Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature* **436,** 876–880.

Li, W., Meyer, C. A., and Liu, X. S. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21**(Suppl. 1), i274–i282.

Lieb, J. D., Liu, X., Botstein, D., and Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genet.* **28,** 327–334.

Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., Fraenkel, E., Bell, G. I., and Young, R. A. (2004). Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303,** 1378–1381.

Sabo, P. J., Humbert, R., Hawrylycz, M., Wallace, J. C., Dorschner, M. O., McArthur, M., and Stamatoyannopoulos, J. A. (2004). Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl. Acad. Sci. USA* **101,** 4537–4542.

Scacheri, P. C., Davis, S., Odom, D. T., Crawford, G. E., Perkins, S., Halawi, M. J., Agarwal, S. K., Marx, S. J., Spiegel, A. M., Metzer, P. S., and Collins, F. S. (2006). Genome-wide analysis of menin binding provides insights into MEN1 tumorigenesis. *PLoS Genetics* **2**(4), e51.

Wu, C., Wong, Y. C., and Elgin, S. C. (1979). The chromatin structure of specific genes. II. Disruption of chromatin structure during gene activity. *Cell* **16,** 807–814.

[15] Extrapolating Traditional DNA Microarray Statistics to Tiling and Protein Microarray Technologies

By Thomas E. Royce, Joel S. Rozowsky, Nicholas M. Luscombe, Olof Emanuelsson, Haiyuan Yu, Xiaowei Zhu, Michael Snyder, and Mark B. Gerstein

Abstract

A credit to microarray technology is its broad application. Two experiments—the tiling microarray experiment and the protein microarray experiment—are exemplars of the versatility of the microarrays. With the technology's expanding list of uses, the corresponding bioinformatics must evolve in step. There currently exists a rich literature developing statistical techniques for analyzing traditional gene-centric DNA microarrays, so the first challenge in analyzing the advanced technologies is to identify which of the existing statistical protocols are relevant and where and when revised methods are needed. A second challenge is making these often very technical ideas accessible to the broader microarray community. The aim of

0076-6879/06 $35.00
DOI: 10.1016/S0076-6879(06)11015-0

this chapter is to present some of the most widely used statistical techniques for normalizing and scoring traditional microarray data and indicate their potential utility for analyzing the newer protein and tiling microarray experiments. In so doing, we will assume little or no prior training in statistics of the reader. Areas covered include background correction, intensity normalization, spatial normalization, and the testing of statistical significance.

Introduction

Microarray technology (Fodor *et al.*, 1991; Schena *et al.*, 1995) allows for the parallel quantitative assessment of biochemical reactions. On the order of 10^6 measurements can be taken simultaneously with current technology (Cheng *et al.*, 2005). The initial challenge following a microarray experiment is to determine which of these potentially millions of observations are significant and should be studied in more depth. This challenge has been met by hundreds of practitioners in both biomedical and mathematical sciences and literally hundreds of papers have been published on the topic. This chapter aims to illustrate some prevailing ideas and techniques found in the microarray analysis literature. In addition to covering statistics used for traditional microarray experiments, we include those techniques exploited in protein and tiling microarray analyses as well. These latter experiments share some mechanistic aspects with the traditional DNA microarrays, but in several respects, are quite different. Therefore, some of the bioinformatics research done for traditional microarrays is relevant, whereas some of it is not. We will guide our discussion with this as our theme, and focus on two main areas of study: microarray normalization and the assessment of statistical significance.

Prior to delving into the heart of our discussion, we will first introduce some naming conventions, followed by statistical preliminaries. Following these prerequisites, a brief discourse on how microarray data are obtained is given. The first major area of study reviewed is *microarray normalization* or, more concisely, *normalization*. Normalization deals with the technical aspects of the microarray technology that can potentially confound and/or bias the results of the experiment. It does so by correcting measured values so as to remove these effects. Normalization is discussed later. The second area focused on is the assessment of statistical significance. Statistical significance can mean different things for different microarray experiments, depending on their respective goals, and is discussed. In a majority of traditional DNA microarray experiments, significance indicates the presence of differential mRNA expression between two or more biological classes for some gene. An experiment might, for example, assess mRNA concentrations for thousands of genes as cells progress through the cell cycle

(Cho *et al.*, 1998). In such a scenario, we would like to know within each stage those genes that exhibit differential expression (higher or lower concentrations) relative to the other stages. For tiling microarrays, as shown later, significance pertains more loosely to genomic regions. In these experiments, we seek chromosomal regions (consisting of multiple probes) that exhibit higher than expected fluorescent intensities on the microarray. Protein microarrays have two main classes of use: analogous to the DNA microarray, *antibody microarrays* can be used to determine protein abundances, whereas *functional protein microarrays* can be used to detect protein–protein interaction partners *in vitro*. For each of these experiments, significance clearly takes on a different meaning.

Definitions

Some common points of confusion within the microarray literature are how various entities are defined. This section explicitly defines some of these entities so as to minimize the potential for confusion. Herein, we define molecules on the microarray at time of its construction as *probes* and those molecules that are subsequently introduced to the microarray as *targets*. We use the words *spot* and *feature* interchangeably to indicate a collection of probes that have the same sequence and are concentrated at a known position in the microarray design. A collection of targets from a single biological source is called a *sample*. A single event consisting of introducing one or more samples to a microarray is termed *probing*. Finally, a set of probings designed to test certain hypotheses is simply an *experiment*.

Statistical Preliminaries

It is impossible to have a discussion on microarray statistics without any prior knowledge of statistics in general. This section provides some basic concepts that will aid our presentation of microarray analysis. Anyone who has taken an introductory statistics course has seen this material already and can safely skip this section.

Summary Statistics

Assume for the moment that a microarray experiment measures the expression level of just a single gene and that the experiment consists of several technically replicate probings from which a measurement is observed. To generalize the measurements for discussion, let each measurement be denoted by the symbol X_i. Here, the subscript i indicates the ith measurement of the gene. For example, $X_4 = 162$ would indicate that the measurement coming from the fourth microarray is 162.

A first natural question to ask of the experiment is ''What is the central tendency of my measurements or, equivalently, how can I best describe my measurements with a single number?'' The most commonly used response to this question is to calculate the *arithmetic mean*, or *average*, of the measurements. To calculate the arithmetic mean, we first sum all the measurements and then divide by the total number of measurements observed. If N is the number of measurements taken, then the mean $\bar{X}$ is calculated as

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{X_1 + X_2 + \ldots + X_N}{N}. \tag{1}$$

We often would like to measure the spread of our measurements in addition to their central tendency. The most commonly used measure of spread is the variance σ^2:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{(N-1)}. \tag{2}$$

Note that the numerator consists of N terms, added together. Each term in the summation corresponds to the ith measurement and is the difference between that measurement (X_i) and the mean of all N measurements, $\bar{X}$. Also note that each term is squared. Doing so ensures that the numerator is positive and that measurements less than the mean contribute positively to the variance just as much as those measurements greater than the mean. This measure of spread is roughly the average squared difference from the mean. We say ''roughly'' here because the denominator in Eq. (2) is ($N - 1$) rather than the N that we might expect from the definition of arithmetic mean [Eq. (1)]. Why this is so is beyond our scope, but with large N this detail makes little difference. Related to the variance is a quantity called the *standard deviation*. A standard deviation, symbolized as σ, of a group of measurements is simply the square root of those measurements' variance.

We often read or hear the phrase ''microarray data are noisy,'' or some similar (potentially less polite) variant. This can be taken to mean several things, but quite often it is the presence of outliers that is being referred to. An *outlier* is a measurement in large disagreement with other measurements of the same phenomenon. In a microarray experiment, the difference could be due to a biological effect, but more likely the outlier is due to some kind of technical malfunction of the instrument and/or its associated protocol(s). Outliers can have large effects on the aforementioned summary statistics. For an example, consider an experiment where five measurements are taken for the same gene. If these measurements are $X_1 = 12$, $X_2 = 9$, $X_3 = 11$, $X_4 = 507$, and $X_5 = 12$, then

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{N} = \frac{12 + 9 + 11 + 507 + 12}{5} = 110.2. \qquad (3)$$

Clearly the quantity 110.2 does not represent the central tendency of data very well. It is not particularly close to any of the measurements. Luckily, there are ways around such pitfalls. One technique is called the trimmed mean. With this approach, some percentage of the most extreme measurements is thrown away prior to calculating the mean. An extreme (and quite common) version of this approach is to calculate the measurements' *median* as a measure of their central tendency. The median is defined as the middle quantity occurring in a sorted list of observations. That is, if N is odd and you first sort your measurements $X_1, X_2, \ldots, X_N$ in either increasing or decreasing order, then the median is the quantity $X_{\frac{N+1}{2}}$. (If N is even, the middle two measurements, $X_{\frac{N}{2}}$ and $X_{\frac{N}{2}+1}$, are averaged.) In our noisy example of five measurements where the mean of 110.2 was obtained, the calculated median is 12. This value intuitively summarizes these data much better.

An analogous calculation can be performed in place of the variance. Recall that the variance is essentially an average squared difference of measurements from the mean [Eq. (2)]. This computation can be made more robust to outliers by first substituting the median for the mean and then computing the median of absolute differences between the measurements and the previously calculated median. This quantity is sometimes referred to as the *median absolute difference* (MAD).

Statistical Significance

The term *p value* comes up frequently in texts about microarray experiments and their analyses. A p value is simply the probability of some null hypothesis being true given a set of assumptions and observations. A typical experiment utilizing DNA microarrays might have thousands of such null hypotheses, one for each gene being studied. These null hypotheses would typically claim that the expression level of some gene is not different between two biological samples. As a result, we declare that any gene for which we can compute a low p value is *significant* and potentially worthy of further study. To call the gene significant, a p value threshold for significance must, of course, be in place. A common interpretation for this threshold is the false-positive rate of the study: the percentage of time replications of the experiment would reject the null hypothesis when it is actually true.

Now, we do not typically know the actual probability of observing something under a given null hypothesis. However, if we know that the numbers being studied follow some known form (e.g., we might know or assume that the gene expression levels are distributed like a bell curve, or a *normal* distribution), then we can use this knowledge to either simulate or directly calculate how likely an average difference between two such groups of measurements would be if there were in fact no difference, for example.

A final note on significance worth noting is that, generally speaking, the more observations we are able to make of some phenomenon, the better is our ability to compute a low p value. To illustrate this point, consider an experiment where we ask, "Is gene A expressed at a higher level in tumors than in healthy tissue?" Let us assume that the answer to this question is, "Yes." If we have one measurement of A from a tumor and one measurement of A from a healthy tissue and the measurement from the tumor is twice as high as its healthy counterpart, we have some limited confidence that the gene is more highly expressed in tumors. This occurrence could be an anomaly, so we still would assign some fairly high probability to the null hypothesis of no difference being. If instead we measure the abundance of gene A in 20 tumors and they are all higher than 20 measurements taken from healthy tissues, we would assign a much lower probability to the null hypothesis because the chance of 20 anomalies is very small.

Multiple Testing

Another issue that comes up frequently in the microarray literature is that of *multiple testing*. Multiple testing simply indicates that more than one statistical test (which generates a p value) is part of the study. For microarray experiments there are thousands and potentially millions of statistical tests being conducted, so clearly we are dealing with multiple testing, but what is our concern when we engage in multiple testing?

In biology, the threshold for considering a p value significant is typically $p < 0.05$ or $p < 0.01$. These criteria arise from a balance between our willingness to accept a 5% or even a 1% false-positive rate and the number of replicate measurements we are able to take. Multiple testing becomes a problem, for example, if we conduct 100 statistical tests and identify that one of them yields a significant p value ($p = 0.04 < 0.05$). It would be tempting to report this seemingly significant finding. The problem here is that within a set of 100 tests, we expect to find 4 of these to yield $p = 0.04$ simply due to random chance (100 tests multiplied by the false-positive rate 0.04 yields 4 tests). This toy example becomes a staggering problem if we are testing, say, 20,000 genes. In this case, at a significance threshold of $p < 0.05$, we will identify roughly 1000 false positives. This number of false positives is potentially more than the actual number of differentially expressed genes that we seek to identify.

The most simple method for dealing with multiple comparisons is to require sufficiently low p values such that the total number of expected false positives is small. The Bonferroni correction (Bonferroni, 1935) does this by controlling the so-called *family-wise error rate* (FWER). The FWER is defined as the probability of detecting a false positive anywhere among the multiple tests. As a result if we want the probability of detecting a false positive among our tests to be less than α, we require that any individual test achieve $p < \frac{\alpha}{N}$ where N is the number of tests. Such corrections pose a problem for microarrays where thousands of genes are being tested for significance and the number of available replicate experiments is small. The problem is more acute for high-density tiling microarrays where the number of tests performed can reach into the millions (see later) and the number of experimental replications is often fewer than five.

Microarray Data

This section reviews briefly how microarray data are obtained.

Data for Traditional, Gene-Centric DNA Microarrays

Each spot on a gene-centric DNA microarray corresponds to a DNA sequence derived from a known or putative gene. That sequence could be the whole spliced form of a gene (such as a cDNA clone) or a tethered 25-bp oligonucleotide sequence, as is the case for Affymetrix GeneChip brand microarrays. Such a microarray typically probes a sample that is derived from a mRNA source.

Subsequent to probing a labeled sample with a microarray, an image representing its surface is generated by subjecting the microarray to a digital-scanning device. Depending on the type of labels used, different scanning technologies are employed. Typically, the samples have been labeled with a fluorescent dye or, alternatively, with radioactive isotopes. For fluorescently labeled samples, the probed microarray is scanned with a laser scanner. There is a wide selection of laser scanners available, including but not limited to ScanArray GX from Perkin-Elmer, GenePix 4200 from Molecular Devices, and DNA microarray scanner from Agilent Technologies. A laser wave length near the absorption maximum of the fluorophor dye used (that was attached to the hybridizing sample) is scanned across the microarray surface from top to bottom and from left to right so that all areas of the microarray are accessed by the laser. The light emitted at each location when laser-excited fluorophors transition to their unexcited state is captured by a detector and translated into a pixel intensity at that location. Microarrays that probe multiple-labeled samples simultaneously must be scanned with a scanner having at least as many unique laser wavelengths as labeled samples. Current scanning resolutions are as high as 1 μm^2 per pixel.

The result of scanning a single probed sample is a monochromatic digital image (usually stored as a TIFF file) of the microarray surface. Bright regions in the image correspond to regions of the microarray with high levels of fluorescence and dim regions likewise correspond to regions devoid of fluorescence. Presumably, the bright regions correspond to spots to which labeled nucleic acid hybridized. If two different samples were labeled with two different dyes and were probed with the same microarray, then the result of scanning is two digital images. There would be one image for each wavelength used.

The microarray images generated by the laser scanner must be further processed with image analysis software. First, the spots of the microarray have to be identified within the image. To do this, rules have to be obtained or assumed that can distinguish between pixels that constitute spots and pixels that belong to background regions. Separating spots from the background is called *segmentation*. Following segmentation, *grid alignment* must be performed. Grid alignment is the process of identifying which spots correspond to which annotation. Basic versions of grid alignment software are usually included with the purchase of a scanner, but there are alternatives, such as TIGR Spotfinder (Saeed *et al.*, 2003), which is freely available under an open-source license, or ScanAlyze (http://rana.lbl.gov/EisenSoftware.htm), which is free for academic and noncommercial use. For most spotted arrays, the grid has to be defined by the user, either manually or semimanually, whereas for many higher density microarrays, such as Affymetrix GeneChip brand microarrays and NimbleGen System's NimbleChips, the alignment of the microarray image to the grid is done automatically by software. This automation is made possible by reserving some spots on the microarray exclusively for grid alignment. Certain labeled cDNA/cRNA molecules that are complementary to the grid alignment probes are spiked into the sample(s), ensuring that the grid alignment probes will appear as bright regions in the scanned image, enabling automatic grid alignment.

After aligning the grid the image analysis software reports back a certain number of key statistics for each of the identified spots. These statistics may include the mean and median pixel intensities within each spot, the standard deviation of those pixels, and sometimes also other information, such as mean intensity ratios in the case of a two-channel experiment. The area of each spot (number of pixels) is also frequently reported. Importantly, if the software considers a particular spot aberrant, for example, irregular in its shape, or if its measured intensity is lower than the surrounding background intensity, the spot may be flagged as irregular. Such flagged spots are often excluded from further statistical analyses.

The end result is a tab-delimited plain text file containing all raw data for each microarray feature within a single row. The tab-delimited, text-based format is easily amenable to further analysis by importing it into a microarray analysis software package such as ExpressYourself (Luscombe *et al*., 2003) or MIDAS (Saeed *et al*., 2003) or, of course, into your own microarray analysis pipeline. For simple calculations, a spreadsheet program (e.g., Gnumeric, OpenOffice or Microsoft Excel) could also be used.

For spotted DNA microarrays, it is common that each gene under study is represented by a single spot. An important difference exists for Affymetrix GeneChip brand microarrays. For this technology, each gene is represented by a *probe set*, typically consisting of 10–20 features on the microarray. Within the probe set, each feature contains probes of different sequence. To assess the differential expression of a single gene, multiple spots from each microarray need to be considered.

Data for Tiling Microarrays

The two most widely utilized high-density oligonucleotide platforms are those produced by Affymetrix, using masks to synthesize the oligonucleotides on the microarray (Lipshutz *et al*., 1999), and those manufactured by NimbleGen Systems, which use a system of mirrors controlled by a digital light processor for synthesis (Nuwaysir *et al*., 2002). Affymetrix microarrays currently utilize 25-bp oligonucleotide probes for each spot. For every spot corresponding to some 25-bp stretch of genomic DNA (perfect match), there is a corresponding spot (mismatch) where the middle nucleotide of the probe has been substituted with its reverse complement. This perfect match/mismatch setup is also the standard for the Affymetrix GeneChip system as well. The purpose of the mismatch probe in both traditional and tiling applications is to measure the nonspecific binding of the probes within a spot [there is some debate about the usefulness of mismatch probes, however (Irizarry *et al*., 2003)]. Currently, Affymetrix microarrays are capable of including on the order of 10^6 spots.

Maskless microarrays manufactured by NimbleGen Systems are synthesized such that each microarray can be completely customized with unique probe sequences. These microarrays allow for oligonucleotide lengths of up to 70–80 nucleotides (in fact, isothermal arrays exist where each feature corresponds to a oligonucleotide probe of a different length). Current maskless microarray designs have approximately 390,000 spots per microarray. One important difference between these two high spot density platforms is that Affymetrix brand microarrays can only be hybridized with a single target nucleic acid population, whereas maskless arrays allow the hybridization of two samples simultaneously using different labels, typically

Cy5 (red) and Cy3 (green). This is potentially beneficial when looking for differential expression between samples or for ChIP-chip (Horak and Snyder, 2002; Iyer *et al.*, 2001), where chromatin-immunoprecipitated DNA is labeled differently from some reference DNA.

Tiling microarrays (Bertone *et al.*, 2004; Cawley *et al.*, 2004; Cheng *et al.*, 2005; Kapranov *et al.*, 2002) use high-density capabilities to tile the nonrepetitive sequence of a genome. The word *tile* indicates that probes are selected for inclusion on the microarray at some roughly uniform interval over a potentially large genomic space. In the context of mRNA transcript mapping, this high resolution enables the unbiased detection of individual exons of a spliced transcript. This experiment is not practical on a whole-genome scale in a mammalian species with lower resolution polymerase chain reaction amplicon microarrays due to cost (Bertone *et al.*, 2005).

Tiling microarrays are an evolving medium, and data format standards have not yet materialized. However, several summary statistics about each spot are typically included in a tab-delimited text file. These statistics usually include the mean and/or median pixel intensity of each spot, the number of pixels within each spot, and a standard deviation of the pixel intensities of the spot. It is worth including a cautionary note about tiling microarray data here. Tiling microarrays generate very large data sets. As such, they are difficult or impossible to import into desktop spreadsheets such as Microsoft Excel. Therefore, more robust tools are often needed.

There is one more major difference between traditional DNA microarrays and tiling microarrays to consider. The signal intensity measured at a spot containing short oligonucleotide probes is arguably too unpredictable to score each probe separately. This variability is due to a number of factors, including cross-hybridization and differential binding affinity due to probe sequence and other sequence-based artifacts. In addition, higher standards of statistical significance are typically required for tiling arrays because of the much larger number of spots being queried and therefore require more evidence than that given by a single spot. Thus the methodologies that have been adopted for the analysis of tiling microarrays is to incorporate the intensities of a number of spots that lie within a contiguous genomic region. This methodology is often referred to as a *genomic sliding window* approach.

Data for Protein Microarrays

There are two types of protein, microarrays as defined by their goals (Zhu and Snyder, 2003). One type is protein detection microarrays, or antibody microarrays (Lueking *et al.*, 1999), which use antibodies for its probes and are used to detect and quantify proteins in solution. This design is very similar to its DNA-based counterparts, which quantify mRNA concentrations. The other major class of protein microarrays is functional protein microarrays (Zhu *et al.*, 2001), which aim to identify protein binding or modification capabilities. In such a design, each spot consists of some known protein or protein domain. The target that is introduced will typically consist of a single macromolecule. This target may be labeled so as to detect molecular interaction partners or, as is the case for kinase activity assays, may be probed in the presence of hot ATP to detect phosphorylation events.

An aspect of protein microarrays of note is that the spots therein will usually not contain equal amounts of protein from spot to spot. This discord can cause differences in measured intensity between spots that are not due to molecular activity, but rather to an aspect of the microarray construction.

Regarding software and generated data, protein microarrays utilize the same scanners and scanning software as their DNA-based counterparts and therefore the raw data files they produce are technically very similar. This is an advantage, as some existing computational protocols and interfaces developed for DNA microarrays may be integrated easily with protein microarray analysis.

Microarray Normalization

Once data have been obtained, a usual next step is to perform microarray normalization.

Motivation

Technical aspects of the microarray experiment can cause systematic biases and artifacts to be present in their data. In a two-sample DNA microarray experiment, the probed biological samples may contain different concentrations of RNA, leading to an overall bias in favor of greater measurements in one channel. In addition, the fluorescent dye molecules Cy3 and Cy5 are known to have slightly different properties, leading to a similar problem. Complicating these troubles is that they may be more or less present depending on the intensity of the spot being measured and/or its physical location on the microarray. The following section illustrates an example of how such biases can affect biological conclusions made from microarray data when proper data normalizations are not carried out. We include this example as a cautionary tale and as a motivation for microarray normalization, in general.

Most spotted microarrays are built by depositing solutions of cDNA clones at known locations on a microarray surface. This deposition process is controlled robotically with little human intervention and is therefore completely regular and predictable. Furthermore, the printing process is such that spots close to each other on the microarray surface are printed closely in time as well. Given that a microarray hybridization can be uneven across the surface of the microarray, this leads one to speculate that neighboring spots on the microarray surface might be coordinately affected. An example situation would be if labeled sample were more abundant in one region of the microarray than in others. Spots in that region would have systematically higher observed intensities than those spotted elsewhere.

Indeed, it does appear that such a spatial effect exists. For printed cDNA microarrays, the effect was first reported by examining the relationships between observed spot intensities and the locations of spots in the design of a microarray (Kluger *et al.*, 2003; Qian *et al.*, 2003). Similarities were examined between gene expression profiles (across a large number of probings) for genes that are printed on the microarrays at varying distances. It was found that genes that are close in the microarray design (on average) have higher similarities between their expression profiles than those further away. That is, it might appear that genes that are close on the microarray surface seem more likely to be coexpressed. Note that without knowledge of the microarray design, the genes would be identified as exhibiting coordinated mRNA expression.

It turns out that for the microarray design used in the aforementioned study, genes were printed in an order related to their chromosomal arrangement for organizational convenience. This printing strategy yielded a microarray such that genes located 22 open reading frames (ORFs) away in genomic space are printed as immediate neighbors on the constructed microarrays more often than would occur if they were printed in a random order. Interestingly, by examining the relationships between gene expression and chromosomal localization, a striking similar frequency was found: genes that are approximately 22 ORFs away on the same chromosome are more likely to be coexpressed, whereas genes that are about 11 ORFs away are less likely to be coexpressed (Qian *et al.*, 2003). Furthermore, it was determined that genes on microarrays with a different layout have a different frequency. This last piece of evidence suggests the existence of an artifactual effect related to microarray architecture. One of the aims of microarray normalization is to reduce the effect of such artifactual components of observed data.

Most microarray studies examine the relationship between two biological samples by comparing their relative mRNA expression levels. The idea behind such two-channel experiments is straightforward: labeled (typically red with Cy5 and green with Cy3) nucleic acids in the samples are probed simultaneously with a microarray slide, and relative abundances are derived from comparative fluorescence of the nucleic acid molecules hybridized at each microarray feature. For a given spot i, the relative concentration between the two samples is commonly represented as the log ratio, λ_i, of the measured fluorescence intensities between the two dyes. We summarize the log ratio as

$$\lambda_i = \log\left(\frac{R_i}{G_i}\right) \tag{4}$$

where R_i and G_i denote the observed intensities (mean or median of spot pixels' intensities) for probe i when scanned with red and green lasers, respectively. Note that a log ratio of zero indicates that R_i and G_i are equal. Further, a set of observed log ratios (with measurement error) should center about zero for probes representing genes of equal expression in the two samples. Measurements deviate from this situation proportionately to their degree of up- or downregulation relative to the two samples.

The log ratio measured between a gene in two samples is in itself a normalization technique. Microarray manufacture is not errorfree. Any given spot may be printed poorly on one microarray and printed perfectly on the next. If these two microarrays were used to measure the concentration of the gene corresponding to that spot, the poorly printed spot would likely lead to an artificially low measurement for one sample relative to the sample hybridized to the higher quality spot. If instead both samples were hybridized to both microarrays, then the hybridization of one sample to the poor spot is directly comparable to the hybridization of the other sample to the poor spot and likewise for the higher quality spot. This self-normalization is particularly useful when the two samples hybridized to the microarray are paired in other respects beyond the fact that they were measured with the same instrument. A good example of paired samples is an mRNA sample taken from a tumor biopsy before treatment and an mRNA sample taken after treatment. Regarding log ratios, it should be noted that the Affymetrix GeneChip system only allows hybridization of a single sample to a microarray. Therefore, log ratios are not meaningful as a spot quality normalization. It is believed, however, that Affymetrix microarray construction is much more uniform in terms of quality control than its spotted microarray counterpart so such self-consistency concerns are relatively minor. Log ratios can still be relevant for Affymetrix microarrays in the case of paired samples, such as in the cancer experiment mentioned earlier. Most tiling and protein microarrays yield just a single intensity measurement as well so the log ratio is not always a natural measurement for these experiments either.

Although the log ratio provides an intuitive measure of relative gene expression, it must often be corrected for inconsistencies resulting from the experiment (see earlier discussion). Such corrections are collectively termed normalization. Normalization adjusts the measured intensities for each sample and for each spot as corrective measures. The aim is to compensate for artifactual effects by applying transformations so that equally expressed genes have log ratios approaching zero. (For single-channel experiments, no such baseline generally exists.) Measurements for all spots on the microarray are scaled relative to this baseline. In practice, implementing good normalization has proved challenging; researchers have developed many competing methods, which can lead to divergent results (Hoffmann *et al.*, 2002). The following sections describe some of the more widely implemented strategies.

Background Correction

For many types of microarrays, a measurement of the local background of each spot is recorded in addition to the foreground intensity of the spot. This measurement is, in common practice, the mean or median of all pixels residing in the surrounding regions of the spots (see earlier discussion). It is believed that any measured intensity from this background region is also measured in the foreground pixel intensities of the spot as well. This background fluorescence is attributable, in general, to glass fluorescence and unincorporated label molecules. The background intensities have no biological interpretation so we would ideally like to remove their contribution from spot intensities before proceeding. The easiest way to make this correction is to subtract the mean (median) of all local background pixels measured in the red channel (denoted ρ_i) from the red intensity of each spot, do likewise for the green channel (γ_i), and then compute the background adjusted log ratio as

$$\hat{\lambda}_i = \log\left(\frac{R_i - \rho_i}{G_i - \gamma_i}\right). \tag{5}$$

Equation (5) assumes that $\rho_i < R_i$ and that $\gamma_i < G_i$. Any spot not in agreement with these assumptions should be flagged as a bad spot and subsequently ignored, as it does not make sense for a background region to have a higher intensity than the spot.

We need not rely upon just the background values provided with each spot in a microarray results file. The values of ρ_i and γ_i could actually be computed as the mean or median of all spot background measurements in a localized region before applying Eq. (5). An example of this would be to utilize a spot's eight nearest-neighboring spots' backgrounds to calculate its local background intensity. Such an approach is advisable so as to avoid aberrantly high local background values due to scratches or other artifacts present in a microarray scan. This is of particular importance when dealing with protein microarrays, as these devices can yield spots that smear to bigger sizes due to phosphorylation activity, for example. These smears will often be measured as part of a spot's background, causing it to be erroneously high.

Unfortunately, tiling microarrays will usually seek to maximize feature density in an effort to reduce cost and as such, features are packed immediately next to one another and background calculations may not be possible. In these cases, we can only hope that background intensities are minimal, or at least, constant throughout the microarray.

Normalization via Total Intensity

Following background subtraction, we would like to normalize sample intensities so that their intensity distributions have desirable properties. One commonly desired property within two-sample probings is to have a distribution of log ratios representing nondifferentially expressed genes to center about zero. This is usually reasonable, as in most experiments we do not expect a centering around any other value.

In a differential expression experiment, microarrays should hybridize similar numbers of labeled molecules from each sample, so the total hybridization signals summed over all probes should be the same for both channels. Using these assumptions, we can calculate a scaling factor C_{total} that can be used to correct any observed deviance from this assumption. If M is the total number of features on the microarray, then we have

$$C_{\text{total}} = \log\left(\frac{\sum_{i=1}^{M} R_i}{\sum_{i=1}^{M} G_i}\right). \tag{6}$$

We can then compute the normalized log ratios as

$$\hat{\lambda}_i = \lambda_i - C_{\text{total}}. \tag{7}$$

The result is a distribution of log ratios that are centered somewhere near zero. This method performs well in most standard microarray experiments with sufficiently large numbers of spots (>20,000), as in these scenarios, outlier signals make negligible contributions to the total intensities.

A similar approach to Eq. (6) can be used to normalize intensities from one single channel microarray to others. In this application, every intensity in one channel is divided by the summed intensity (e.g., $\sum_{i=1}^{M} R_i$) of spots from the same microarray. Then, these normalized intensities can be used to compare and contrast different samples hybridized to different microarrays. This latter calculation may be useful in experiments where just a single probing is carried out on each microarray. This is always the case for Affymetrix GeneChip technology and is almost always the situation for protein microarrays and for tiling microarrays.

Normalization via Gene Set

The previous method performs fairly well in standard microarray experiments where the number of genes studied is large and overall gene expression differences between the two samples are not excessive. However, the approach must be applied cautiously, as it may mislead researchers into believing that similar numbers of genes are always up- and downregulated. This clearly is not true in some circumstances.

In the following method, sometimes called the *gene set method*, some set of genes is assumed not to be expressed differentially between the samples being studied. This set of genes is typically made up of housekeeping genes. The procedure is analogous to that in Eq. (7), with the only difference being the numbers that are summed are those from the gene set, not all spots. We call this value C_{geneset}. Captured in this statistic is the overall deviation that you would expect given no differential expression. Ideally, C_{geneset} is equal to zero, but effects such as unequal RNA concentrations and differences between the fluorescent dyes can cause C_{geneset} to be nonzero. Once C_{geneset} has been calculated, all log ratios (not just those in the gene set) are normalized by C_{geneset} using the relationship

$$\hat{\lambda}_i = \lambda_i - C_{\text{geneset}} \tag{8}$$

where $\hat{\lambda}_i$ denotes the normalized log ratio for probe *i*. Using control spots in this way has an added benefit for sets of microarrays where the spots present on each microarray are not the same. In such a scenario, a common set of control spots can be used to normalize the intensity distributions of the microarrays so that they are similar from microarray to microarray. This is a typical situation for tiling microarrays that require several microarrays having different designs to probe for large fractions of the sequence of a genome.

Normalization via Spiked Controls

A way to guide normalization further is to spike known quantities of external controls into the biological samples prior to fluorescence labeling. Normalization is then based on balancing the signal intensities for those probes corresponding to the control RNA molecules as in Eq. (8).

There are two advantages of this technique. First, the spike-ins are completely controlled—we are sure that they should show no differential expression between two or more samples. Second, different scale factors can be calculated for genes having different expression levels if several different spike-in concentrations are used. A disadvantage, though, is that control probes must be built into the array at the onset. Further, the scaling factor is calculated using a comparatively small number of probes that may be sparsely distributed on the array depending on the design and the correction techniques for spatial microarray biases (discussed later) currently cannot be incorporated easily. A final point of concern is that spiked controls may interact with unintended spots on the microarray in addition to the control spots. For traditional DNA microarrays and tiling microarrays this is manifest as cross-hybridization. For protein microarrays, spiked proteins may interfere with desired protein-binding interactions.

Normalization via Quantiles

Another popular alternative for intensity normalization is so-called *quantile normalization* (Bolstad *et al.*, 2003). In this approach, the first step is to construct a synthetic microarray such that the ‘‘measurement of each spot,’’ S_i, is the mean or median of its measurements across all P probings in the experiment. Mathematically, if we use the mean in constructing this synthetic microarray and $X_{i,j}$ is the measurement from the jth probing for spot i, then we have

$$S_i = \frac{\sum_{j=1}^{P} X_{i,j}}{P}$$

The S_i values are then sorted in increasing order, as are the intensities within each probing. The final step in this normalization is to replace each observed intensity by that intensity S_i that occupies the same position within its sorted list. If $X_{1043,2} = 87$ is the third largest observation within probing number two, it is replaced by the third largest value of S. A major advantage of this approach is that it requires no extra probes or spike-ins and yet still can correct for biases that may be present more or less at different intensity levels. This advantage makes this method broadly applicable to any microarray experiment with little concern over experimental nuances.

Correcting Signal Intensity Bias

Numerous reports have indicated that log ratios resulting from a two-sample probing can have a systematic dependency on signal intensity because of differences in the fluorescent properties of the red and green dyes (Quackenbush, 2002; Yang *et al.*, 2002a,b).

Lowess regression (Cleveland, 1981) analysis allows its users to fit a nonlinear curve to a ratio vs intensity distribution. We call the logged product of the measured R_i and G_i intensities ϕ_i and plot each λ_i as a function of its respective ϕ_i. The basic idea of Lowess is then to first find a curve that passes through the ''middle'' of this ratio versus intensity distribution. The output of Lowess is a value L_i paired with each ϕ_i. Once L_i is calculated, it can be used to correct for intensity biases. The corrected log ratio is

$$\hat{\lambda}_i = \lambda_i - L_i. \tag{9}$$

The question remains as to how L_i is calculated. This is somewhat beyond the scope of this chapter, but we will sketch the calculation here. For every ϕ_i, a neighborhood of ϕ values is found. The size of this neighborhood is a variable that can be adjusted but is typically set to be 10% of all spots. Once the neighborhood of spots is found, a line is plotted through the values corresponding to the ϕ values in the window. This line is used as a function to compute L_i from ϕ_i. The method can be generalized. In fact, a commonly used variant of this method called Loess (no ''w'') performs the same functionality but replaces the locally fitted line with a locally fitted quadratic curve.

This technique has no analog for single-channel experiments as in most tiling and protein microarray experiments. The technique can be forced if one microarray is considered a baseline and then all other microarrays are normalized relative to the baseline. This is potentially problematic for tiling microarrays where each microarray may contain different probes and therefore have different expected intensity distributions.

Correcting Array Location Bias

It has become increasingly clear that there are often substantial spatial biases caused by uneven hybridization conditions across a microarray slide. Uncorrected, this can have an effect on results. An example of this is the apparent coexpression of groups of genes, which is actually caused by the proximity of their corresponding spots on the microarray surface (see earlier discussion).

For spotted microarrays, the effect is frequently corrected using subgrid normalization in which local subsets of spots are grouped by their depositing print tip. These groups are then normalized separately using, for example, the method outlined earlier. This approach should be used with caution, as we have observed that most spatial variations do not follow the boundaries of print-tip groups (sometimes referred to as *blocks*).

As an alternative, a variation of the Lowess analysis introduced earlier can be used to correct spatial biases. In this alternative, a surface is fit to the log ratios as a function of their spatial coordinates as opposed to fitting curve to log ratios as a function of total intensity. The corrected intensity is obtained analogously. This procedure can be applied to single-channel intensities as well.

It should be noted here that during the design of a microarray, no regions should be overpopulated with spots that might display coordinated expression level changes. In this unfortunate scenario, the corrective methods will eliminate biologically meaningful variations in the measurements. This limitation can be overcome easily by randomizing spot locations during microarray manufacture.

This procedure may prove difficult for microarrays where a small fraction of spots show measurable signal because there are too few intensities to fit the surface to. Tiling microarrays will usually fall into this category as much of the genomic sequence is inactive at any given time. Functional protein microarrays may fall into this category as well, as a given protein is likely to have just a handful of binding partners.

Normalization by Spot Concentration

Concentrations of probes within each spot will affect measured intensities. For most traditional and tiling microarrays, this is not an issue. However, for protein microarrays, it is difficult to control the amount of protein present at each spot and therefore it is advisable to divide any measurement by the concentration of the spot. The concentration measurements can be obtained by hybridizing a protein microarray with a labeled universal protein marker.

This section briefly described the most common techniques for normalizing microarray data. Many of these methods have been implemented in published software tools that facilitate microarray normalization; examples include Express Yourself (http://bioinfo.mbb.yale.edu/expressyourself/) (Luscombe *et al.*, 2003) and SNOMAD (pevsnerlab.kennedykrieger.org/snomadinput.html) (Colantuoni *et al.*, 2002).

Future improvements in microarray technology may eliminate the need to correct for intensity and spatial bias, or even for normalization all together. However, current technologies still produce substantial artifacts, even if they are not evident from visual inspection of a scanned image.

Scoring for Significance

Following microarray normalization, the intensities are in a more suitable form for statistical testing. This section begins by exploring some of the more common approaches for testing the significance of differences between measured intensities generated from two biological conditions. The discussion is then generalized to the multiple condition case and to tiling and protein microarrays.

Fold Change

Assume for simplicity that we are interested in assessing differential expression for just a single gene between two biological conditions. Call these conditions A and B. Further, assume that we have multiple measurements for the gene within each condition. Let $M > 0$ be the number of measurements obtained for condition A and $N > 0$ be the number of measurements for condition B. Note that M need not be equal to N but ideally they would be equal. To designate the ith measurement from condition A, we will use the notation A_i. We adopt the same convention for measurements of B.

Perhaps the simplest technique for comparing A and B is to compute an average fold change between the two. Call this fold change statistic S_{fold} and define it as

$$S_{\text{fold}} = \max\left\{\frac{\sum_{i=1}^{M} A_i}{\sum_{i=1}^{N} B_i}, \frac{\sum_{i=1}^{N} B_i}{\sum_{i=1}^{M} A_i}\right\}. \tag{10}$$

In addition to calculating S_{fold}, we also choose cutoff values to deem the statistic potentially interesting. A good way to choose this cutoff is to have control features present on the microarray that are not expected to display differential expression. With enough unique controls, the 95th percentile of their S_{fold} statistics could be a useful cutoff. By the quantity *95th percentile*, we mean that 95% of all S_{fold} values are below this quantity. Such a cutoff would suggest that values above this threshold would occur just 5% of the time for genes not showing differential expression. More commonly, such controls do not exist and an arbitrary cutoff is selected. Often, this cutoff is set at two.

t *Test*

The fold-change method utilizes just a single summary statistic (the sum) for each condition. No information about how widely the measurements vary is considered. In addition, there must be negative control spots in the microarray design to assess how likely an observed fold change would be if the gene was not expressed differentially. Application of the t test addresses both of these issues.

The first step in carrying out a t test is to calculate the mean of measurements from A and the mean of measurements from B. We will symbolize these quantities $\bar{A}$ and $\bar{B}$, respectively. We will also need to calculate the conditions' variances, σ_A^2 and σ_B^2. The next value calculated is the standard error, SE

$$SE = \sqrt{\frac{\sigma_A^2(M-1)+\sigma_B^2(N-1)}{(M+N-2)} \times \frac{M+N}{MN}}. \tag{11}$$

The details of what this quantity represents are beyond our scope. For our purposes, it is worthwhile to note, however, that as σ_A and/or σ_B get larger, so does the standard error. SE is large when data are highly variable.

The next calculation we must make is the t statistic. This value is simply the difference between the two cell type means, divided by the standard error calculated in Eq. (11):

$$S_t = \frac{\bar{A} - \bar{B}}{SE}. \tag{12}$$

We note that as the difference between $\bar{A}$ and $\bar{B}$ becomes large, so too does the absolute value of S_t. In addition, as the uncertainty of these means grows (manifest as the variances, σ_A^2 and σ_B^2), the statistic gets smaller. Another way to view this statistic is that it expresses the differences between two means in units of (roughly) standard deviations. This is an advantage over the simpler S_{fold} statistic where variances are not considered. Another nice property about the t statistic is that it is very well studied by statisticians. In fact, we know how likely a given value of S_t is given M, N, and the null hypothesis that there is no real difference between the two means. Therefore, we can assign a p value for any value of S_t without the requirement of negative control spots.

The corresponding p values of the t statistic should be interpreted carefully, however. The knowledge we have about these probabilities assumes that the observations from each cell type are distributed normally (bell curve shaped). Unfortunately, replicate measurements coming from a microarray experiment do not always behave this way (Thomas *et al.*, 2001) and should be considered when utilizing the t test.

Another potentially troublesome aspect of the t test is that two quantities can lead to large values of S_t. The first is the value we are chiefly interested in, the difference between two conditions. The second quantity that can lead to large S_t is a small SE term. A problem with most microarray experiments is that there are few replicates available from which to calculate the standard error. This leads to the situation where SE can be quite small just by chance, resulting in high S_t values regardless of differences between the two groups of measurements. This is a situation we may not want to deem significant and worthy of further study. A useful guard against this situation is to require low p values computed with the t test *and* some fold-change criterion to consider genes for further study (Rinn *et al.*, 2004).

Significance Analysis of Microarrays (SAM)

The statistic used in SAM (Tusher *et al.*, 2001) is a slight variant of the one given in Eq. (12). The only difference is the so-called ''fudge factor'' f:

$$S_{\text{sam}} = \frac{\bar{A} - \bar{B}}{SE + f}. \tag{13}$$

The purpose of f is to disallow inflated test statistics solely due to standard errors close to zero. Effectively, it sets a lower bound on the denominator of Eq. (13). This factor gives an advantage over the t statistic but it is arguably not the greatest contribution of SAM.

In SAM, the concept of permutation testing was introduced as a means to calculate a *false discovery rate* (FDR). To perform this technique, we first fix a p value threshold T. Next, we identify those genes that have p values less than T. These are our positives. Then, for each gene, the class associations are randomized, that is, we randomly assign measurements for that gene to one of the two classes being compared. Using Eq. (13), S_{sam} is computed for each of these randomized genes. Once the S_{sam} statistics are computed along with their associated p values, the number of these p values less than T is counted. The randomization procedure is repeated a number (100 or 1000) of times and a count is made for each repetition. The median of these counts divided by the total number of genes in the study is then the reported FDR. The intuition of this is that the randomized genes represent genes that do not experience differential expression; therefore, any time one of their p values falls below T, this event can be considered a false discovery.

The notion of a FDR is an important one for microarray experiments having thousands of genes that need to be tested. It helps interpret results of an experiment in light of the multiple testing problem.

Cyber T

Equation (13) introduced the fudge factor f. The purpose of adding this factor was to guard against selecting genes that have a low mean difference and unusually low variances. Another way to protect against such situations is to apply another variation of the t test, called Cyber T (Baldi and Long, 2001). In this test, standard error is replaced by an expression that is a function of both the standard error of the gene and the standard error computed over all genes. The assumption here is that most genes should have similar standard errors; by utilizing this assumption, we can lessen the degree to which unexpectedly low or high gene-specific standard errors affect the t statistic. This method has been demonstrated to be quite powerful for detecting differences between two samples in experiments using Affymetrix GeneChip brand microarrays (Choe *et al.*, 2005).

Wilcoxon Rank Sum Test

An alternative method for computing significance levels when t test assumptions do not hold is the Wilcoxon rank sum test. This test, like many other so-called nonparametric tests, transforms measurements to their magnitude ranks and calculates probabilities based on rank-based statistics. This test was introduced in the microarray literature in Troyanskaya *et al.* (2002). (As an aside, it should be noted that when the assumptions of the t test hold, that test should be used, as it is more likely to detect a difference if it exists.)

The basic idea of the Wilcoxon rank sum test is to count the number of times a measurement from one group is greater than a measurement from a second group. The properties of how this value behaves under the null hypothesis of no difference between the groups' medians are well known so we can directly calculate a p value from this number. The actual computations of the procedure are not straightforward and lie beyond the scope of this chapter.

Wilcoxon Signed Rank Test

The previously described Wilcoxon rank sum test is generally applicable for comparing two sets of numbers. When the two sets of numbers are paired in some way (such as gene expression levels before and after a treatment), a more powerful nonparametric test is available. This test is called the Wilcoxon signed rank test. To begin, the difference D_i for the ith spot is calculated for each pair in a set of N measurements:

$$D_i = X_i - Y_i \qquad (14)$$

where X_i and Y_i are the paired measurements. Next, each D_i is assigned a rank value R_i of its absolute value

$$R_i = \sum \text{Rank of } |D_i| \tag{15}$$

Next, we sum the ranks of those D_i values that are positive

$$R_+ = \sum R_i \text{ with } D_i > 0 \tag{16}$$

and do the same summation for ranks that have negative D_i values

$$R_- = \sum R_i \text{ with } D_i < 0. \tag{17}$$

Now if we sum all ranks regardless of whether D_i is negative or positive, we will obtain the quantity $1 + 2 + \ldots + N = \frac{N(N+1)}{2}$. If there is no difference between the paired values being compared, then both R_+ and R_- should be roughly half of this previous quantity: $\frac{N(N+1)}{4}$ Therefore, if we take one of the R values as in

$$S = \min(R_+, R_-), \tag{18}$$

we known that under the null hypothesis of zero difference between the two groups, S is expected to be $\frac{N(N+1)}{4}$. We then determine how far away S is from this expected value. Again, the statistic is well studied, and given S and the number of measurements N, we can compute a corresponding p value.

The Wilcoxon signed rank test has utility in experimental designs having perfect match and mismatch probes. In fact, this a commonly used statistic for Affymetrix tiling microarray analysis.

Analysis of Variance (ANOVA)

Previous sections showed how to test for the differential spot intensities measured between two conditions. Frequently, however, a study consists of three or more conditions and the researcher would still like to deduce which genes differ in expression levels between the conditions under study. The standard statistical tool for solving such problems is the ANOVA.

To begin, we need a null hypothesis. For ANOVA, our null hypothesis will be that for all conditions, the gene under study has the same expression level. It may seem strange that a model for assessing equality of means is called analysis of variance. However, the basic idea of ANOVA is to compare the variance of within-condition means to the variance calculated within each condition. (The variance of within-condition means will hereafter be called the between condition variance, and the variance within the samples as the within condition variance.)

Consider measurements $X_{i,j}$ for a single gene where subscript i indicates that the measurement is from the ith biological condition being studied and

j denotes the jth measurement from this condition. If we symbolize the average intensity within condition i as $\bar{X}_i$ and the average of all measurements as $\bar{X}$, we can compute the between condition variance as

$$\sigma^2_{\text{between}} = \frac{\sum_{i=1}^{K} N_i(\bar{X}_i - \bar{X})^2}{K - 1} \tag{19}$$

where K is the number of conditions being studied, N is the total number of measurements, and N_i is the number of measurements taken for condition i. Note that if there are no differences among the conditions, then the variance of their means is small. Likewise, if there are differences the terms $(\bar{X}_i - \bar{X})^2$ become larger. We would like to compare this number to the amount of variation we expect if there are no differences. We can estimate this level of variation by calculating the within condition variance:

$$\sigma^2_{\text{within}} = \frac{\sum_{i,j}(X_{i,j} - \bar{X}_i)^2}{N - K}. \tag{20}$$

We can then compare these two variances [Eqs. (19) and (20)] via a ratio:

$$S_{\text{anova}} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}}. \tag{21}$$

Like previous statistics, we know how this statistic behaves under the null hypothesis of no differential expression and we use this information to calculate its corresponding p value.

The aforementioned discussion on ANOVA is intended to provide a basic feel of the technique and is useful in the case where just one factor (such as biological condition) is expected to affect the measured intensities. Clearly, it can easily be the case that several factors affect microarray measurements. As an example, let us assume that our microarray measurements are expected to vary due to two independent factors in a cancer study. First, we might expect to see differences based on which of several tissue types the measured mRNA came from. Example tissues might include "healthy tissue," "localized cancer," and "metastatic cancer." Second, we could also expect that expression measurements are affected by the race of the individual from which the tissue was obtained. The goal of the study is to identify whether the expression level of some gene changes among the healthy, localized, and metastatic samples.

Given the stated goal of the study, it is tempting to simply apply Eqs. (19), (20), and (21) to elucidate an answer. The problem with doing so is that $\sigma^2{}_{\text{within}}$ is large when there are unaccounted sources of variation. This translates into lower values of S_{anova} and higher p values.

Why would this higher $\sigma^2{}_{\text{within}}$ be the case? Recall that the two factors are independent. Therefore, when we bin data by a single factor (e.g., tissue), each bin contains a number of measurements from each class of the other factor (race). Now if there are differences among the classes of the second factor, this will lead to some spread within each tissue bin. This spread leads to higher values of $\sigma^2{}_{\text{within}}$. To give us the best chance of detecting a difference among the factor we care about, we need to do some additional work.

First, accounting for two sources of variation requires a little more notation. Previously, we used $X_{i,j}$ to indicate the jth measurement of the ith condition. Now because we have an additional source of variation we wish to model, we must extend this to the term $X_{i,j,k}$, which symbolizes the kth measurement of those belonging to both the ith class of one factor and the jth class of the second. For example, $X_{3,1,7}$ could symbolize the seventh measurement taken of those of the third tissue type (e.g., metastatic tissue) and the first race (e.g., African). In addition, we previously used the variable K to indicate the number of classes we were testing between. Now in addition to K, we also need a variable that denotes the number of classes of the other factor we are studying. Let this variable be B. In our example we might have $K = 3$ ("healthy," "localized," and "metastatic") and $B = 4$ ("African," "Asian," "Caucasian," and "Latino").

In studying the differences between the different stages of cancer, we calculate $\sigma^2{}_{\text{between}}$ as before using Eq. (19), where we use tissue labels as the different classes. The main difference in our analysis lies in how $\sigma^2{}_{\text{within}}$ is calculated. If we let $\bar{X}_{i,j}$ be the mean of all measurements where factor 1 (e.g., tissue type) is i and factor 2 (e.g., race) is j, then $\sigma^2{}_{\text{within}}$ is calculated as

$$\sigma^2_{\text{within}} = \frac{\sum^{i,j,k}(X_{i,j,k} - \bar{X}_{i,j})^2}{N - BK}. \tag{22}$$

We can then use Eq. (21) as before and use knowledge of its distribution under the null hypothesis to obtain a corresponding p value. Intuitively, all we have done in moving from one factor to two is to adjust the within condition variance so that it does not include potential variation from known sources such as age, race, or gender. Accounting for these sources of variation gives us an enhanced ability to detect differences between some conditions of interest. This increase in sensitivity comes at a cost, however. To calculate $\sigma^2{}_{\text{within}}$ accurately, there must be a number of measurements available for each combination of the factors we wish to model. As the number of factors in our model increases, so too does the number of replicate experiments needed to estimate $\sigma^2{}_{\text{within}}$.

Given that ANOVA can account for different sources of variability, it is also capable of merging microarray normalization with differential expression detection. To do this, sources of variation within the model are not only those of biological interest (such as cancerous vs healthy tissue), but also those of technical concern (such as microarray used and dye used for labeling) (Kerr *et al.*, 2000). The application of ANOVA to microarray data in this context is reviewed nicely in Kerr (2003).

Extensions to Tiling Microarrays

The tests described earlier can be applied to tiling microarrays as well. Recall that in a tiling microarray, we are looking for regions of consecutive probes (in genomic space) that exhibit intensities higher than some background level. To assess this, a windowing approach is often taken where we do not simply assess a single feature by itself, but rather we assess that feature along with a window of neighboring features. To apply the t statistic, for example, we may test the intensities of each window to a random sampling of intensities from any genomic region, to intensities from within putative promoters (which are not expected to be transcribed), or to a control set of features. For Affymetrix tiling microarrays that contain a mismatch probe for every perfect match probe on the microarray, the mismatch probes can serve as this control set to which the comparison can be made. The extension of this approach to fold change, SAM, etc. is straightforward.

Following scoring each window in this manner, the resulting statistics are thresholded by some criteria (set by negative and positive control probes or theoretical considerations). The result is a set of putatively "on" and "off" probes. Spots that meet the threshold criterion and that are within a short distance of each other in genomic space are combined (the spacing between probes above threshold must be less than *maxgap* bp apart) to form larger continuous regions. These combined fragments are then filtered to remove short fragments (require a length longer than *minrun* bp) that are likely to be spurious results.

Extensions to Protein Microarrays

For antibody microarrays that assess concentrations of proteins in solution, the methods described in this section can be applied directly to testing abundance differences between two or more biological conditions. For functional protein microarrays, however, the question is usually one of event detection. In these cases, control experiments must be designed so that they represent the activities of proteins in some baseline state. Once a suitable control is identified, then the methods described here are suitable as well.

Summary

The microarray platform is emerging as a standard tool in biological and biomedical research. This is partly because of its ever-expanding utility, as evidenced by both tiling and protein microarray applications. As is true for any standard tool, it is important that the microarray technology be well understood by its practitioners. For microarrays, part of this technological understanding is resident in the understanding of microarray statistics. Here, in this chapter, widely used methods for microarray normalization and significance testing are presented with the aim of providing this understanding in at least a broad sense. We have indicated where and when gene-based microarray statistics can be useful for tiling and protein microarrays in our discussion. The information conveyed was intended to provide at least a motivation and intuition for what happens to microarray data after it leaves the bench.

Acknowledgment

This work was supported by NIH Grant HG02357. Microsoft and Excel are registered trademarks in the United States and/or other countries.

References

Baldi, P., and Long, A. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**(6), 509–519.

Bertone, P., Gerstein, M., and Synder, M. (2005). Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res.* **13**(3), 259–274.

Bertone, P., Stolc, V., Royce, T., Rozowsky, J., Urban, A., Zhu, X., Rinn, J., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**(5705), 2242–2246.

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2), 185–193.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *In* "Studi in Onore del Professore Salvatore Ortu Carboni. Rome," pp. 13–60.

Cawley, S., Bekiranov, S., Ng, H., Kapranov, P., Sekinger, E., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T.

(2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**(4), 499–509.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D., and Gingeras, T. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**(5725), 1149–1154.

Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., and Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**(1), 65–73.

Choe, S., Boutros, M., Michelson, A., Church, G., and Halfon, M. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.* **6**(2), R16.

Cleveland, W. S. (1981). Lowess: A program for smoothing scatterplots by robust locally weighted regression. *Am. Stat.* **35,** 54.

Colantuoni, C., Henry, G., Zeger, S., and Pevsner, J. (2002). SNOMAD (Standardization and NOrmalization of MicroArray Data): Web-accessible gene expression data analysis. *Bioinformatics* **18**(11), 1540–1541.

Fodor, S., Read, J., Pirrung, M., Stryer, L., Lu, A., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**(4995), 767–773.

Hoffmann, R., Seidl, T., and Dugas, M. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.* **3**(7), RESEARCH0033.

Horak, C., and Snyder, M. (2002). ChIP-chip: A genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* **350,** 469–483.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2), 249–264.

Iyer, V., Horak, C., Scafe, C., Botstein, D., Snyder, M., and Brown, P. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**(6819), 533–538.

Kapranov, P., Cawley, S., Drenkow, J., Bekiranov, S., Strausberg, R., Fodor, S., and Gingeras, T. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296** (5569), 916–919.

Kerr, M. (2003). Linear models for microarray data analysis: Hidden similarities and differences. *J. Comput. Biol.* **10**(6), 891–901.

Kerr, M., Martin, M., and Churchill, G. (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **7**(6), 819–837.

Kluger, Y., Yu, H., Qian, J., and Gerstein, M. (2003). Relationship between gene co-expression and probe localization on microarray slides. *BMC Genom.* **4**(1), 49.

Lipshutz, R., Fodor, S., Gingeras, T., and Lockhart, D. (1999). High density synthetic oligonucleotide arrays. *Nature Genet.* **21**(Suppl. 1), 20–24.

Lueking, A., Horn, M., Eickhoff, H., Bussow, K., Lehrach, H., and Walter, G. (1999). Protein microarrays for gene expression and antibody screening. *Anal. Biochem.* **270**(1), 103–111.

Luscombe, N., Royce, T., Bertone, P., Echols, N., Horak, C., Chang, J., Snyder, M., and Gerstein, M. (2003). Express Yourself: A modular platform for processing and visualizing microarray data. *Nucleic Acids Res.* **31**(13), 3477–3482.

Nuwaysir, E., Huang, W., Albert, T., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J., Ballin, J., McCormick, M., Norton, J., Pollock, T., Sumwalt, T., Butcher, L., Porter, D., Molla, M., Hall, C., Blattner, F., Sussman, M., Wallace, R., Cerrina, F., and

Green, R. (2002). Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.* **12**(11), 1749–1755.

Qian, J., Kluger, Y., Yu, H., and Gerstein, M. (2003). Identification and correction of spurious spatial correlations in microarray data. *Biotechniques* **35**(1), 42–44, 46, 48.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genet.* **32** (Suppl), 496–501.

Rinn, J., Rozowsky, J., Laurenzi, I., Petersen, P., Zou, K., Zhong, W., Gerstein, M., and Snyder, M. (2004). Major molecular differences between mammalian sexes are involved in drug metabolism and renal function. *Dev. Cell* **6**(6), 791–800.

Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. (2003). TM4: A free, open-source system for microarray data management and analysis. *Biotechniques* **34**(2), 374–378.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235), 467–470.

Thomas, J., Olson, J., Tapscott, S., and Zhao, L. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.* **11**(7), 1227–1236.

Troyanskaya, O., Garber, M., Brown, P., Botstein, D., and Altman, R. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18**(11), 1454–1461.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**(9), 5116–5121.

Yang, I., Chen, E., Hasseman, J., Liang, W., Frank, B., Wang, S., Sharov, V., Saeed, A., White, J., Li, J., Lee, N., Yeatman, T., and Quackenbush, J. (2002a). Within the fold: Assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.* **3**(11), research0062.

Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., and Speed, T. (2002b). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**(4), e15.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R., Gerstein, M., and Snyder, M. (2001). Global analysis of protein activities using proteome chips. *Science* **293**(5537), 2101–2105.

Zhu, H., and Snyder, M. (2003). Protein chip technology. *Curr. Opin. Chem. Biol.* **7**(1), 55–63.