Microarray Gene Expression

James J. Chen
U.S. Food and Drug Administration, Jefferson, Arkansas, U.S.A.

Chun-Houh Chen
Academia Sinica, Taipei, Taiwan, Republic of China

INTRODUCTION

Deoxyribonucleic acid (DNA) microarray technology provides tools for studying the expression levels of a large number of distinct genes simultaneously.[1] The technology is based on the hybridization of known DNA segment on the array with an unknown labeled DNA or ribonucleic acid (RNA) sample based on base-paring rules. Gene expression levels are quantified from the image of hybridized microarray excited by a laser scanner. Signal intensities reflect the amount of transcript present for the gene in the mRNA sample. A typical microarray experiment data set includes expression levels of thousands of genes in a number of experimental samples (conditions). These samples may correspond to different toxins or serial time points taken during a biological process. The expression data can be summarized by a matrix with rows representing genes and columns representing samples. This matrix is referred to as gene expression matrix. Gene expression matrix can be analyzed by comparing expression profiles of an individual gene (row) under various conditions or comparing expression profiles of sample (column) to understand their similarities and differences and to find the relationships among genes and samples condition.

Microarray experiment is a multistep process, where each step is potentially a source of variation. Usually, raw data generated from image analysis can not be directly compared. Data need to be normalized in order to minimize systematic biases so that biological differences can be distinguished. Statistical analyses generally are of two types. The first type of analysis involves identifications of the genes that are differentially expressed among different sample groups. The second type seeks to determine the relationships between genes or gene clusters to identify biological functions, or to predict specific biological outcomes (or diseases) from the analysis of expression patterns. Many statistical methods are used in the microarray data processing and analysis. Microarray technology, experimental design, normalization, significance testing, and classification and prediction are discussed in the following sections.

DNA MICROARRAY EXPERIMENT

Background

A gene consists of a segment of DNA. A DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. The expression of genetic information stored in the DNA molecule occurs in two stages: 1) transcription, during which DNA is transcribed into mRNA (messenger RNA), a single-stranded complementary copy of the base sequence in the DNA molecule; 2) translation, during which mRNA is translated into protein, which are action molecules of the cell responsible for nearly all cellular processes. The majority of genes are expressed as the protein they encode. Although regulation of protein synthesis is not solely controlled by mRNA levels, the changes in protein abundance are determined in part by changes in the levels of mRNA. By measuring transcription levels of genes in organism under various conditions, at different developmental stages and different tissues, one can characterize the dynamic function of each gene in the genome. DNA microarray is a device for measuring transcription abundance of a large number of genes simultaneously under experimental conditions. This section describes two types of DNA microarray: complementary DNA (cDNA)-based and oligonucleotide-based arrays.

cDNA Microarray

There are, mainly, two platforms for cDNA microarray: nylon membrane filter arrays and chemically coated glass arrays. The nylon membrane arrays are suited for radioactivity labeling, while glass only supports fluorescence-based detection. A cDNA microarray consists of thousands of single-stranded known DNA fragments attached at fixed spots by a robotic arrayer. The known DNA fragments are selected from a cDNA library prepared from reverse transcription of mRNA from various tissues and organisms. Ideally, each gene fragment should represent a unique gene or alternative

Encyclopedia of Biopharmaceutical Statistics
DOI: 10.1081/E-EBS 120018493

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

splice variant. There are rat, mouse, and human arrays commercially available. Also, tissue-specific arrays can be constructed from a cDNA library extracted from the tissue.

In the experiment, the mRNA extracted from the tissue cell under study is purified, reverse transcribed into cDNA target, and labeled with radioactive markers or with green or red fluorescent dyes. Labeled cDNA hybridizes to the spots containing complementary sequence probes on the array based on base-paring rules. After hybridization, the radioactive or fluorescent signal intensities are imaged using a phosphorimager or laser scanner, respectively. One intensity is measured on each spot for the radiation-labeled array (one-channel array) while two intensities are measured on each spot for the fluorescence dye-labeled array (two-channel array). In both cases, the intensities are surrogates for the expression levels of genes in the sample under study. In this entry, the presentation is primarily for the two-channel array; however, the methods and procedures generally are applicable to the one-channel array.

The process of transforming the fluorescence intensity of each spot into a measure of transcript abundance is referred to as image analysis. In addition, image analysis includes evaluation of the quality of the quantified spot intensities and flags unreliable spots or arrays. Image analysis can be separated into three steps: 1) image acquisition, 2) spot location, and 3) intensity extraction. First, the microarray is scanned to produce two images, one for each channel. These images are stored as a 16-bit tag image file format (TIFF) file (a digital record of fluorescence intensities of pixels for the array), representing a number between 0 and $2^{16}-1$. As soon as the image has been captured, the next step is to locate each spot on the array. This can be performed automatically by the image analysis software. The image analysis software must segment the pixels as foreground (feature) or background. Once the foreground pixels and background pixels have been identified, the spot intensities are calculated using the mean (or median) intensity of all foreground pixels subtracting the mean (or median) intensity of the background pixels. The adjusted intensities are the primary data for subsequent analyses. The background-corrected intensity may be negative for some spots. Fig. 1 depicts the scheme of a two-channel microarray experiment.

Oligonucleotide Microarray

The oligonucleotide array technology was introduced by Lockhart et al.[2] and Lipschutz et al.,[3] and is commonly known as Affymetrix GeneChip™ (Santa Clara, CA). This technology does not use gene fragments extracted from a sequenced cDNA library as in cDNA microarray. Typically, each gene will be represented by 16–20 pairs of oligonucleotides referred to as probe sets. Each probe pair consists of 25 base sequences. One is perfectly complementary to a portion of a given transcript referred to as a perfect match (PM) probe. Each PM probe is paired with a mismatch (MM) probe that is created by changing the middle (13th) base pair to control possibility of cross-hybridization. RNA samples are prepared, labeled with array. Arrays are scanned, and images are produced to obtain an intensity for each probe. These intensities indicate how much hybridization occurred for each oligonu-

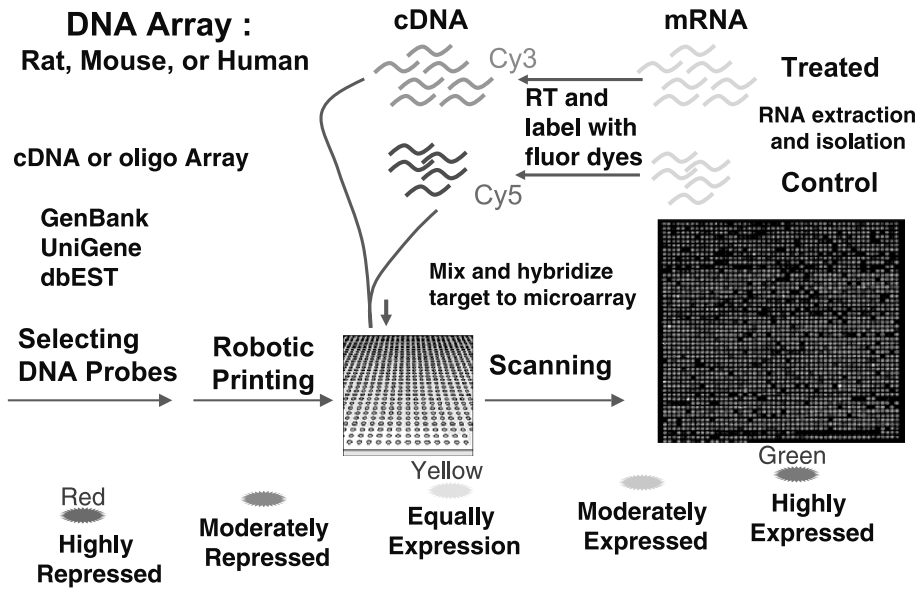

Fig. 1 Two-channel DNA microarray experiment.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

cleotide probe. The level of expression of each gene is calculated as an ''average'' of the difference between PM and MM, via a procedure provided by Affymetrix through their scanning software.

The Affymetrix arrays are constructed on a silicon chip by photolithography and combinatorial chemistry. The target labeling is performed using amplified RNA rather than cDNA. Hybridization to the array is noncompetitive (only a single sample is applied to each chip), and is detected by addition of a fluorescently labeled streptavidin compound that binds to the biotin group in the aRNA molecules.[4]

cDNA methods are more flexible in design. Researchers have control over the probe content on the array. It can be applied to any organism. The hybridization is based on over a thousand bases rather than 25 bases, thus reducing the chance of cross-hybridization artifacts. In contrast, oligonucleotide arrays can accommodate the genes not represented in cDNA library. It has lower variability from chip to chip, and can be used by researchers for data comparison across research groups.

Recently, a hybrid technology consisting of 50- to 80-mer oligonucleotides representing each gene was introduced. This technology combines the uniformity of GeneChip™ with the specificity of cDNA microarray. That is, the length of sequence is optimized to minimize cross-hybridization compared to shorter-length probes. Also, the oligo-array has higher sensitivity compared to cDNA probes.

EXPERIMENTAL DESIGN

Variability, Replication, and Experimental Unit

Microarray experiment is generally a comparative experiment, in which the experiment of interest is the comparison of the relative expression levels among the samples rather than the determination of absolute intensity measures of each sample. Furthermore, the underlying principle of a two-channel experiment is competitive hybridization between two samples. The measured intensities reflect relative abundances of the two samples. Data generated are inherently comparative. In a one-channel experiment, the intensity is an absolute measure of gene expression; however, this measure should not be regarded as the precise measure of transcript abundance. Inferences are made about the expression levels for a gene in different samples but not about the level of expression of one gene in relation to other genes.

Biological experimental data are inherently variable. Variability in gene expression measurements can be classified according to the source. Processes and biological variations are two general sources of variation. Biological variation refers to the variations from different RNA sources. Process variation refers to the variation arise from the use of microarray system. This variability is independent of the RNA source. Experimental design methods are used to identify and minimize experimental variations and provide sound statistical inference. The basic principles of experimental design are randomization, blocking, and replication. Replication is essential in the experimental design. Yang and Speed[5] described two types of replications: technical replicates and biological replicates. Technical replicate refers to replication in which the mRNA is from the same pool (the same extraction). The term biological replicates refers to hybridizations that involve mRNA from different extractions. If the mRNA is labeled separately and comes from different extractions, then the sample can be regarded as independent biological samples. The experimental units are the biological samples. The use of biological replicates allows the generalization of experimental results from sample to population.

Experimental design for one-channel microarray experiments such as Affymetrix oligonucleotide or nylon membrane array experiment is similar to clinical trials or other biological experiments. The same principle used in the biomedical experiment can be applied to microarray experiments. For example, the biological samples (experimental units) should be randomized to a treatment using a predetermined scheme. Blocking can be used to increase the precision of estimates. (A block is a subset of experimental units that are more homogeneous than the entire experiment itself.) In the two-channel experiment, two samples are labeled differently for competitive hybridization. One important design issue is to determine which samples are to be labeled with a corresponding fluor, and which are to be hybridized together in the same array. In addition, there can be constraints on the number of slides, the amount of RNA available, or cost considerations.[5]

Types of cDNA Experimental Design

The choice of experimental design does not only depend on the number of different samples to be compared, but also on the aim at the comparisons of interest. The simplest microarray experiment is intended to study the changes in gene expression levels between a control and a treated sample. Each microarray in the experiment is probed with two cDNA samples. The expression levels of the two samples can be compared for each gene in an array. In practice, however, the expression levels from the two measurements are not directly comparable because of dye effects. One dye may be consistently brighter than the other because of different labeling efficiencies, or different scanning sensitivities in the two dyes. Fig. 2[6] is a density plot (in log based-2 scale) of a same–same

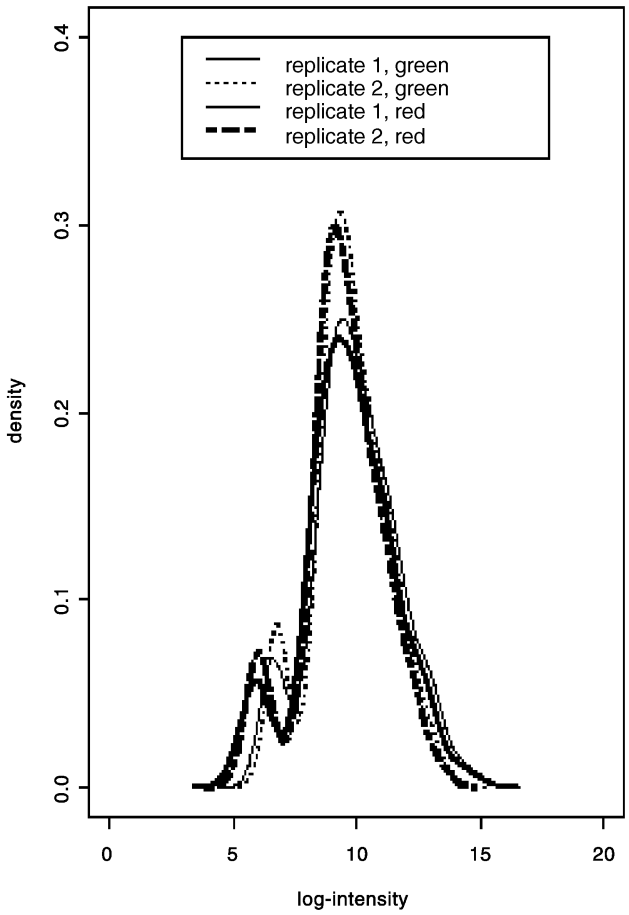

Fig. 2 Density plot of a same–same array.

(control vs. control) array. That is, two control samples are labeled separately, one with Cy5 (red dye) and one with Cy3 (green dye). Also, the upper half and lower half of the array are duplicate. In Fig. 2, the solid lines represent the genes from the upper half, and the dotted lines represent the lower half. The dark and light curves represent the red and green fluorescent readings, respectively. The plots show that there were systemic location effect differences in dye intensity across the individual DNA probes. (The left tail from the low intensity spots represents the empty genes.)

To adjust for dye biases, the so-called dye-swap design with two arrays is often used. On array 1, the control sample is assigned to the green dye, and the treated sample is assigned to the red dye; the dye assignments are reversed on array 2. Dye-swap design is a loop design (described below) for the experiment with a control and a treatment study. This design is useful to reduce dye biases, but it is unlikely to eliminate the bias in every spot of every array. Dye bias and variation can be reduced via the normalization method (see the section ''Normalization'').

Many experiments involve more than one treatment sample. Each experiment will have its own study objectives. Some studies may be interested in the comparisons of differences in the expression profiles among different toxins or different cell lines. In other studies, different samples may come from different time points or different dose levels. The main objectives of these two types of studies can be different. Kerr and Churchill[7] described the two designs for a single factor study: (augmented) reference design and loop design.

Reference design

In the reference design, all samples of interest (control and treatments) are hybridized on different arrays labeled with the same color dye, while a reference sample labeled with the other color dye is used on every array to hybridize with either a control or a treatment sample. One can denote a dye assignment in which the first sample is labeled with red and the second sample labeled with green as R/G. A reference design for an experiment with four groups—T0, T1, T2, and T3—can be expressed as [T0/Ref, T1/Ref, T2/Ref, T3/Ref]. The reference sample (Ref) can be the same as the control sample (T0) or a universal reference sample. This design has several advantages. The experiment is easy to conduct. The relative expression levels of the treated to control can be directly computed as observed responses. Because array consists of one measurement of relative change of the sample to a common reference for each gene, statistical methods are readily applicable. In this design, treatment effects are confounded with dye effects. It assumes there are no treatment $\times$ dye interaction effects.

Reference design can be inefficient because fully half of the data are dedicated to an extraneous sample. This type of design is an indirect comparison in which the reference sample serves as an intermediate between two treatments. Suppose the variance of a single log comparison $\log(T_i/\text{Ref})$ is σ^2, then the variance of the comparison $\log(T_i/T_j) = \log(T_i/\text{Ref}) - \log(T_j/\text{Ref})$ is $2\sigma^2$.

Loop design

In the loop design, the control and each treated group are labeled once with the red and once with the green. A loop design for the four groups—T0, T1, T2, and T3—can be [T0/T1, T1/T2, T2/T3, T3/T0] or [T1/T0, T2/T1, T3/T2, T0/T3]. The loop design uses the same number of arrays as a reference design does, but it collects more information on the treatments. The loop design compares two adjacent treatments directly. The variance between two groups hybridized in the same array $\log T_i/T(i+1)$ is σ^2 (but with only one array). Yang and Speed[5] provided

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

more details on the variance of reference design and loop design, and the time-course experiments and multifactorial designs.

Sample size

Before conducting a microarray experiment, one important issue that needs to be resolved is the number of arrays required to carry out a statistically interpretable result for the experiment. Lee et al.[8] suggested three replicates in view of the noise-to-signal ratio. Black and Doerge[9] proposed an approach using the lognormal and gamma model to determine sample size, while Pan et al.[10] proposed using normal mixture model normal. These approaches require modeling of the distribution of genes, and they do not consider gene-specific variances in expression.

Calculation of the number of arrays needed in a microarray experiment is similar to the sample size and power calculations in clinical trials or other experiments. However, microarray experiments involve hundreds or thousands of genes, and only a fraction of genes is expected to be expressed differentially. For genes that are affected by a treatment, the effect size can be different for various genes in the study. To estimate the number of arrays needed, the primary scientific aim of the experiment needs to be stated. Here, sample size estimation focuses on the identification of differentially expressed genes in a control and a treatment group comparison in the framework of hypothesis testing. Analysis is based on the one-sample or the two-sample t-test, one gene at a time. The problem is formulated as identification of at least $100\lambda\%$ of truly altered genes at a $1-\beta$ power at the given type I error rate and effect size.

For a given significance level α, the power of a two-sample t-test[11] for gene i is

$$1-\beta_i = F_t\left[\sqrt{(r/2)}\Delta_i - t_{\alpha/2}\right] \quad (1)$$

where Δ_i is the effective size (the desired change between the altered and unaltered gene to detect) and F_t is the cumulative t_{2r-2} distribution (Student's t-distribution with $2r-2$ degrees of freedom) and $t_{\alpha/2}$ is the $100(\alpha/2)$ percentile of the t_{2r-2} distribution. The sample size needed in each group is

$$r = 2(t_{\alpha/2} - t_{1-\beta_i})^2/\Delta_{i,2}^2$$

The power and the sample size given above is for the ith gene with the effect size Δ_i, $i = 1,\ldots,g$. In practice, only a fraction of the genes will be affected by a treatment in the experiment. The problem is formulated as the number of arrays needed to detect at least $100\lambda\%$ of the truly altered genes at the desired power $1-\beta$, where λ is a prespecified fraction, $0 < \lambda \leq 1$. The method described below assumes an equal effect size for all altered genes, and the effect size is standardized by its standard deviation. Let us denote the power for identifying an altered gene as γ $(= 1-\beta_i)$. Let k denote the number of truly altered genes, and the notation $t = b$ is the largest integer less than t. The power for identifying at least $k\lambda + 1 = b$ altered gene can be computed by summing the binomial probabilities:

$$1-\beta = \sum_{l=b}^{k} \frac{k!}{l!(k-l)!}\gamma^l(1-\gamma)^{k-l}$$

Given k, λ, and β, the γ can be estimated by solving the above equation. Thus the sample size is calculated by substituting γ into $1-\beta_i$ in Eq. 1 so as to achieve the objective of an overall $1-\beta$ power level and identify at least $100\lambda\%$ truly altered genes. The number of array needed is

$$r = 2(t_{\alpha/2} - t_\gamma)^2/\Delta_i^2$$

The power and the sample size for a one-sample t-test can be similarly obtained. The power for the one-sample t-test is

$$1-\beta_i = F_t\left[\sqrt{r}\Delta_i - t_{\alpha/2}\right] \quad (2)$$

where F_t is the cumulative t-distribution with $(r-1)$ degrees of freedom. The sample size estimate is

$$r = (t_{\alpha/2} - t_\gamma)^2/\Delta_i^2$$

NORMALIZATION

As discussed, there are many sources of variation in the measured intensity levels. Normalization of the expression data should be performed prior to statistical analysis. The purpose of normalization is to identify and adjust process variations (e.g., amount of RNA preparation for each sample, different labeling efficiencies and scanning properties of the dyes, print-tip, or spatial effects). Normalization can be regarded as a preprocessing step or a transformation before the data analysis. Normalization strategies depend on the experimental design and the data collection process. The primary consideration is to determine the (sub)set of genes for use as the baseline for a normalization method. That is, normalization can be based on either entire set of genes or certain subset of genes, such as housing keeping genes that are not regarded to change the expression level under the experimental condition or some added set of control genes. Various normalization methods have been proposed.[6,12–19] Two assumptions are commonly made on normalization: 1) most genes do not change their expression level;

2) the total intensity across all spots in the array should be the same. The global normalization method uses all genes on the array for adjustment; it uses the mean or median as the normalization factor based on the two assumptions. For each array, the data are normalized by dividing individual gene intensity by the mean or median intensity of the array. However, the global mean/median normalization method fails to account for spatial- and intensity-dependent biases that have been observed in many experiments.

Two known statistical methods proposed for normalization are: the analysis of variance (ANOVA) model[14,15] and the scatter-plot smoother ''lowess'' fit.[19] Although both procedures consider two-channel fluorescence array, they can be applied to one-channel data with modifications. The ANOVA procedure[15] recommends modeling the logs of individual red and green intensities. The ANOVA model uses mean estimates to perform global adjustments for array, dye, treatment, gene effects, and their appropriate interactions simultaneously. However, as in the case of global normalization, the ANOVA method is a constant adjustment, with respect to specific effects such as dye biases. Alternatively, the lowess fit[19] uses local regression to account for intensity- and spatial-dependent dye biases. The lowess fit is a scatter-plot smoother that performs robust locally linear fits. Yang et al.[19] recommended performing an array-by-array normalization on the log-ratios instead of the log-ratio of two intensities. Unlike the ANOVA approach, this lowess approach fits different lowess cures for different arrays. These two procedures can be formulated by a generalized additive model[20] (shown below). This entry does not address location-specific (e.g., print-tip) effects. They are discussed in detail by Yang et al.[19] and Chen et al.[6]

Generalized Additive Model

Consider a two-channel microarray experiment. Let y_{ijks} denote the base-2 logarithm of the (background corrected) intensity measurements for the ith gene ($i = 1, \ldots, g$) in the jth treatment ($j = 1, \ldots, t$) for the kth dye ($k = 1, 2$) on the sth array ($s = 1, \ldots, a$). Let us assume that the three-factor interaction is negligible. Consider the standard linear model,[14]

$$y_{ijks} = m + G_i + T_j + D_k + A_s + (GD)_{ik} + (GA)_{is} + (TD)_{jk} + (DA)_{ks} + (GT)_{ij} + \epsilon_{ijks}$$

The interaction of genes and treatments $(\text{Gene} \times \text{Trt})_{ij}$, which captures the expression of the gene i specifically attributable to the treatment j, is the effect of interest. The treatment effect T_j is typically nested within the interaction of dye and array effects $(DA)_{ks}$. In the ANOVA model, the parameters are conventionally estimated in terms of cell means under the normal model. The normalized data from the estimates of the interaction (Gene × Trt) are analyzed as raw data using statistical methods such as parametric or bootstrap methods. The ANOVA procedure can be regarded as a normalization by adjusting several factors simultaneously.

The ANOVA model can be decomposed into two submodels

$$y_{ijks} = f_{ks}(G_i, D_k, A_s) + [T_j + (TD)_{jk} + (GT)_{ij}] + \epsilon_{ijks}$$

The first submodel is a regression model $f_{ks}(G_i,D_k,A_s)$ consisting gene, dye, and array-related effects. The second submodel includes the remaining effects. This model can be estimated by a two-step procedure.

The first step involves fitting a smooth function for $f_{ks}(G_i, D_k, A_s)$ to account for the nonlinear effects of gene, dye, array, and interaction factors. Let us denote $\bar{y}_{i.k.} = \Sigma_{j,s} y_{ijks} = \Sigma_{j,s}$ and $\bar{y}_{i..s} = \Sigma_{j,k} y_{ijks} = \Sigma_{j,k}$. A generalized additive model describing gene-specific array and dye effects is

$$f_{ks}(G_i, D_k, A_s) = \alpha_{ks} + f_s(\bar{y}_{i..s}) + f_k(\bar{y}_{i.k.}) + \epsilon_{ijks}$$

where $E\{f_s(\bar{y}_{i..s})\} = E\{f_k(\bar{y}_{i.k.})\} = E(_{ijk}) = 0$. This model does not fit any interaction effect among genes, arrays, and dyes. It can be estimated using a suitable robust scatterplot smoother, for example, the lowess method of Cleveland[21] used by Yang et al.[19] Let us denote the fitted regression function as $\hat{y}_{ijks} = \hat{f}_{ks}(G_i, D_k, A_s)$. The second step is to apply the linear ANOVA model in estimating the remaining effects including the interaction $(GT)_{ij}$ from the residuals,

$$y_{ijks} - \hat{y}_{ijks} = T_j + (TD)_{jk} + (GT)_{ij} + \epsilon'_{ijks}$$

Let us denote the gene by treatment interaction estimates as $r_{ij} = \widehat{GT}_{ij}$. The r_{ij}'s are the normalized data and can be treated as raw data in data analysis. For example, treatment effects can be analyzed by the gene-specific ANOVA model

$$r_{ij} = m + T_j + e_{ij}$$

That is, the normalized data are analyzed one gene at a time. Because the distribution of the residuals (normalized data) is typically not normal, nonparametric approaches such as randomization tests or bootstrap procedures are recommended for subsequent analyses.

The generalized additive model consists of two components: an array-specific function $f_s(\bar{y}_{i..s})$ and a dye effect function $f_k(\bar{y}_{i.k.})$. It may not be necessary to adjust for both array and dye effects in the application. The model for an array-specific normalization is

$$y_{ijks} = f_s(\bar{y}_{i..s}) + \epsilon_{ijk}$$

The second step ANOVA model is

$$y_{ijks} - \hat{f}_s = (\text{Gene} \times \text{Trt})_{ij} + (\text{Dye})_k + (\text{Dye} \times \text{Array})_{ks} + (\text{Gene} \times \text{Dye})_{ks} + \epsilon'_{ijks}$$

Normalization of Intensity Ratios

Consider a reference design that includes control, treatment, and reference samples. The control and treatment samples are labeled with the same color dye, and reference samples are labeled with another color. The experiment of interest is the control and treatment comparison. Let y_{ijks} be the log intensity from gene i, treatment j, dye k, and replicates s, where $j = 1,2,3$ represent control, treatment, and reference samples, respectively. Consider the ANOVA model

$$y_{ijks} = \mu + G_i + T_j + D_k + (GT)_{ij} + (GD)_{ik} + \epsilon_{ijks}$$

This model has no array effect because every array can be determined by a treatment and a replicate combination. The data can be analyzed in terms of the log-ratios.[19] Let $y'_{ijs} = y_{ij1s} - y_{i32s}$ for $j = 1,2$. The above ANOVA model can be simplified as

$$y'_{ijs} = m + G_i + [T_j + (GT)_{ij}] + \epsilon_{ijs}$$

Applying the generalized additive model, the first-step model is

$$y'_{ijs} = f(\bar{y}'_{i..}) + \epsilon_{ijs}$$

where $\bar{y}'_{i..} = \Sigma_{j,s} y'_{ijs}/(ra)$. The second-step ANOVA model is

$$y'_{ijs} - \hat{f} = T_j + (GT)_{ij} + \epsilon'_{ijs}$$

In this analysis, only one lowess function $f(\bar{y}'_{i.})$ is fitted to all data (there are ra arrays) in the first-step normalization. One alternative is to fit the lowess function $f_s(\bar{y}'_{i.s})$, $s = 1, \ldots, ra$. That is, there is one lowess function for each array, which is the array-by-array normalization procedure of Yang et al.[19] However, the use of array-wise lowess normalization may oversmooth treatment effects and bias test results.[20]

IDENTIFYING DIFFERENTIALLY EXPRESSED GENES

The simplest microarray experiment is performed to identify a subset of genes that are differentially expressed between control and treated groups. In general, a gene is said to be differentially expressed if the ratio in absolute value of the expression levels between the treated group to the control exceeds a certain threshold, e.g., 2-fold or 4-fold change. This approach is deficient in some respects. For example, the ratio at the lower levels can be more diverse than that at the higher levels. More seriously, it does not take into account the variability of the expression levels and does not allow direct comparison of different sources of variance.

Because of the lack of replications, the early statistical approach for assessing differentially expressed genes is based on modeling the distribution of entire gene sets in the array. Chen et al.[12] assumed that the expression levels are independently normally distributed with a constant coefficient of variation. Other distributions such as lognormal, gamma, or mixture models have also been proposed.[9,12,22,23] Recently, statistical tests commonly utilized for identifying differentially expressed genes were analyzed using variants of t-test or ANOVA-test statistics.[24–26]

Statistical Models

Let us denote the background-subtracted and normalized intensity (in log based-2 scale) for control and treated groups as Y_{ijc} and Y_{ijt}. A model with two sources of variation for gene expression intensity[27] is

$$Y_{ijc} = \mu_{ic} + b_{ic} \cdot \eta_{ijc} + \epsilon_{ijc}$$
$$Y_{ijt} = \mu_{it} + b_{it} \cdot \eta_{ijt} + \epsilon_{ijt}$$

where (μ_{ic}, μ_{it}) represents the paired true expression levels for control and treated samples for spot (gene) i and replicate j, $i = 1, \ldots, g$ and $j = 1, \ldots, r$; b_{ic} and b_{it} are constant. The errors (η_{ijc}, η_{ijt}) and $(\epsilon_{ijc}, \epsilon_{ijt})$ are referred to as the multiplicative error and additive error, respectively. For each gene i and replicate j, the multiplicative error (η_{ijc}, η_{ijt}) and the additive error $(\epsilon_{ijc}, \epsilon_{ijt})$ are assumed to be independently and identically bivariate-normally distributed, $(\eta_{ijc}, \eta_{ijt}) \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{\Phi})$ and $(\epsilon_{ijt}, \epsilon_{ijt}) \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Phi} = \begin{bmatrix} \phi_c^2 & \tau\phi_c\phi_t \\ \tau\phi_c\phi_t & \phi_t^2 \end{bmatrix} \quad \text{and}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_c^2 & \rho\sigma_c\sigma_t \\ \rho\sigma_c\sigma_t & \sigma_t^2 \end{bmatrix}$$

The errors (η_{ijc}, η_{ijt}) and $(\epsilon_{ijc}, \epsilon_{ijt})$ are independent of one another. The correlations τ and ρ are zero for the one-channel experiment. The distributions of Y_{ijc} and Y_{ijt} are

$$Y_{ijc} \overset{\text{i.i.d.}}{\sim} N(\mu_{ic}, b_{ic}^2\phi_c^2 + \sigma_c^2), \quad \text{and}$$

$$Y_{ijt} \overset{\text{i.i.d.}}{\sim} N(\mu_{it}, b_{it}^2\phi_t^2 + \sigma_t^2)$$

The covariance between Y_{ijc} and Y_{ijt} is $(b_{ic}b_{it}\tau\phi_c\phi_t + \rho\sigma_c\sigma_t)$.

The model given above is a fixed effect model. The multiplicative errors can be used to model an array-specific (random) effects model by letting $\eta_{ijc}, \eta_{ijt}) \equiv (\eta_{jc}, \eta_{jt}) \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{\Phi})$. All genes on the same array have the same variance. The covariance between the spots i_1 and i_2 on the same array (j) for the control and treated groups are $\text{Cov}(Y_{i1jc}, Y_{i2jc}) = b_{i1c}b_{i2c}\phi_c^2$, and $\text{Cov}(Y_{i1jt}, Y_{i2jt}) = b_{i1t}b_{i2t}\phi_t^2$, respectively. When $b_{ic} = b_{it} = 1$, it becomes a linear mixed effects model.[17]

Significance Testing

Let $\bar{Y}_{ic}$ and $\bar{Y}_{it}$ denote the means of the r replicates in the control and r replicates for the treatment, respectively; similarly, s_{ic}^2 and s_{it}^2 represent the sample variances. Identification of differentially expressed genes between the control and treatment groups can be carried out for each gene by computing the two-sample t-statistic

$$t_{i,2} = (\bar{Y}_{ic} - \bar{Y}_{it})/\hat{\sigma}_{i,2}$$

where $\hat{\sigma}_{i,2}^2 = [(s_{ic}^2)/r + s_{it}^2/r]$. Under the model of an equal variance, the variance estimate is

$$\hat{\sigma}_{i,2}^2 = (r/2)[(r-1)s_{ic}^2 + (r-1)s_{it}^2]/(2r-2)$$

If there is no difference between the two groups, then $t_{i,2}$ has a t-distribution with $(2r-2)$ degrees of freedom.

The comparison between the control and treatment groups can also be tested by a one-sample t-statistic. Let $T_{ij} = Y_{ijc} - Y_{ijt}$ and $\bar{T}_i$ be the mean of T_{i1},SSPI. . ., $T_{i,r}$. If there is no difference between the two groups, then the one-sample statistic

$$t_{i,1} = (\bar{Y}_{ic} - \bar{Y}_{it})/\hat{\sigma}_{i,1}$$

has a t-distribution with $r-1$ degrees of freedom, where $\hat{\sigma}_{i,1}$ is the standard error estimate of $\bar{T}_i$. Both one-sample and two-sample t-tests have been used for a two-group comparison[24,26] under the assumption of an equal variance. The one-sample t-test is a more powerful test if the samples are correlated, such as two-channel data from a dye-swap experiment.

Generally, the distribution of the normalized intensities is not normal. The permutation tests provide an alternative method to determine the significance. The principle of the permutation test is conceptually straightforward. It first computes the empirical value of the test statistic for the observed data, and then performs all possible permutations under the null hypothesis and computes the test statistic for each permutation. The p-value of the test is calculated by summing the probability of the observed outcome and the probabilities of all outcomes with the test statistic values greater than the observed value. Permutation tests do not require assumption on the distributions; they are easy to perform, using high-speed personal computers or workstations. However, when the number of replicates is small (less than or equal to 5 for the two-sample test and less than or equal to 4 for the one-sample test), permutation test is not recommended.[28]

The p-value (equivalently, the t-value) represents the order of evidences for differential expression. Because the t-statistic is sensitive to small s_i (nearly constant expression genes), the penalized t-statistic can also be proposed to ranking genes,[29,30]

$$t_i = (\bar{Y}_{ic} - \bar{Y}_{it})/(\hat{\sigma}_i + s_0)$$

where s_0 is the penalty factor. Tusher et al. chose s_0 to minimize the coefficient of variation of the absolute t_i, while Efron chose s_0 to be the 90th percentile of $\hat{\sigma}_i$. The p-value of the penalized t-statistic can be computed either by permutation or bootstrap method.

Multiple Hypothesis Testing

A microarray experiment often involves comparisons of hundreds or thousands of genes. Because a large number of genes is compared, the simple use of a significance test without adjustment for multiple comparison artifacts could lead to a large chance of false positive findings. This issue is referred to as multiple hypotheses testing or testing multiple endpoints. The familywise error rate (FWE) and false discovery rate (FDR) approaches have been proposed to control the false positive rates in the analysis of gene expression data.[31–34]

Let us consider testing m null hypotheses (genes). Assume that there are m_0 true null hypotheses; the remaining $m - m_0$ nontrue null hypotheses may not be from the same population. According to the true state of nature, either the null or nonnull hypothesis is true. The results from m tests can be summarized as a $(m - m_{0+1}) \times 2$ table:

	Declared significant	*Declared nonsignificant*	*Total*
True null hypotheses	V	$m_0 - V$	m_0
Nontrue null hypotheses			
Hypothesis 1	U_1	$1 - U_1$	1
.	.	.	.
Hypothesis $(m - m_0)$	$U_{(m-m_0)}$	$1 - U_{(m-m_0)}$	1
Total	R	$m - R$	m

The binary random variable $U_l = 1$ if the lth nontrue hypothesis is rejected, and $U_l = 0$ if otherwise

$(1 \leq l \leq m - m_0)$. Comparison error rate (CWE) α is the probability of rejecting a true null hypothesis. FWE is the probability of rejecting at least one true null hypothesis in a given family of hypothesis tests, $\Pr(V \geq 1)$. FDR is the expected proportion of errors among the rejected hypotheses

$$\text{FDR} = E(V/R \mid R > 0)\Pr(R > 0)$$

Storey[34] defined a positive FDR as conditional on the event that positive findings have occurred,

$$\text{pFDR} = E(V/R \mid R > 0)$$

Another error measure is the conditional cFDR,

$$\text{cFDR} = E(V/r \mid R = r)$$

where r is the observed number of rejected null hypotheses.

The Bonferroni technique is the most well-known approach to controlling the FWE. If the number of true hypotheses is m, then using the significance level of $\alpha_0 = m$ for the individual tests, we shall obtain an FWE of less than or equal to α_0. After rejecting the hypothesis with the smallest p-value, the Holm[35] step-down procedure can then be applied to the remaining hypotheses for further improvement of the power. Westfall and Young[32] proposed the resampling method to compute the distribution of the minimum of the p-value. The resampling method is less conservative than the Bonferroni method.

The controlling FWE approach could present a problem in the analysis of microarray data, because this procedure tends to screen out all but a handful of genes that show extreme differential expressions. Instead of controlling the FWE, the investigator might be more interested in identifying all potential genes that are differentially expressed, even if some genes are falsely identified. Investigators often perform further tests, such as reverse transcription polymerase chain reaction (RT-PCR) or Northern blots, to validate the positive results. Benjamini and Hochberg[33] proposed controlling the FDR as an alternative to controlling the FWE. When the number of true hypotheses is not m, the FDR is smaller or equal to FWE. This implies that if a procedure controls the FWE, then it controls the FDR as well. Tusher et al.[29] proposed a significance analysis of microarray (SAM) method to estimate an FDR, instead of controlling FDR, by data permutations. Storey[34] proposed a method to estimate both FDR and pFDR. The FDR and pFDR methods are available in the SAM software (http://otl.standford.edu).

CLUSTER ANALYSIS AND PATTERN RECOGNITION

One important goal in the analysis of gene expression data is to determine the relationships between genes or gene clusters for identifying biological functions or predicting specific biological outcomes (or diseases) from the analysis of expression patterns. Clustering and visualization methods have been developed and applied to organizing gene expression data by grouping genes with similar patterns of expression. Usually, genes with previously unknown functions are simply annotated with functions of the known genes in the same cluster because they share similar expression profiles across different experimental conditions. Genes expressed with similar patterns may be coregulated by common upstream regulatory motifs or may belong to the same biochemical pathways. A cluster of unknown genes with homogeneous expression profiles may indicate the discovery of a novel function group.

Before any clustering and visualization procedures can be applied to the gene expression matrix, careful extraction of array images and appropriate normalization of system variations are necessary. A screening (data filtering) procedure is often applied in selecting a subset of genes that satisfy certain criteria of expression patterns across all arrays (treatments) for further statistical analyses. A commonly used criterion is to include genes with the expression $|\log_2 \text{Trt/Ref}| \geq 1$ in at least some experimental conditions. Minimum variation (standard deviation or range) criteria are also frequently applied. The screening process is needed for reducing the number of genes from tens of thousands to thousands or hundreds for clearer visualization, faster computation, and better interpretation. In addition to array normalization and centering, individual gene normalization is often used to scale the relative magnitude of a gene's expression profile (vector) to one. Gene normalization do not only prevent clustering procedures from being dominated by genes with extreme expression patterns; they also prevent visualizations from losing color resolution.

In this section, a gene expression time course data set from a serum stimulation of primary human fibroblasts[36] is adopted to illustrate concepts of statistical visualization and clustering tools in gene expression analyses. The experiment used cDNA arrays with 9800 spots, representing approximately 8600 distinct human transcripts. Briefly, fibroblasts were grown in culture and sera were derived for 48 hr. Serum was added back and samples were taken during various periods: at the start (designated as 0), 15 min, 30 min, 1 hr, 2 hr, 3 hr, 4 hr, 8hr, 12 hr, 16 hr, 20 hr, and 24 hr (with one array for each of these 12 samples). Genes were selected for their analysis if the expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points. The selection resulted in a

gene expression matrix of size 517 (genes) × 12 (arrays). A further randomly selected subset of 100 genes is used in the illustration.

Gene Expression Matrix Map with Hierarchical Clustering

The most frequently used visualization tool for gene expression profiles is the gene expression matrix map. There are many synonyms for gene expression matrix map in the literature. Wen et al.[36] made one of the earliest applications of matrix map to gene expression data, and Eisen et al.[37] laid down the fundamentals of cluster analysis of gene expression patterns with matrix map. Numerical data matrix of gene expression intensities (single channel) or log ratios (two channel) are projected through properly chosen single directional (single channel) gray spectrum or bidirectional (two-channel) gray scales. Fig. 3a is the log ratio (base 2) gene expression matrix with the 100 randomly selected genes on 12 time course arrays. A green–black–red spectrum is used to represent negative–zero–positive log ratio values with brighter (darker) colors representing extreme (mild) differentially expressed gene-array combinations. The 100 row color strips represent 100 gene expression profiles that are randomly permuted, while the order of 12 column strips representing 12 arrays remain intact to depict the time course nature of the experiments. Very little information can be visually extracted from Fig. 3a because rows (genes) are randomly permuted. It is necessary to rearrange the 100 rows to satisfy the concept of relativity,[38] such that genes with similar (different) expression profiles are placed in closer (distant) rows before any interesting pattern can be summarized from this expression matrix map.

The most commonly used permutation mechanism for rearranging row orders (gene) and column order (array) in an expression matrix map is a hierarchical clustering tree with a binary dendrogram representing the association structure of pair-wised genes or arrays. A proximity (similarity or distance) matrix, such as correlation matrix, is first generated. Fig. 3b is the Pearson correlation matrix of the 100 gene expression profiles with a blue–white–red bidirectional spectrum representing the negative–zero–positive correlation coefficients. A similarity proximity matrix is usually converted into a distance matrix before the dendrogram for that matrix is constructed. The second step is to identity the pair of genes with the smallest distance and to group them with a link. Distances between the newly formed group with the rest of the 98 genes are computed resulting in a new proximity matrix of size 99×99. The algorithm proceeds in a recursive manner to build the tree structure step by step, with one less cluster member at each step. The final tree architecture is displayed as the left panel of Fig. 3c. The relative positions of terminal leaves in this tree form a linear order of the 100 genes, which is used to sort the randomly arranged input correlation matrix map into the structured matrix in the middle panel. The randomly ordered gene expression map in Fig. 3a is rearranged into the right panel of Fig. 3c.

There is a possible flip associated with every intermediate node in the tree architecture. A tree with 100 genes has 99 intermediate nodes with 2^{99} possible flip combinations and 2^{99} different permutations in rearranging the gene orders. External and internal references can

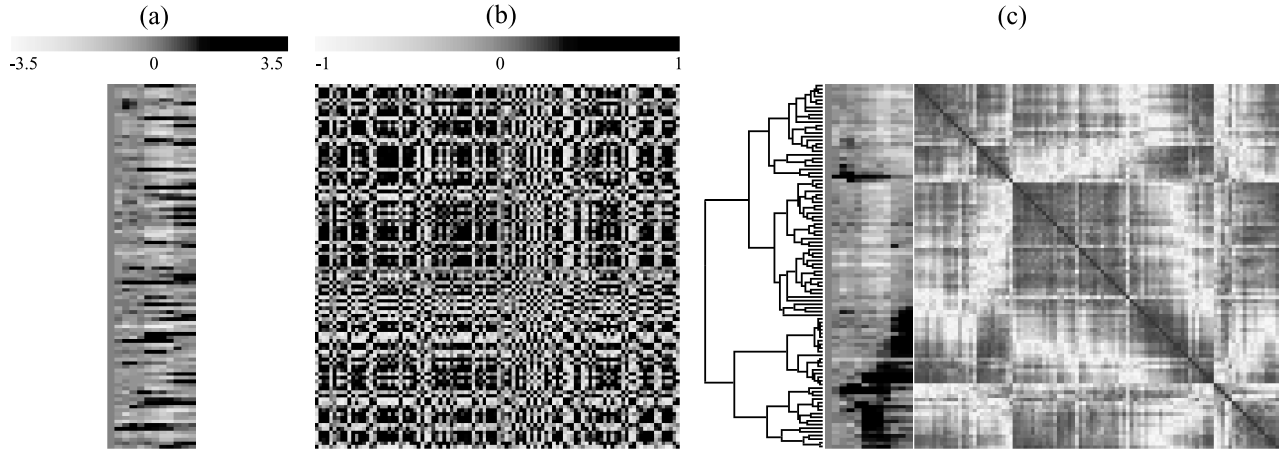

Fig. 3 Gene expression matrix map with hierarchical clustering of 100 randomly selected genes from a time course data set with serum stimulation of primary human fibroblasts.[37] (a) Gene expression matrix map for the 100 genes on 12 time points (arrays) with a green-black-red color spectrum; (b) correlation matrix map for the 100 genes with a blue-white-red color spectrum; (c) hierarchical clustering tree with sorted expression matrix map and correlation matrix map.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

be applied to these flips, such that certain criteria may be satisfied.[39,40] Chen[38] proposed a GAP (Generalized Association Plot) algorithm for identifying more global and smoother clustering patterns. Variants of hierarchical clustering algorithms have been developed, according to the direction (agglomerative vs. divisive) of trees that are constructed and the way new distances are computed (single linkage, complete linkage, average linkage, centroid method, and Ward's method). Sokal and Sneath[41] describe methodologies for constructing hierarchical clustering trees and nonhierarchical clustering algorithms in details.

Nonhierarchical Clustering

K-means[42] and self-organizing maps (SOM)[43,44] are the two most commonly used nonhierarchical clustering algorithms in gene expression pattern analysis. In the nonhierarchical clustering procedure, genes are divided into k ($k_1 \times k_2$ for SOM) partitions or groups, with each partition representing a cluster of genes. Therefore as opposed to the hierarchical clustering, the number of clusters must be known (decided) a priori.

In K-means clustering, the user first specifies k initial cluster centroids or seeds. The proximities (similarity or distance) from each gene to all k centroids are calculated. Each gene is then assigned to the closest cluster (centroid). The k new centroids are formed with new cluster members, and the genes are reallocated to one of the new k clusters. This iterative process stops if there is no reallocation of genes, or if the reassignment satisfies the criteria set by the stopping rule. Variants of K-means algorithms differ with respect to: 1) the method used for obtaining initial cluster centroids or seeds; 2) the rule used for reassigning genes; and 3) the proximity for computing the closeness between each gene and every centroid. Because K-means method outputs only the cluster memberships for all genes, this membership information is usually displayed in colors or symbols (Fig. 4a with $k = 5$).

The SOM method combines the features of dimension reduction in multidimensional scaling (MDS) and

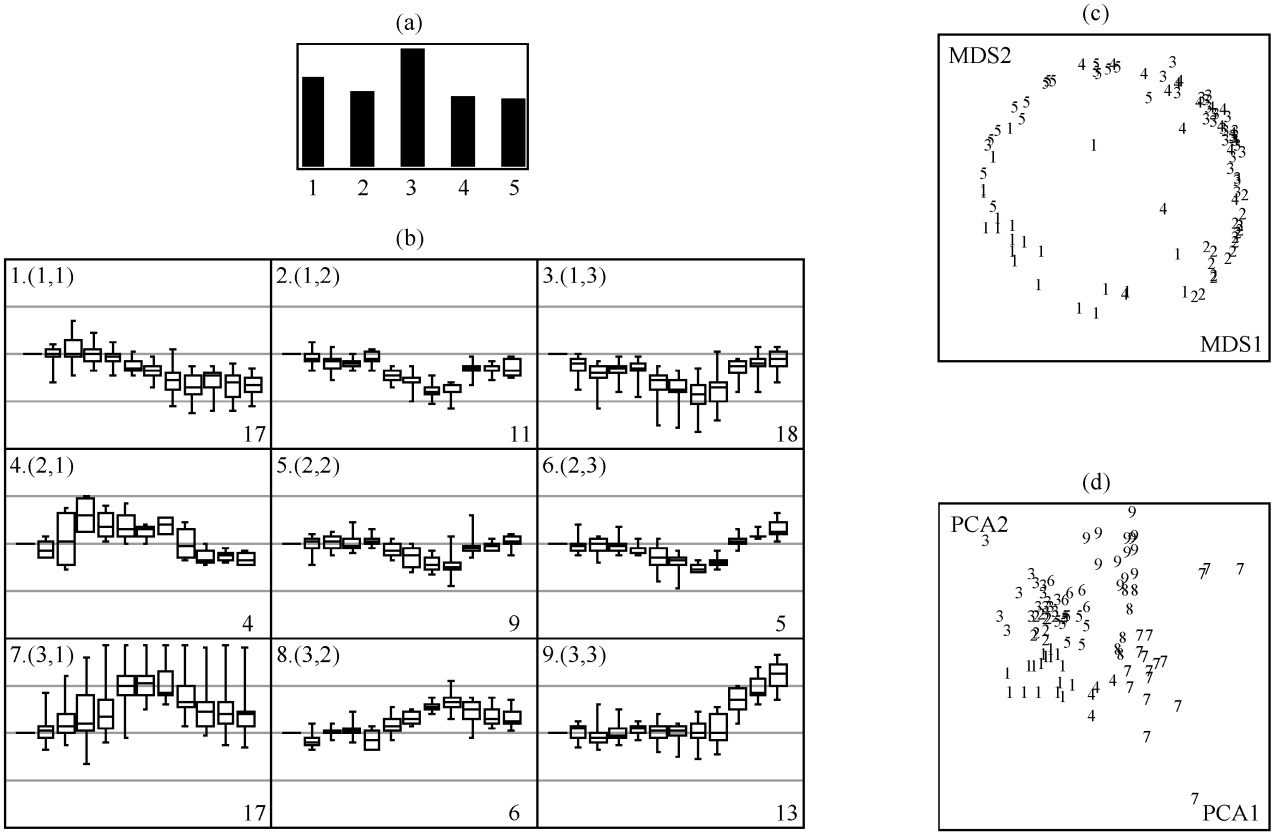

Fig. 4 K-means clustering and self-organizing maps (SOM) with multidimensional scaling (MDS) and principal component analysis (PCA) for the time course data set with 100 genes. (a) K-means ($k = 5$) clustering with cluster membership by symbols; (b) SOM with a 3×3 grid structure, expression profiles for genes grouped at each grid point are displayed as box-plots with cluster membership by symbols; (c) MDS with K-means membership symbols; (d) PCA with SOM membership symbols.

prespecified number of clusters in K-means. The goal is to represent genes in the original high-dimensional input expression space by grid points in a low-dimensional output space. Commercial packages usually limit the output space to a one-dimensional grid line or two-dimensional grid net for visualization purpose. Similar to K-means algorithm, these grid points are projected to the original high-dimensional input space as centroids for the gene expression profiles to be iteratively assigned to a cluster membership. The computation of cluster membership is confined to the grid structure in the low-dimensional output space. Final cluster membership is displayed using the output space grid structure. Gene expression profiles for genes belonging to each grid point are displayed as a multivariate statistical plot at the relative grid location. Commonly used statistical plots include parallel coordinate plot, side-by-side box-plots, and gene expression matrix map. An SOM analysis with a two-dimensional 3×3 grid and box-plots for the 100 randomly selected genes is illustrated in Fig. 4b.

Simulation studies have shown that K-mean algorithms and other nonhierarchical clustering algorithms perform poorly when random initial seeds are used. Their performance is improved when the results from hierarchical methods are used to form the initial partition.[45] In other words, hierarchical and nonhierarchical techniques should be applied as complementary clustering techniques, rather than as competing techniques. The k-means or SOM output membership information can be used in dimension reduction plots, principal component analysis (PCA), or MDS (to be shown in the subsection ''Dimension Reduction Analysis and Visualization'').

Parallel to the applications of unsupervised clustering algorithms in exploring gene expression profiles for annotating genes with novel functions, several supervised discriminant algorithms have been adopted for classification of cancer subtypes and gene functions. These algorithms have found applications in the following: Golub et al.[46] applied a neighborhood analysis for selecting genes that had strong discriminating power in separating acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) patients; Brown et al.[47] compared the effectiveness of gene function prediction of support vector machines (SVM)[48,49] to Fisher's linear discriminant analysis,[50] nonparametric kernel classification,[51] and decision tree learner;[52] Yeang et al.[53] compared various classification algorithms for the classification of multiple tumor types using gene expression profiles.

Dimension Reduction Analysis and Visualization

There are two major statistical techniques for simultaneously visualizing thousands of gene expression profiles in a single display—dimension reduction visualization and dimension-free visualization. Because of its high-dimensional nature, dimension reduction techniques such as PCA[54,55] and MDS[56,57] are often applied to gene expression matrices for projecting the original high-dimensional structure onto a lower dimensional space (usually two or three) for visualizing and computational purposes. Dimension reduction visualization is often adopted for presenting gene grouping structure for methods such as K-means and SOM.

For a given gene expression matrix X of size g (genes) by a (arrays), PCA performs an eigenvalue decomposition of the $a \times a$ covariance matrix $\Sigma = X'X/g$ as $\Sigma \mathbf{v}_k = \lambda \mathbf{v}_k$, where $\mathbf{v}_k$'s are the eigenvector of Σ with eigenvalue λ_k, $k = 1,\ldots,a$. The PCA summarizes the dispersion of gene expression profiles as data cloud in a small number of major axes (principal components) of variation among the arrays. Alter et al.[54] named these major axes as the eigenarrays of the original expression matrix. They normalized the arrays by filtering out the eigenarrays that were inferred to represent noise or experimental artifacts. Hastie et al.[58] recursively applied the PCA to gene expression matrices for identifying subsets of genes with coherent expression patterns and large variation across conditions. Fig. 4d is the plot of the first two principal components of the 100 by 12 gene expression profiles with colors correspond to gene clusters identified by the SOM analysis in Fig. 4b.

Given a set of n genes (or n arrays), among which the associations are to be studied, MDS first computes proximity (distance or similarity measurement) for all pairs of objects (genes or arrays) as δ_{rs}, $1 \le r, s \le n$. MDS then converts the input proximities into disparities $\hat{d}_{rs}$, $1 \le r, s \le n$ with a transformation function f. The function f can be the identity function if proximity δ_{rs} itself is a Euclidean distance. MDS finally represents objects as points in a Euclidean space, so that the perceived distances d_{rs} between points can reflect proximities δ_{rs} between objects. For visualization purposes, the dimension of the projected space is usually kept as low as possible. Thus it is unavoidable that part of the information in the original proximity matrix of similarity (or dissimilarity) will be lost in the MDS configuration plot. Kruskal[59] proposed a scheme for measuring how much a lower-dimensional geometrical representation falls short of a PM. This measure, called STandardized REsidual Sum of Squares (STRESS), is defined as

$$\text{STRESS} = \left\{ \frac{\sum_r \sum_s (d_{rs}^{(q)} - \hat{d}_{rs}^{(q)})^2}{\sum_r \sum_s (d_{rs}^{(q)})^2} \right\}^{1/2}$$

where q (usually 2 or 3) is the dimension of the solution space. A two-dimensional MDS configuration plot

using the Pearson correlation matrix of the 100 gene expression profiles is displayed in Fig. 4c, with colors corresponding to gene clusters identified by the K-means analysis in Fig. 4a.

Gene Network Modeling and Integrated System Biology

One of the fast growing research related to gene expression data analysis is the integration of expression profiles to the existing biological databases for validating and predicting gene networks. The analysis of gene expression profiles becomes much more powerful if knowledge about other biological databases such as interaction (gene–gene, gene–protein, protein–protein) and pathways (metabolic, regulatory) can be incorporated into the modeling frameworks. DeRisi et al.[60] used global analysis of changes in gene expression to validate known metabolic structure for diauxic shift in yeast (*Saccharomyces cerevisiae*). Marcotte et al.[61] proposed integrating information about gene expression, protein homologues, phylogenetic profiles, metabolic function, and experimental data into genomewide prediction of protein functions. D'haeseleer et al.[62] discussed applications of clustering algorithms and reverse engineering, such as Boolean networks to measure the output of the gene regulatory network from gene expression data. Jenssen et al.[63] used pure information retrieval tools to extract biomedical knowledge from publicly available gene and text databases to create a gene-to-gene cocitation network for more than 10,000 named human genes.

CONCLUSION

DNA array experiments are novel biotechnologies that have been increasingly used in biological and medical research. The mass of data generated from a typical DNA array experiment raises numerous statistical and computational challenges in data displays and processing, experimental design, and multiple testing, cluster analysis, and prediction. There are inherent biases in microarray data due to spatial, intensity, and dye effects within an array and array-to-array variation across experimental samples. Normalization is an integral part of microarray data analysis. Design consideration for one-channel experiment is similar to the parallel design in clinical trial. For two-channel experiment, the loop design appears to be more efficient than the reference design when the number of treatment is small. When the number of treatments is large, optimal design will depend on the objectives of the experiment. Several authors have considered other design issues, such as pooling mRNA samples.[64,65] Two statistical approaches to data analysis are discussed—significance testing for the identification of differentially expressed genes and cluster analysis for grouping the genes according to the expression profiles. Significance testing focuses on traditional gene by gene analysis, while cluster analysis focuses on the study of relationships among a set of genes or association between genes and samples. The field of microarray expression analysis is booming, and new analytic methods are under development. The literature is filled with novel analysis methods. New procedures for studying gene expression are being continuously developed. Generally accepted normalization procedures and data analysis methods, based on sound statistical principles, may be possible.

REFERENCES

1. The chipping forecast. Suppl. Nat. Genet. **1999**, *21*.
2. Lockhart, D.J.; Dong, H.; Byrne, M.C.; Follettie, M.T.; Gallo, M.V.; Chee, M.S.; Mittmann, M.; Wang, C.; Kobayashi, M.; Hortin, H.; Brown, E.L. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. **1996**, *14*, 1675–1680.
3. Lipschutz, R.J.; Fodor, S.; Gingeras, T.; Lockhart, D. High-density synthetic oligonucleotide arrays. Nat. Genet. **1999**, *21*, 20–24.
4. Gibson, G.; Muse, S.V. *A Primer of Genome Science*; Sinauer Associates: Sunderland, MA, 2002.
5. Yang, Y.W.; Speed, T.P. Design issues for cDNA microarray experiments. Nat. Rev., Genet. **2002**, *3*, 579–583.
6. Chen, Y.-J.; Kodell, R.L.; Sistare, F.; Thompson, K.; Morris, S.; Chen, J.J. Normalization methods for cDNA microarray data Analysis. J. Biopharm. Stat. **2003**, *13*, 54–57.
7. Kerr, M.K.; Churchill, G.A. Experimental design for gene expression microarrays. Biostatistics **2001**, *2*, 183–201.
8. Lee, M.L.; Kuo, F.C.; Whitemore, G.A.; Sklar, J. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cNDA hybridization. Proc. Natl. Acad. Sci. U. S. A. **2000**, *97*, 9834–9839.
9. Black, M.A.; Doerge, R.W. Calculation of the Minimum Number of Replicate Spots Required for Detection of Significant Gene Expression Fold Changes for cDNA Microarrays. In *Technical Report*; Department of Statistics, Purdue University: Indiana, 2002.
10. Pan, W.; Lin, J.; Le, C.T. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach research. Genome Biol. **2002**, *3*, 0022.1–0022.10.
11. Desu, M.M.; Raghavarao, D. *Sample Size Methodology*; Academic Press: New York, 1990.
12. Chen, Y.; Dougherty, E.R.; Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Opt. **1997**, *2*, 364–374.
13. Schuchhardt, S.; Beule, D.; Malik, A.; Wolski, E.; Eickhoff, H.; Lehrach, H.; Herzel, H. Normalization strategies for cDNA microarrays. Nucleic Acids Res. **2000**, *28* (10), e47.
14. Kerr, M.K.; Martin, M.; Churchill, G.A. Analysis of

variance for gene expression microarray data. J. Comp. Biol. **2000**, *7* (6), 819–838.
15. Kerr, M.K.; Afshari, C.A.; Bennett, L.; Bushel, P.; Martinez, J.; Walker, N.J.; Churchill, G.A. Statistical analysis of a gene expression microarray experiment with replication. Stat. Sin. **2002**, *12*, 203–217.
16. Goryachev, A.B.; Macgregor, P.F.; Edwards, A.M. Unfolding of microarray data. J. Comput. Biol. **2001**, *8*, 443–461.
17. Wolfinger, R.D.; Gibson, G.; Wolfinger, E.D. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **2001**, *8*, 625–637.
18. Irizarry, R.A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y.D.; Antonellis, K.J.; Scherf, U.; Speed, T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **2002**, *in press.*
19. Yang, Y.W.; Dudoit, S.; Luu, P.; Peng, V.; Ngai, J.; Speed, T.P. Normalization of cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. **2002**, *30*, e15.
20. Tsai, C.; Hsueh, H.; Chen, J.J. A Generalized Additive Model for Microarray Gene Expression Data Analysis. In *Technical Report E7221-3*; National Center for Toxicological Research: Jefferson, AR, 2002.
21. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. **1979**, *74*, 829–836.
22. Newton, M.A.; Kendziorski, C.M.; Richmond, C.S.; Blattner, F.R.; Tsui, K.K. On differential variability of expression ratio: Improving statistical inference about gene expression change from microarray data. J. Comput. Biol. **2001**, *8*, 37–52.
23. Allison, D.B.; Gadbury, G.L.; Heo, M.; Fernandez, J.R.; Lee, C.K.; Prolla, T.A.; Weindruch, R. A mixture model approach for the analysis of microarray gene expression data. Comput. Stat. Data Anal. **2002**, *39*, 1–20.
24. Draghici, S.; Kuklin, A.; Hoff, B.; Shams, S. Experimental design, analysis of variance and slide quality assessment in gene expression arrays. Curr. Opin. Drug Discov. Dev. **2001**, *4*, 332–337.
25. Herwig, R.; Aanstad, P.; Clark, M.; Lehrach, H. Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. Nucleic Acids Res. **2001**, *29*, e117.
26. Dudoit, S.; Yang, Y.H.; Callow, M.J.; Speed, T.P. Statistical methods for identifying differential expressed genes in replicated cDNA microarray experiments. Stat. Sin. **2002**, *12*, 111–139.
27. Ideker, T.; Thorsson, V.; Siegel, A.F.; Hood, L.E. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J. Comput. Biol. **2000**, *7*, 805–817.
28. Tsai, C.; Chen, Y.J.; Chen, J.J. Testing for Differentially Expressed Genes with Microarray Data. In *Technical Report E7221-2*; National Center for Toxicological Research: Jefferson, AR, 2002.
29. Tusher, V.G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. **2001**, *98*, 5116–5121.
30. Efron, B.; Tibshirani, R.; Goss, V.; Chu, G. *Microarrays an Their Use in a Comparative Experiment*; Stanford University: California, 2000. Preprint 37B/213.
31. Hochberg, Y.; Tamhane, A.C. *Multiple Comparison Procedures*; John Wiley & Sons: New York, 1987.
32. Westfall, P.H.; Young, S.S. *Resampling-Based Multiple Testing*; John Wiley & Sons: New York, 1993.
33. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc., B **1995**, *57*, 289–300.
34. Storey, J.D. A direct approach to false discovery rate. J. R. Stat. Soc., B **2002**, *64*, 479–498.
35. Holm, S. A simple sequentially rejective multiple test procedure. Scand. J. Statist. **1979**, *6*, 65–70.
36. Wen, X.; Fuhrman, S.; Michaels, G.S.; Carr, D.B.; Simth, S.; Barker, J.L.; Somogyi, R. Large-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. **1998**, *95*, 334–399.
37. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. **1998**, *95*, 14863–14868.
38. Chen, C.H. Generalized association plots: Information visualization via iteratively generated correlation matrices. Stat. Sin. **2002**, *12*, 7–29.
39. Hartigan, J.H. Representation of similarity matrices by trees. J. Am. Stat. Assoc. **1967**, *62*, 1140–1158.
40. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, B.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. **1999**, *96*, 6745–6750.
41. Sokal, R.R.; Sneath, P.H. *Principal of Numerical Taxonomy*; Freeman: San Francisco, 1963.
42. Tavazoie, S.; Hughes, J.D.; Campbell, M.J.; Cho, R.J.; Church, G.M. Systematic determination of genetic network architecture. Nat. Genet. **1999**, *22*, 281–285.
43. Kohonen, T. *Self-Organizing Maps*; Springer: Berlin, 1995.
44. Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E.S.; Golub, T.R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. **1999**, *96*, 2907–2912.
45. Panel on Discriminant Analysis and Clustering. Discriminant analysis and clustering. Stat. Sci. **1989**, *4*, 34–69.
46. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gassenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; Bloomfield, D.D.; Lander, E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **1999**, *286*, 531–537.
47. Brown, M.P.S.; Grundy, W.N.; Lin, D.; Cristianini, N.; Sugnet, C.W.; Furey, T.S.; Ares, M., Jr.; Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. **2000**, *97*, 262–267.
48. Vapnik, V. *Statistical Learning Theory*; Wiley: New York, 1998.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

49. Scholkopf, C.; Burges, J.C.; Smola, A.J. *Advances in Kernel Methods*; MIT Press: Cambridge, MA, 1999.
50. Fisher, R.A. The use of multiple measurements in taxonomic problems. Ann. Eugen. **1936**, *7*, 179–188.
51. Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press: New York, 1990.
52. Quinlan, J.R. Induction of decision trees. Mach. Learn. **1986**, *1*, 81–106.
53. Yeang, C.H.; Ramaswamy, S.; Tamayo, P.; Mukherjee, S.; Rifkin, R.M.; Angelo, M.; Reich, M.; Lander, E.; Mesirov, J.; Golub, Y. Molecular classification of multiple tumor types. Bioinformatics **2001**, *17* (Supplement 1), S316–S322.
54. Alter, O.; Brown, P.O.; Bostein, D. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. **2000**, *97*, 10101–10106.
55. Holter, N.S.; Mitra, M.; Maritan, A.; Cieplak, M.; Banavar, J.R.; Fedoroff, N.V. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. Proc. Natl. Acad. Sci. U. S. A. **2000**, *97*, 8409–8414.
56. Borg, I.; Groenen, P. *Modern Multidimensional Scaling: Theory and Applications*; Springer-Verlag: New York, 1997.
57. Bittner, M.; Meltzer, P.; Chen, Y.; Jiang, Y.; Seftor, E.; Hendrix, M.; Radmacher, M.; Simon, R.; Yakhinik, Z.; Ben-Dor, A.; Sampas, N.; Dougherty, E.; Wang, E.; Marincola, F.; Gooden, C.; Lueders, J.; Glatfelter, A.; Pollock, P.; Carpten, J.; Gillanders, E.; Leja, D.; Dietrich, K.; Beaudry, C.; Berens, M.; Alberts, D.; Sondak, V.; Hayward, N.; Trent, J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature **2001**, *406*, 536–540.
58. Hastie, T.; Tibshirani, R.; Eisen, M.B.; Alizadeh, A.; Levy, R.; Staudt, L.; Chan, W.C.; Botstein, D.; Brown, P. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. **2000**, *2*, 0003.1–0003.21.
59. Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **1964**, *29*, 1–27.
60. DeRisi, J.L.; Iyer, V.R.; Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **1997**, *278*, 680–686.
61. Marcotte, E.M.; Pellegrini, M.; Thompson, M.J.; Yeates, T.O.; Eisenberg, D.A. Combined algorithm for genome-wide prediction of protein function. Nature **1999**, *402*, 8386.
62. D'haeseleer, P.; Liang, S.; Somogyi, R. Genetic network inference: From co-expression clustering to reverse engineering. Bioinformatics **2000**, *16*, 707–726.
63. Jenssen, T.K.; Laegreid, A.; Komorowski, J.; Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. **2001**, *28*, 21–28.
64. Zin, W.; Riley, R.M.; Wolfinger, R.D.; White, K.P.; Passador-Gurgel, G.; Gibson, G. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **2001**, *29*, 389–395.
65. Kendziorski, C.M.; Lan, H.; Attie, A.D. The Efficiency of Pooling mRNA in Microarray Experiments. In *Technical Report #168*; Department of Biostatistics, University of Wisconsin: Madison, WI, 2002.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016