3

Introduction to Microarray Experimentation and Analysis

Peter Gieser, Gregory C. Bloom, and Emmanuel N. Lazaridis

1. Introduction

Microarray experiments try to measure simultaneously the quantity of many specific messenger RNA (mRNA) sequences contained in a sample. These quantities are called *gene expression*. The sample mRNA can be extracted from human tissue, plant material, or even yeast. Because thousands of these sequences can be measured in a single experiment, scientists have a large window into the workings of a biological system. This is in contrast to use of more traditional approaches such as Northern blots, which limit research to one-gene-at-a-time experiments.

There are many ways that microarrays can be used to further scientific research. One application is in the area of human cancer where, for example, we seek to identify colon cancer patients who are at risk for metastasis. While surgical extirpation of colorectal cancer remains the primary modality for cure, patients who have metastasized to distant sites at the time of surgical intervention frequently die from their disease. Unfortunately, there is no accurate means of identifying the patients who are at risk for metastasis using current staging systems, which are based only on clinicopathologic factors. Moreover, attempts at improving these staging systems, using molecular techniques to assay the expression of single or a small number of genes, have been relatively unsuccessful. This is likely because the process of metastasis is complex and linked to the expression of numerous gene families and biological pathways. Because microarray technology provides a more comprehensive picture of gene expression, experiments involving colon cancer and metastatic tumor specimens can be used to derive a molecular fingerprint in primary tumors portending metastasis.

From: *Methods in Molecular Biology, vol. 184: Biostatistical Methods*
Edited by: S. W. Looney © Humana Press Inc., Totowa, NJ

2. Robotic Spotting vs Photolithographic Technology

The term *microarray instance* refers to a single image of a particular hybridized microarray chip, slide, or filter. Currently, there are two major technologies for generating microarray instances. The older technology uses a robotic arm to first place specific cDNA probes, representing known genes or expressed sequence tags (ESTs)[1] of interest, onto a substrate. An RNA sample is then reverse transcribed and tagged with a fluorescent dye. This mixture is washed over the substrate, where the tagged cDNA hybridizes to the complementary cDNA probe. When scanned by a laser of the appropriate wavelength, the amount of fluorescence (as seen by a confocal microscope) is a measure of the quantity of tagged cDNA that has adhered to each probe. This, in turn, is used to directly infer the amount of a particular gene present in the original sample. In addition, a second RNA sample, tagged with a different fluorescent dye, can be mixed with the first sample. By scanning at two different wavelengths, information from each sample can be generated using only one slide. A newer, proprietary, technology by Affymetrix has emerged that also generates microarray instances. The Affymetrix GeneChip system works by creating a defined array of specific oligonucleotides, or *oligos*, on a solid support via appropriate sequencing of masks and chemicals in a photolithographic process, not unlike the way in which semiconductors are manufactured. A Biotin-tagged cRNA sample is washed over the chip and hybridized to the complementary oligo probes. A laser scans the chip and the fluorescence is measured. In contrast to the spotting technology, a mathematical model is required to combine the information from multiple oligos into a single gene expression level.

These two methods are similar in that they both contain probes in an array on a solid surface and are exposed to a sample for hybridization. Both are scanned and result in an image representation of the data.

The differences in the methods are key. One difference is the manner in which the expression level for a gene is established. The Affymetrix system uses a designed set (typically 40) of 25-mer oligos per gene,[2] which must be combined to quantify gene expression. This is in contrast to the spotting technology, where each gene is typically represented by only one target sequence. Another difference is the ability of the spotted array to generate two or more microarray images from the same slide by scanning it at different frequencies, corresponding to the fluorescent labels employed.

---

[1] Successive references to genes or gene expression implicitly include ESTs.

[2] The 40 features in a gene set are typically composed of 20 perfect match–mismatch oligo pairs. Each perfect match oligo is a true cDNA probe for its associated gene product. Each mismatch feature contains a single nucleic acid substitution in the center of the strand relative to the perfect match. Affymetrix includes mismatch features on chips to provide a means for quality control prior to and during quantitation.

There are statistical implications that need to be considered in working with the different technologies. As spotted arrays have the ability to utilize multiple samples per slide, they are more flexible than the Affymetrix arrays in accommodating different types of experimental designs. However, spotted arrays can require human intervention that is hard to account for in statistical models. Supervision in the form of discarding malformed spots is common, but what are the implications? What methods are used to find the spots, and what impact does the choice of a particular method have on the final analysis? As argued in Chapter 2 by Bloom, Gieser, and Lazaridis, imaging choices made within a reasonable envelope can have substantive effects on analytic results. The Affymetrix arrays, although generally more consistent and well-defined on the substrate, use only one sample chip, and that limits the choice of experimental design. The additional complexity of trying to put together the information from the individual oligos also provides a statistical challenge that is not present with the spotted arrays.

It is an important fact that microarray technologies continue to develop, resulting in additional complexities for the analyst. For example, it has been suggested that 60-basepair oligo sequences may improve sensitivity and specificity for gene expression, relative to the 20-basepair oligos currently used in Affymetrix chips. It is also of note that fewer oligo probes for each gene set may be assembled in future versions of Affymetrix chips. Data from spotted arrays are being impacted by developments in substrate technology seeking to improve the imaging properties of glass slides. Other ongoing technological developments include the use of ink-jet spotters as a means to place probes on a substrate, and increases in the acceptable density of spots or features on a microarray instance. Changes in technology imply that experiments performed at one time may need to be treated differently from those performed at a later date. Because one goal of many microarray projects is to compile gene expression information over multiple years, there is a need for analytic models that can handle the complexities resulting from further technological developments.

3. Imaging Analysis of Microarray Images

As mentioned previously, we use the term *microarray instance* to refer to a single image of a particular hybridized microarray chip, slide, or filter. Such terminology heeds the fact that the first (electronic) capture of information from a physical experiment is in the form of an image, which is typically obtained using a laser scanning device. Because so many imaging issues can impact what quantities are derived from a microarray image, we advocate treating microarray images as "the data." Although Chapter 2 suggests a new paradigm for microarray data analysis based on this thinking, the current approach at most institutions derives only one set of data from any single microarray instance. Thus, we limit our discussion of imaging analysis of microarray images to a few, brief remarks.

The primary elements of imaging analysis that can affect quantitation are choice of background adjustment and spot characterization methods, along with choice of their associated parameters. Typical background adjustment algorithms may account for global (image-wide) and/or local (in the vicinity of a spot or feature) background phenomena. What regions of an image are chosen to calculate the parameters of these algorithms may vary substantially by technology and analyst. For example, Affymetrix technology packs oligonucleotide probes so densely on the surface of a chip that one must employ minimally hybridized probe regions to calculate average background intensity. Global background values can be calculated directly using this scheme; calculation of local background adjustments requires application of a spatial model that uses a method such as kriging. Because of the manufacturing and hybridization processes employed by oligonucleotide chip technology, local background adjustments may be relatively less important than in the context of spotted arrays, in which the application of a coating to a glass slide as well as other technological issues can introduce substantial amounts of local noise. Chapter 2 presents an example wherein use of two, virtually indistinguishable, threshold values for local background lead to substantially different inference.

Spots may also be characterized differently in different applications. Three common approaches involve quantitating image intensity in rectangular regions, in ellipsoids, or even in regions determined by an edge detection scheme (which will also depend on choice of background threshold). In addition, spotted microarrays may exhibit doughnut spotting, that is, bright spots with a dark hole in their centers resulting from the process by which probe material was placed on the substrate by a robotic needle. An algorithm to identify and discount these dark regions may be appropriate in certain cases.

4. Statistical Analysis of Microarray Data

4.1. General Overview

Having illustrated that quantities derived from microarray instances can be affected by imaging choices, we proceed to discuss the statistical issues that are the focus of this chapter, restricting our presentation to cases wherein only one set of data is derived from any single microarray instance. In such a set, each arrayed gene is associated with one estimate of its (relative) expression. We employ the generic term *objects* to mean a set of genes, microarray instances, drugs, or so on, to which an analytic method for clustering or classification can be applied.

Generally, *microarray data* consist of observations on n-tuples of objects. A common form for these data gives one observation of estimated gene expression for each combination of elements of a set of genes and a set of microarray instances.

An important analysis characteristic is the degree to which external information is employed to assist a method in determining appropriate object clusters or classes. Methods that rely on external information or user interaction are called *supervised*. Methods that refer only to the data at hand are called *unsupervised*. In a similar manner, some methods may be trained using external data (and in that sense are supervised), but may still be applied to new data in an unsupervised manner.

All methods assume a certain degree of structure in the data to be analyzed. Whether the underlying model is explicitly recognized or not, some methods are structurally heavy, leading to a substantial influence on clustering or classification, while structurally light methods tend to have less influence. Clearly, a good model will reflect the structure of the data, and the best models will represent the underlying biology.

4.2. Data Adjustment

Because the data represented by a single microarray instance are a reflection of the relative amounts of gene expression in a tested sample, an important question is how to standardize these quantities to allow for comparison across multiple instances. It is unfortunate that some authors call this process *normalization*, as that term additionally suggests the transformation of data to satisfy Gaussian distributional properties. We discuss data transformation for modeling purposes toward the end of this chapter. Several methods for standardizing across microarray instances are available, but each has its disadvantages.

The most basic technique involves standardizing data from each chip according to the average intensity of the pixels across the whole scan of the chip, or across all the spots. This approach has the advantage of simplicity, and requires no special experimental considerations, but is fraught with danger. For example, the goal of many experiments involves depressing or stimulating transcription of a large number of gene products, so that different overall average intensity between images may not be an imaging artifact.

Another basic technique is to standardize data from each chip according to the average intensities of internal controls, or "housekeeping" genes, that are not expected to change across a particular experiment (such as cellular gene products in a viral DNA chip). This method avoids the major problem of the previous approach, but still relies on the assumption that the internal controls are unaffected by differences in samples and experimental conditions across microarray instances. If a large quantity of internal control spots is assembled on a microarray, then the approach of Amaratunga and Cabrera ***(1)*** may be considered. These authors employed the intensity histograms of pixels associated with the internal controls to estimate a transformation of each microarray instance to a standard intensity curve. Alternatively, one can try dividing the

average intensity measurement of each spot by the average intensity over a small set of controls. Most published analyses to date employ this approach, but this can be dangerous because the intensity transformation between two images (and especially between two fluorescent channels) is frequently nonlinear.

A third technique standardizes data according to the fluorescence of one or more known elements added to the experimental sample just prior to hybridization. Required is the assumption that the addition of a "spike" to the mixture does not change the sample's other properties. For example, the Affymetrix procedure is to spike human samples with herring sperm DNA, relying on the supposition that herring sperm and human DNAs are sufficiently lacking in homology. One major problem is the likelihood that the image intensity of a spike will demonstrate substantial variability. To account for this, the Affymetrix protocol recommends spiking the sample with a set of staggered concentrations of control cRNAs. These controls are used to determine the sensitivity of the chip by noting the smallest concentration level that can be detected. We note that they are not used in the Affymetrix analytic procedure except as a global filter for adequacy of the microarray instance.

Finally, any of a variety of statistical regression models can be used to standardize data by looking at pools of experimental samples in an associated experimental design. As an illustration of how one such approach might work, suppose there are two biological samples to be analyzed using a one-channel scan microarray system. Instead of running each sample on each chip separately, suppose the first sample is run on the first chip, and a mixture of the two samples is run on the second chip. Denoting the expression vector of each sample by X_i and the expression vector from each microarray instance by Y_j, one might (somewhat naively) expect that the above experiment would satisfy the relationships $X_1 = Y_1$, and $X_2 = 2(Y_2) - X_1$. Unfortunately, this may not always be the case, possibly owing to effects of RNA concentration and complexity in the mixture. Experiments suggest a trend correlated with spot average intensity, whereby mixtures of samples on a single chip tend to underestimate the average of samples run on different chips. However, because this relationship may be predictable, it is possible that a regression model might be used to adjust for possible bias.

4.3. Combining Oligonucleotide Information in a Probe Set

A special requirement of oligonucleotide chips is a robust method to combine the measured average intensities of oligo features in a probe set into a single value estimating expression of the associated gene product. Typically, each probe set is composed of 40 features, arranged in 20 perfect match–mismatch oligo pairs. Each perfect match oligo is a piece of cDNA probe for the

associated gene product. Each mismatch feature contains a single nucleic acid substitution in the center of the strand relative to the perfect match. Affymetrix includes mismatch features on chips to provide a means for quality control prior to and during quantitation. We note that the existence of perfect match and mismatch oligos in each probe set on an Affymetrix chip is not an important component of this problem because there is substantial evidence to suggest that molecules with high affinity to a mismatch oligo may have low affinity to the corresponding perfect match. Thus, we advise against the use of algorithms that combine perfect match and mismatch information to create summary statistics intended to estimate gene expression, such as pairwise differences in average feature intensity.

The current software provided by Affymetrix to investigators returns the average of feature intensities for a subset of features in each probe set. The subset is chosen in each array by considering the mean and standard deviation of differences between paired perfect match and mismatch average intensities, after excluding the maximum and the minimum. Probes whose probe pair differences deviate by more than three standard deviations from the mean are excluded in gene expression estimates. If two microarray instances are to be compared, the intersection of acceptable probes in each instance is employed to evaluate gene expression difference. Not only does this procedure potentially exclude informative probes with large responses in individual arrays, it also implicitly assumes that the information provided by each acceptable oligo is of equal importance. Clearly superior to this approach would be a weighted sum of oligo-specific values, with parameters chosen to reflect the extent of information in each oligo.

At least two procedures have been employed to choose such weights. Li and Wong ***(2)*** assume that the average intensity of each probe in a probe set increases linearly with respect to increases in underlying, unknown gene expression, but with probe-specific sensitivity. This assumption leads to a weighted sum conditional least squares estimate of gene expression. In what follows we ignore their use of mismatch feature intensities. Letting i equal index array instances, j index probes, and n index genes, their basic model is $y_{ij} = \theta_i\phi_j + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim N(0, \sigma^2)$, and $\sum_j \varphi_j^2 = J$, J being the number of probes in a given probe set. Least squares estimation may be performed by iteratively calculating the gene-specific parameters (θ_i) and the probe-specific weights (ϕ_j), identifying and excluding during the iterative procedure any outlier microarray instances and probes (outliers relative to the model) as well as probes with high leverage (which may be untrustworthy because of their influence on the model estimates). Possible drawbacks to this approach arise from reliance on the parametric model, on its distributional assumptions, and on the criteria one employs to exclude outlying or untrustworthy data.

In our setting we employ a nonparametric approach to weighing oligo features using a minimum risk criterion, an approach that is easily described and implemented. We think of oligos in a probe set as players on a team, who have been selected on evidence that as individuals they are top performers. The performance of any given player can be gauged according to how well that player estimates a value of interest, in this case, the expression of the gene associated with the probe set. Each microarray instance in a particular analysis corresponds to a single game, resulting in a score for each of the members of the team. Indexing players by i and games by j, denote the average intensity of each oligo at each microarray instance by y_{ij}. Our main question involves a coaching decision, whereby the analyst (coach) seeks to obtain a better estimate of (unobserved) gene expression, θ, for a particular probe set. Using the least squares criterion, it can be shown that a coach should use $\bar{y}_{gj}$ to estimate θ in the situation where all the players perform equally well across games. In the microarray context, relative performance of players must be estimated by the coach, who does not know the value of θ that would be needed to calculate an exact loss function. Instead, we argue that within each game, the rational coach would evaluate each player against a best estimate of gene expression derived from the rest of the team. Thus, the problem of minimizing loss over players and games reduces to calculating a set of parameters, ϕ_i, such that

$$\sum_i \sum_j \left(\phi_i y_{ij} - \frac{\sum_{k:k \neq i} \phi_k y_{kj}}{\phi_g - \phi_i} \right)^2$$

is minimized. The fact that this procedure is equivalent to minimizing the leave-one-out cross-validation estimate of variance for the mean of coach-adjusted player estimates, $\phi_i y_{ij}$, suggests the situations in which this approach may perform substantially better than a parametric one, including settings in which limited information is available about probe weights. Designed laboratory experiments, in particular, typically result in few microarray instances over a multiplicity of conditions. In addition, situations in which assumptions of a linear model with constant variance may be violated will often arise in laboratory experiments because of the hybridization performance of different molecular mixtures across samples being compared.

4.4. Differential Gene Expression

After proper standardization, a natural question to ask is which genes are differentially expressed across two or more samples. For concreteness, consider a spotted microarray in which two samples have been cohybridized.

Traditionally, the raw ratio (between the two samples) of standardized spot average intensities has been used to make inferences about which genes are significantly differentially expressed. Newton et al. ***(3)*** point out that this is problematic because a given fold change may have a different interpretation for a gene whose absolute expression is low in both samples as compared to a gene whose absolute expression is high in both samples. They suggest that there is room for improvement in the initial signal processing that may have bearing on downstream tasks such as clustering.

The solution Newton describes is based on hierarchical models of measured expression levels that account for two sources of variation. One source is measurement error, the fluctuation of the spot intensity around some mean value that is a property of the cell type, the particular gene, and other factors. The second source is gene variation, the fluctuation of the mean intensity value between the different genes. This formulation allows the computation of probabilistic statements about actual differential expression. The key findings are that observed ratios are not optimal estimators, focusing on fold changes alone is insufficient, and confidence statements about differential expression depend on transcript abundance.

The specific sampling model used by Newton is based on the Gamma distribution. Given genes are modeled as independent samples from distinct Gamma distributions with common coefficient of variation (i.e., constant shape parameter). Specifically, if R and G are the measured expression levels for a gene across the two samples, let $R \sim \text{Gamma}(a,\theta_R)$ and $G \sim \text{Gamma}(a,\theta_G)$. Then the scale parameters are assumed to follow a common $\text{Gamma}(a_0,\upsilon)$ distribution. This model is stated to be reasonably flexible and skewed right, while exhibiting increasing variation with increasing mean. It turns out that given these model components, the Bayes estimate of differential expression is $p_B = (R + \upsilon)/(G + \upsilon)$, which has the classic form of a shrinkage estimator. The implication of this is that for strong signals p_B will be close to the naïve estimator R/G, but there is attenuation of p_B when the overall signal intensity is low. Clearly, p_B naturally accounts for decreased variation in differential expression with increasing signal on the log scale. One problem with this method, however, is that spots that cannot be distinguished from the background in either channel are omitted from analysis. It could be argued that these are in fact the most important cases of all! Another problem is that the restrictive parametric model may not fit the distribution of actual gene expressions on a given chip.

To determine significant differential expression, an additional layer is added to the model in the form of a latent variable z that indicates whether or not true differential expression exists. The Expectation-Maximization (EM) algorithm

is then used to estimate the parameters and compute the posterior odds of change at each spot.

An alternative method for exploring differential expression is via *robust* analysis of variance (ANOVA). This has been described by Amaratunga and Cabrera ***(1)***.

4.5. Principal Components

Although principal components analysis (PCA) is not a model-based method, it still plays an important role in facilitating model-based analyses. PCA is a technique commonly used for dimension reduction. Generally, PCA seeks to represent n correlated random variables by a reduced set of d ($d < n$) uncorrelated variables, which are obtained by transformation of the original set onto an appropriate subspace. The uncorrelated variables are chosen to be good linear combinations of the original variables, in terms of explaining maximal variance, orthogonal directions in the data. Data modeling and pattern recognition are often better able to work on the reduced form, which is also more efficient for storage and transmission. In particular, pairs of principal components are often plotted together to assist in visualizing the structure of high-dimensional data sets, as in the biplot.

Suppose we have a set of microarray data in standard form, X, a matrix with as many rows as there are genes (M responses) and as many columns as there are microarray instances (n observations). Standard PCA seeks the eigenvectors and associated eigenvalues of the covariance matrix for these data. Specifically, if Σ is the covariance matrix associated with the random vector $\vec{Z} = (Z_1, Z_2, ..., Z_n)$ and Σ has eigenvalue–eigenvector pairs $(\lambda_1, \vec{e}_1), (\lambda_2, \vec{e}_2), \ldots, (\lambda_n, \vec{e}_n)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, then the ith principal component is given by $Y_i = e_{1i}Z_1 + e_{2i}Z_2 + \cdots + e_{ni}Z_n$. Note that the principal components are uncorrelated and have variances equal to the eigenvalues of Σ. Furthermore, the total population variance is equal to the sum of the eigenvalues, so that $\lambda_k/(\lambda_1 + \cdots + \lambda_n)$ is the proportion of total population variance due to the kth principal component. Hence if most of the total population variance can be attributed to the first one, two, or three components, then these components can "replace" the original n variables without much loss of information. Each component of the coefficient vector also contains information. The magnitude of e_{ki} measures the importance of the k^{th} variable to the i^{th} principal component, irrespective of the other variables. In the context of microarrays, the lack of data in the instance direction deemphasizes the data-reduction aspect of principal components. Instead, the interest is generally in the interpretation of the components.

The sample principal components are calculated as described earlier, replacing the (generally unknown) population covariance matrix with the sample covariance matrix. Hilsenbeck et al. ***(4)*** applied this technique to three microarray instances generated from human breast cell tumors. These instances

correspond to estrogen-stimulated, tamoxifen-sensitive, and tamoxifen-resistant growth periods. Their results yielded three principal components interpreted as (1) the average level of gene expression, (2) the difference between estrogen-stimulated gene expression and the average of tamoxifen-sensitive and tamoxifen-resistant gene expression, and (3) the difference between tamoxifen-sensitive and tamoxifen-resistant gene expression.

Another use of principal components is as a basis for clustering. The correlation of each gene with the leading principal component provides a way of sorting (or clustering) the genes. Raychaudhuri et al. ***(5)*** analyzed yeast sporulation data, which measured gene expression at seven time points (***6***). They determined that much of the observed variability can be summarized in just two components: (1) overall induction level and (2) change in induction level over time. Then they calculated the clusters according to the first principal component and compared them to the clusters reported in the original paper.

It is also possible to use PCA to reduce the dimensionality of the analytic problem with respect to the gene space. A number of ways have been proposed to do this.

One use of this idea is in *gene shaving* ***(7)***. This method seeks a set of approximate principal components that are defined to be *supergenes*. The genes having lowest correlation with the first supergene are shaved (removed) from the data and the remaining supergenes are recomputed. Gene blocks are shaved until a certain cost–benefit ratio is achieved. This process defines a sequence of blocks with genes that are similar to one another. A major problem with this approach is the shifting definition of supergenes over the course of the analysis. Although gene shaving incorporates the ideas of principal components, it is important to recognize that the shaving algorithm itself is *ad hoc*.

Another application of the gene space reduction idea solves the problem of using gene expression values as predictors in a regression setting. Because correlated predictors are known to cause difficulties, principal components regression (PCR) uses the gene principal components as predictor surrogates. A second method that uses this idea is partial least squares regression (PLSR). PLSR is employed to extract *only* the components (sometimes called *factors*) that are directly relevant to both the predictors and the response. These are chosen in decreasing order of relevance to the prediction problem.

Both PCR and PLSR produce factor scores as linear combinations of the original predictor variables, so that there is no correlation among the factor score variables used in the predictive regression model. For example, suppose we have a data set with response variable Y and a large number of highly correlated gene expression predictor variables X. A regression using factor extraction for these types of data computes the factor score matrix $T = XW$ for an appropriate weight matrix W, and then considers the linear regression model $Y = TQ + \varepsilon$, where Q is a matrix of regression coefficients (loadings) for T, and

ε is the error term. Once the loadings, Q, are computed, the preceding regression model is equivalent to $Y = XB + \varepsilon$, where $B = WQ$, which can be used as a predictive regression model for gene expression data on the original scale.

PCR and PLSR differ in the methods used in extracting factor scores. In short, PCR produces the weight matrix W reflecting the covariance structure between the predictor variables, while PLSR produces the weight matrix W reflecting the covariance structure between the predictor and response variables.

Partial least squares regression produces a weight matrix W for X such that $T = XW$. Thus, the columns of W are weight vectors for the X columns producing the corresponding factor score matrix T. The weights are computed so that each of them maximizes the covariance between the response and the corresponding factor scores. Ordinary least squares procedures for the regression of Y on T may then be performed to produce Q, the loadings for Y. Thus, X is broken into two parts, $X = TP + F$, where the factor loading matrix P gives the factor model and F represents the unexplained remainder.

Whether centering and scaling of the data (normally done when determining principal components) makes sense for microarray data is an open question, and all the caveats that go along with PCA are still in effect ***(8)***.

A Bayesian application, using a related idea, models a binary response using a probit model and decomposes the linear predictor vector using the SVD ***(9)***. Specifically, if z_i is the binary variable reflecting status for each patient, then let $\Pr(z_i = 1 | \beta) = \Phi(x_i'\beta)$. Using the decomposition, obtain $X'\beta = (F'D)\gamma$, where $X = ADF$ from the SVD; A is the SVD loadings matrix; F is the SVD orthogonal factor score matrix (as before); D is the diagonal matrix of singular values; and $\gamma = A'\beta$ is the vector of parameters on the subspace formed using the new linear basis. A key part of this method is determining reasonable prior distributions for the vector of parameters on the subspace formed using the new linear basis. A reasonability criterion is defined in terms of the interpretability of the priors when back-transformed to the original space.

4.6. Latent Class Models

There are two major types of analyses that fall broadly into this category. A clustering method entails placing objects that are close together into clusters according to a specified metric. A classification method is where clustering is performed by estimating the probability of each object's membership in a latent (i.e., unobservable) class. All classification methods can be used to generate clusters, but clustering methods do not imply a particular class definition. Both clustering and classification methods can be used to discriminate among estimated differentiated sets of objects. However, because clustering methods do not explicitly model theoretical class constructs, they provide no basis for

determining misclassification, the association of an object with a class to which it does not belong. This is a major disadvantage of such analyses.

Analytic methods can be geared toward clustering or discriminating among various classes of either genes or microarray instances. These are one-way analyses. Methods can also be geared toward jointly clustering or discriminating among both gene and microarray instances. These are two-way analyses.

Generally, clustering methods use similarity measures to associate similar objects and to disassociate sets of similar objects from each other. Tibshirani et al. ***(10)*** review the various methods of clustering and show how they can be used to order both the genes and microarray instances from a set of microarray experiments. They discuss techniques such as hierarchical clustering, *K*-means clustering, and block clustering. Dudoit et al. ***(11)*** compare the performance of different discrimination methods for the classification of tumors based on gene expression data. These methods include nearest-neighbor classifiers, linear discriminant analysis, and classification trees.

Although these *ad hoc* analyses are still popular owing to the current dearth of more-sophisticated techniques and software, statisticians are feverishly working to come up with model-based approaches so that inference will be possible. It is these model-based approaches that are the focus of this chapter.

Classic multidimensional latent class models specify the form of the class conditional densities. A common specification associates each class with a multivariate normal distribution ***(11)***. In this case, the maximum likelihood discriminant rule is

$$C(x) = \text{argmin}_k \{(x - \mu_k)\Sigma_k^{-1}(x - \mu_k)' + \log |\Sigma_k|\}$$

Three special cases of interest are (1) when the class densities have the same covariance matrix, $\Sigma_k = \Sigma$, the discriminant rule is based on the square of the Mahalanobis distance and is linear:

$$C(x) = \text{argmin}_k (x - \mu_k)\Sigma_k^{-1}(x - \mu_k)'$$

(2) when the class densities have diagonal covariance matrices, $\Delta_k = \text{diag}(\sigma^2_{k1},...,\sigma^2_{kp})$, the discriminant rule is given by additive quadratic contributions from each variable:

$$C(x) = \text{argmin}_k \sum_{j=1}^{p} \left\{ \frac{x_j - \mu_{kj}}{\sigma^2_{kj}} - \log\left(\sigma^2_{kj}\right) \right\}$$

and (3) when the class densities have the same diagonal covariance matrix, $\Delta = \text{diag}(\sigma^2_1,...,\sigma^2_p)$, the discriminant rule is linear:

$$C(x) = \text{argmin}_k \sum_{j=1}^{p} \left\{ \frac{x_j - \mu_{kj}}{\sigma^2_{kj}} \right\}$$

For repeated measurement experiments, Skene and White ***(12)*** describe a flexible latent class model that also assumes normality. In the context of

microarray experiments, let Y_{mg} denote the log-transformed average spot intensity response of gene g on microarray slide m. Let L represent a discrete latent variable with levels 1,..., J. Assume that corresponding to each level of L is a profile of gene expression defined across multiple slides, $p_L = \{p_{1L},\ldots,p_{ML}\}$ which is defined in terms of deviation from average gene expression. Assume also that genes in the same biological pathway may have different expression intensities, d_g, depending on such factors as gene copy number, transcription efficiency, and so on. Thus, it makes sense to let $\mu_{mgj} = a + d_g + p_{mj}$ be a model for the mean response in microarray instance m for gene g in latent class j. Conditional on latent class membership, the error is assumed normal so that $Y_{mg} \mid L = j \sim N(\mu_{mgj},\sigma^2)$. The difficulty with this formulation is that estimation is a problem when the number of parameters is large, which is frequently the case with microarray data. Research is currently ongoing into techniques that might overcome this limitation.

Similar latent class model forms can be considered in a Bayesian context by placing Dirichlet priors on the class membership probabilities and appropriate conjugate priors elsewhere, then conducting Markov Chain Monte Carlo (MCMC) to generate samples from the full posterior to estimate class and gene parameters. The approximate EM solution is used as the starting point for the MCMC algorithm. An extension of this idea is to simultaneously divide the genes into classes with substantial internal correlation as well as allocate microarray instances to latent sample classes. Such models seek to identify gene classes with high sample class discriminatory power. Typically, we transform data to the log scale prior to application of the model.

For example, consider the following analysis of a time-course experiment on fibroblasts ***(13)***. In this experiment, cells were first serum-deprived and then stimulated, to investigate growth-related changes in RNA products over time. Other aliquots were additionally treated with cycloheximide. Samples of untreated and treated cells were collected at 12 and 4 time points, respectively, as were samples of unsynchronized cells. Microarrays included 8613 gene products, but analyses included only 517 of these. In published work, the authors used a hierarchical clustering algorithm to identify 10 patterns of gene expression in a subset of 517 genes, which was filtered before application of the clustering algorithm on the basis of the existence of "significant" univariate observed variability fold changes in gene expression over time.

Using a Bayesian latent class model of the above form, we analyzed the complete set of 8613 gene products, over all the time points and experimental conditions. We employed a normal error model on the log-transformed data with mean conditional on latent gene class and time (plus experimental condition), with gene-specific intensity modifiers to represent the degree to which each gene is a good marker of its associated latent gene class. Because genes

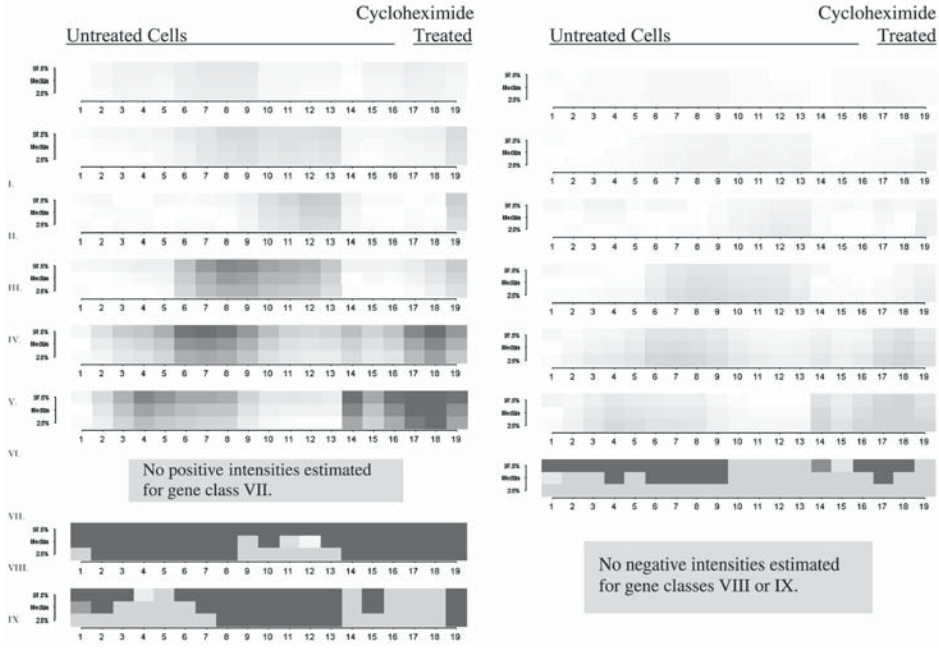

Fig. 1. Latent class patterns including median estimates and 95% confidence intervals from the analysis of serum-stimulated fibroblasts. This figure demonstrates both time-dependent increases and decreases in gene expression, as represented by positive (mostly red) and negative expression (mostly green) patterns.

could be estimated to have decreases in expression (negative intensity modification parameters) or increases in expression (positive intensity modification parameters), in **Fig. 1** we present both the estimated patterns and their inverses. In other words, red and green colors represent, respectively, higher or lower levels of expression relative to untreated cells at time 0. We use a white background to aid in visualization; brighter colors represent greater relative deviation from baseline.

This analysis also adjusted for data quality issues. Specifically, we treated an observation (one spot on one microarray instance) as missing whenever the interpixel correlation between the two scans was <0.6. Thus, our analysis accounts for and is unbiased by differential hybridization of samples. A consequence of this is that pattern VIII is separated from pattern VI because of insufficient information at samples 1 and samples 9 through 13, as evidenced by a wide confidence interval ranging from very bright green (2.5th percentile) to very bright red (97.5th percentile). Probabilities of gene membership in these latent patterns for every gene in the data are estimated by the model.

This analysis illustrates some additional advantages of the latent class approach. First, estimated time-course patterns are smoother than those originally proposed by hierarchical clustering analysis, even though no smoothness criterion is imposed by the model. Although no assumptions were made regarding the correlations within treatment group across time, the estimates show expression patterns that are reasonably correlated with known stages of cellular growth and mitosis, suggesting that this model is uncovering the underlying biology. Second, no prefiltering of genes is required by the technique; prefiltering is often necessary to apply clustering methods. Third, the model adjusts for data quality issues using Bayesian statistical approaches, to reduce the potential biases that can be introduced by experimental variability, especially at low spot intensities.

Standard statistical approaches can also be employed to diagnose lack-of-fit of this model to the data. For example, residual plots can be employed to identify deviations from the model fit. Residual distances between the observed and predicted values, conditioning on the Gaussian model, follow a chi-square (χ^2) distribution in analogous fashion to the residuals from standard multivariate analyses ***(14)***. In that context,

$$n\left(\overline{X} - \mu\right)^T \sum^{-1} \left(\overline{X} - \mu\right) \sim \chi^2 (p)$$

where $(\overline{X} - \mu)$ is the vector of differences between observed and expected mean values, Σ^{-1} is the generalized inverse of the covariance matrix, and $\chi^2 (p)$ refers to the χ^2 distribution on p degrees of freedom, p being functionally related to the dimensionality of the data. We calculated values of the form $\left(\overline{X} - \mu\right)^T \sum^{-1} \left(\overline{X} - \mu\right)$ and compared their quantiles to those of the appropriate χ^2 distribution in **Fig. 2**. Points falling on the straight line give evidence that the model fits the signal in the data, because the residuals seem to follow the appropriate χ^2 distribution. Points deviating to the right of the line signify overdispersion relative to the residual χ^2, meaning that there may be some statistically significant additional signal in those gene products that is not being described in the current model. Only 18 outliers (3.4%) from this model were identified. The most discrepant gene, *AA058863*, is a Soares retina N2b4HR *Homo sapiens* cDNA clone containing ALU sequences, which may account for observed expression pattern differences. Another set of gene products was identified through the residual analysis that could be associated with a still unidentified pathway. Another set of gene products, including WEE1-like protein kinase, were underexpressed around times 7 and 8 to a much greater degree than would be expected by the model. Whereas the hierarchical clustering analysis groups WEE1 with expressed genes such as *p57Kip2* and *p27Kip1*, the latent class analysis identified differences between the cyclin-dependent kinase inhibi-

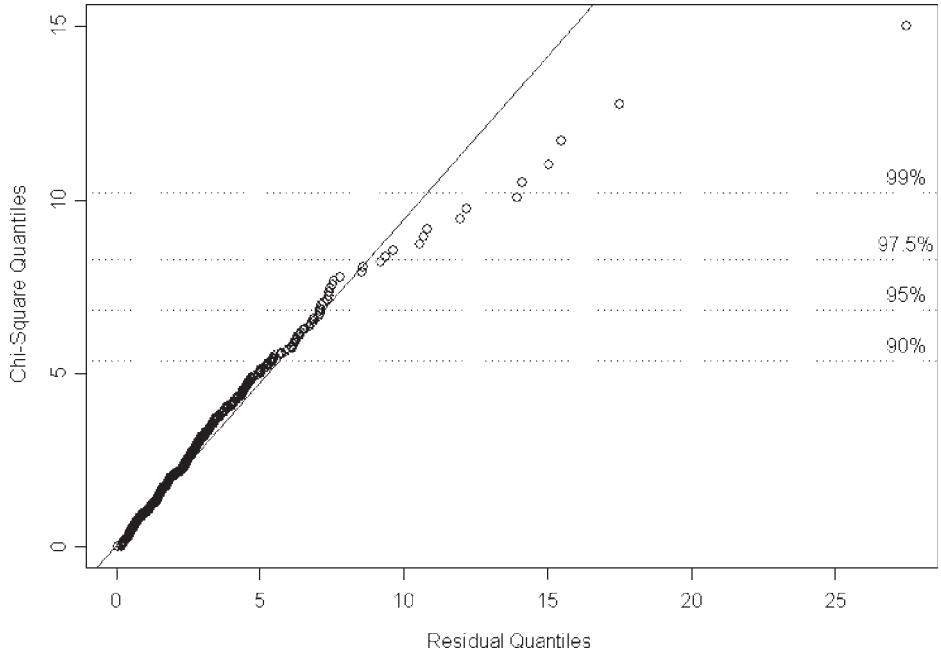

Fig. 2. QQ-plot of residuals from exploratory latent class model vs χ^2 distribution, showing outlying expressed genes deviating from the line towards upper right.

tors and the mitotic inhibitor WEE1, suggesting that WEE1 may have a slightly different function. Changes based on such observations can be fed back into the analytic model.

As previously discussed, a key distinction between formal latent class models and clustering algorithms is that the former provide an inferential framework while the latter do not. An inferential framework allows one to make probability statements concerning the uncertainty associated with identified classes and their members, as well as to estimate the number of classes needed to describe a particular data set. In the clustering arena, an approach for determining the validity of a particular clustering has recently been developed and is described in Bittner et al. ***(15)***. The method is based on evaluating cluster membership after introducing random perturbations to the data set. Hierarchical clustering is performed on the perturbed data set and compared to the original tree. Comparisons involve cutting the original and perturbed trees into k clusters followed by computing the proportion of paired samples clustering together in the original tree that do not cluster together in the perturbed tree. The average over multiple perturbed data sets (for a given k) yields the weighted average discrepant pairs statistic ($WADP_k$), which is then plotted vs k. Local minima on the $WADP_k$ curve indicate reproducible levels of structure.

4.7. Differential Equations

In this subheading, we discuss what are perhaps the most structural of microarray data modeling approaches. Various authors are seeking to model genetic networks using sets of nonlinear differential equations. Examples can be found in Reinitz and Sharp *(16)* and Wahde and Hertz *(17)*. Parametric forms are employed to model the rates of change of expression in certain genes. Unfortunately, lack of data restricts the estimates to first-order terms in the differential equations, so that only gene interaction weights can be estimated. Two assumptions in the Wahde and Hertz *(17)* model are of note. The first is in the choice of a particular structure for the combination of gene effects (the activation function; a genetic algorithm was employed for estimating the parameters), which assumes linear transcription. The second is the use of average trajectories of coarse clusters of genes with similar expression patterns as nodes. For these authors, this simplification is inspired by the fact that for reliable determination of the network parameters, a minimum requirement is that the number of useful data points exceeds the number of parameters. Gene expression data series often consist of only a few measurements, and clustering of genes into sets with similar temporal expression patterns substantially reduces the number of parameters requiring calculation.

Another example is in Chen et al. *(18)*. These authors also propose a differential equation model for gene expression, and provide a method to construct the model from temporal expression data. They make a number of assumptions, among which are a linear transcription function for each gene and feedback of the gene translation product on the transcription rates. They discern in their model transcription, translation, and degradation of RNA and proteins. Parameters can be estimated for these processes using Fourier transforms for stable systems, an approach that is specific for genes with periodic expression. These are important in cell cycle studies, for example, and all genes considered in the model are assumed to show this kind of expression pattern. As before, reduction of the problem, this time by employing periodicity and requiring stability in their system, is essential to solving this system with today's technology. The reduction in complexity resulting from these assumptions is substantial enough that many more features of gene expression could be parametrized and estimated than in the Wahde and Hertz approach.

D'Haeseleer et al. *(19)* provide a third example. These authors begin analysis by calculating cubic spline smoothers, fitting the curves of the gene expression time-series data. Of note is the fact that they use data from three different experiments that employed different time scales. Using a first-order interaction matrix as in Wahde and Hertz *(17)*, they estimate the interaction parameters using a least squares fit to the smooths. Limitations of this approach

include a lack of reducibility in the gene interaction structure and restriction to the primary linear components of the system.

Clearly, this kind of approach is promising, but must undergo substantial development before it can become widely applicable in data analysis. Significant research is needed on the interface of applied mathematics and statistics, so that known physical functioning of biological pathways can be appropriately reflected in otherwise data-driven statistical models.

4.8. Additional Issues

In this subheading we address three questions frequently raised regarding microarray data analysis. We consider these issues after describing the bulk of methods in the preceding subheadings, primarily because answers to them depend on what downstream analytic techniques one seeks to apply.

The first concerns whether microarray data should be transformed to a scale other than the one in which they were collected. The answer is simple: it depends on what analysis one seeks to perform. Although there are good biological reasons to consider transforming data to a $\log_{10}$ scale, we have found no situations in which distributional assumptions in a statistical model could be guaranteed through data transformation. In part this is a reflection of Kolmogorov's lament concerning the perpetual lack of fit one observes in employing simple forms to model large data sets. Our advice to end-user analysts is to pick methods that are relatively robust to a reasonable set of data transformations. We have had better luck with latent class models in this regard than with principal components and related dimension reduction techniques, which are widely known to be sensitive to choice of scale.

A second question concerns whether it matters which microarray instance is used as a reference in standardizing several for analysis. Again the answer depends on what downstream analysis one seeks to do. In our work, we prefer that standardization occur synchronously with the actual data analysis, by the same model that will answer a biological question of interest. A consequence of this preference is that we find ourselves restandardizing sets of data multiple times in multiple ways over the course of an analysis. Of course, methods that are more inferentially robust in the context of data transformations are also more robust to peculiarities of standardization.

Finally, we are often asked how one should visualize microarray data. The answer to this question has two parts. First, there is no biological question of interest that can be answered by looking at a microarray image, unless of course the array was specifically designed to do so. Usually, consideration of colorful array images is worthwhile only if one suspects quality control issues resulting from biological experimentation or imaging analysis. Second, there are at least

as many reasonable ways to visualize these data as there are inferential methods. Some methods (such as principal components and its relatives such as multidimensional scaling) are designed to reduce data to a small number of reasonably informative 2- or 3-D scatterplots. Other methods (clustering and latent class analysis) suggest graphical methods to visualize pattern sets. We also employ the full standard statistical repertoire of diagnostic plots in our work. In summary, there is nothing particularly special about microarray data that requires a different treatment from other large data problems with respect to visual presentation.

5. Conclusion

The vast amount of data generated by microarray technology tests statisticians' abilities to extract meaningful information from any given experimental context. The intense interest in and great potential of high-throughput, comprehensive molecular biology technologies is fueling a corresponding surge in statistical research on analytic methods for large, complex data sets. In addition to statisticians, molecular biologists, computer scientists, and imaging scientists all have roles to play in this development. Because of the multidisciplinary nature of this field, one of the greatest challenges to statisticians is in making sophisticated statistical methods accessible to their collaborators. This chapter has sought to give an overview of the most promising approaches for analyzing microarray data.

References

1. Amaratunga, D. and Cabrera, J. (2000) *Analysis of data from viral DNA microchips*. Technical Report, Rutgers University.
2. Li, C. and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98,** 31–36.
3. Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (1999) *On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.* Technical Report, University of Wisconsin.
4. Hilsenbeck, S., Friedrichs, W., Schiff, R., O'Connell, P., Hansen, R., Osborne, C., and Fuqua, S. (1999) Statistical analysis of array expression data as applied to the problem of Tamoxifen resistance. *JNCI* **91,** 453–459.
5. Raychaudhuri, S., Stuart, J., and Altman, R. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing* Jan. 4–9, 2000, Hawaii, pp. 452–463.
6. Chu, S., De Rivi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P., et al. (1998) The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705.

7. Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., et al. (2000) *Gene shaving: a new class of clustering methods for expression arrays*. Technical Report, Stanford University.
8. Hadi, A. and Ling, R. (1998) Some cautionary notes on the use of principal components regression. *Am. Statist.* **52,** 15–19.
9. West, M., Nevins, J., Marks, J., Spang, R., and Zuzan, H. (2000) *Bayesian regression analysis in the "Large p, Small n" paradigm with application in DNA microarray studies.* Technical Report, Duke University.
10. Tibshirani, R., Hastie, T. Eisen, M., Ross, D., Botstein, D., and Brown, P. (1999) *Clustering methods for the analysis of DNA microarray data*. Technical Report, Stanford University.
11. Dudoit, S., Fridlyand, J., and Speed, T. (2000) *Comparison of discrimination methods for the classification of tumors using gene expression data*. Technical Report, University of California, Berkeley.
12. Skene, A. and White, S. (1992) A latent class model for repeated measurements experiments. *Statist. Med.* **11,** 2111–2122.
13. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., et al.(1999) The transcriptional program in the response of human fibroblasts to serum [see Comments]. *Science* **283,** 83–87.
14. Johnson, R. A. and Wichern, D. W. (1992) *Applied Multivariate Analysis*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.
15. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., et al. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406,** 536–540.
16. Reinitz, J. and Sharp, D. (1995)Mechanism of eve stripe formation. *Mech. Dev.* **49,** 133–158.
17. Wahde, M. and Hertz, J. (1999) *Coarse-grained reversed engineering of genetic regulatory networks*. Technical Report, Nordic Institute for Theoretical Physics.
18. Chen, T., He, H., and Church, G. (1999) Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing* 1999, **4,** 29–40.
19. D'Haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing* Jan. 4–9, 1999, Hawaii **4,** 41–52.