of the four individual metrics. One may utilize more quality measurements similar to the four we have discussed, so long as they follow the same scale: from 0 (lowest) to 1 (highest).

Application and assessment of quality metric

The quality metric can be used to redefine the ratio parameter estimator given in Equation (4) by

$$c = \sqrt{\frac{\sum_{i=1}^n w_i \frac{(t_i - 1)^2}{(t_i^2 + 1)}}{\sum_{i=1}^n w_i}} \quad (25)$$

Equation (26) provides a more robust estimator for c when strong noise is present. Since there is no need to apply data filtering steps before evaluating Equation (6) c is free of instability due to such filtering. The same scheme can be applied to various calculations, such as the estimation for m (Equation (8)), estimation for background bias b (Equation (18)) and many others.

The quality weight is useful for downstream data analysis tasks, such as the similarity measure for gene expression profile analysis. A typical similarity measure is the correlation coefficient, which can be easily modified by introducing the weight into the calculation to yield

$$\rho_{xy} = \frac{\sum_{i=1}^n w_i (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n w_i (x_i - \mu_x)^2} \sqrt{\sum_{i=1}^n w_i (y_i - \mu_y)^2}} \quad (26)$$

where x_i and y_i are the log-transformed ratios for genes x and y in n experiments, and $w_i = w_{x_i} w_{y_i}$. For more stringent conditions, one may binarize w_i to be 0 or 1. Then Equation (26) reduces to the Pearson correlation coefficient with only genes having measurement quality 1 being included.

A reliable quality metric makes it possible to use only the best measurements achieved or to discount the weight of inferior measurements in downstream analysis so that the more certain measurements dominate the outcome. A quality metric should be able to identify the array measurements that are precise enough to likely reproduce with a narrow variance. The proposed metric has been tested by examining its ability to find measurements with low variance from a series of eight replications of the same measurement. Quality assessments should fractionate the

measurements into classes where increasing quality of the individual measurements correlate with increased reproducibility of the observed ratio across the replicates.

To produce the replicate series of hybridizations, RNAs from a melanoma (UACC903) and a myeloid (ML1) cell line were labeled and profiled on eight microarrays, each having 6548 genes. As these cell types are quite different, their comparison provides a very wide range of observed abundance ratios with which to test the quality-determining algorithm. The RNAs were labeled in two batches, one sufficient for 5 hybridizations, where Cy5 was used for UACC903 and Cy3 for ML1, and three in which the dye assignments were swapped. All methods for array fabrication, RNA labeling and hybridization were as previously described (Jiang *et al.*, 2001). After hybridization and scanning, ratio and quality data were extracted for all measurements. To evaluate reproducibility, the ratio values were converted to \log_{10} ratio values and the standard deviation of the \log_{10} ratio value for each gene was calculated. Since low quality measurements are expected to severely degrade reproducibility, the overall quality of the series of measurements was judged conservatively. Overall quality for a given gene was represented by the median quality value for the series of replicates.

A histogram of the distribution of standard deviations of the \log_{10} ratio values for all genes is shown in Panel A of Figure 9. The observed distribution of variances reflects the wide distribution of signal strengths typically encountered in array experiments. The abundance of message for the genes being detected varies from near total absence to many copies per cell. This produces signals ranging from indistinguishable from background noise, to hundreds of times this level, thereby generating a broad spectrum of measurement variance. Panels B through D of Figure 9 show the progressive decrease in both median variance and spread of variance associated with fractions of the total set having higher and higher median qualities. If one uses the median variance of each fraction as an estimate of σ for each quality fraction, then the bounds of a 95% confidence interval for the ratio determinations in each fraction can be estimated. The genes with median qualities in the lowest third of possible values have an estimated 95% confidence interval of 0.29–3.44, while those having median qualities in the top third of possible values have the much tighter 95% confidence interval of 0.7–1.43. It is clearly possible to use the proposed metric to identify the most reproducible measurements being produced.

CONCLUSION

The main focus of this paper has been on the estimation and significance determination of signal ratios arising from the two channels of a cDNA microarray. For strong

signals, the previous assumption of a constant coefficient of variation for the distribution of the red and green channel intensities has been sharpened so that it is only assumed to hold for the target signal alone, not the total intensity measurement over the target. For a large SNR, there is little difference between the two suppositions; however, the change is important when there is a low SNR. The cost of the more refined model is loss of an analytic expression for the ratio density and the necessity of using Monte Carlo methods. The improvement in ratio estimation for weak signals is well worth the cost. Finally, it is clear that poor image quality adversely affects ratio measurements. A quality measure has been introduced that facilitates either the deletion of poor-quality spots or the weighting of post-processing statistics according to spot quality.

REFERENCES

- Bevington, P.R. (1969) *Data reduction and error analysis for the physical sciences*. McGraw-Hill, New York.
- Bishop, J.O., Morton, J.G., Rosbash, M. and Richardson, M. (1974) Three abundance classes in HeLa cell messenger RNA. *Nature*, **250**, 199–240.
- Brown, C.S., Goodwin, P.C. and Sorger, P.K. (2001) Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 8944–8949.
- Chen, Y., Dougherty, E.R. and Bittner, M. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics*, **2**, 364–374.
- Chen, J.J., Wu, R., Yang, P.C., Huang, J.Y., Sher, Y.P., Han, M.H., Kao, W.C., Lee, P.J., Chiu, T.F., Chang, F. *et al.* (1998) Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics*, **51**, 313–324.
- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
- Jiang, Y., Lueders, J., Glatfelter, A., Gooden, C. and Bittner, M. (2001) *Profiling Human Gene Expression with cDNA Microarrays*, Current Protocols in Human Genetics, Wiley, New York.
- Kim, J.H., Kim, H.Y. and Lee, Y.S. (2001) A novel method using edge detection for signal extraction from cDNA microarray image analysis. *Exp. Mol. Med.*, **33**, 83–88.
- Lee, M.L.T., Kuo, F.C., Whitmore, G.A. and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.
- Nadon, R., Shi, P., Skandalis, A., Woody, E., Hubschle, H., Susko, E., Rghei, N. and Ramm, P. (2001) Statistical inference methods for gene expression arrays. In Bittner, M.L., Chen, Y., Dorsel, A.N. and Dougherty, E.R. (eds), *Microarrays: Optical Technologies and Informatics, Proceedings of SPIE 4266*. San Jose, CA, pp. 46–55.

- Newton, M.A., Kendziorshi, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.
- Otsu, N. (1979) A threshold selection method from gray-level histogram. *IEEE Trans. Syst. Man. Cyber.*, **SMC-9**, 62–66.
- Schadt, E.E., Li, C., Su, C. and Wong, W.H. (2000) Analyzing high-density oligonucleotide gene expression array data. *J. Cell Biochem.*, **80**, 192–202.
- 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.
- Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
- Wang, X., Ghosh, S. and Guo, S.W. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.*, **29**, E75.