CHAPTER 4

Algorithmic Methods for the Analysis of Gene Expression Data

HONGBO XIE, UROS MIDIC, SLOBODAN VUCETIC, and
ZORAN OBRADOVIC

4.1 INTRODUCTION

The traditional approach to molecular biology consists of studying a small number of genes or proteins that are related to a single biochemical process or pathway. A major paradigm shift recently occurred with the introduction of gene expression microarrays that measure the expression levels of thousands of genes at once. These comprehensive snapshots of gene activity can be used to investigate metabolic pathways, identify drug targets, and improve disease diagnosis. However, the sheer amount of data obtained using the high throughput microarray experiments and the complexity of the existing relevant biological knowledge are beyond the scope of manual analysis. Thus, the bioinformatics algorithms that help to analyze such data are a very valuable tool for biomedical science. First, a brief overview of the microarray technology and concepts that are important for understanding the remaining sections are described. Second, microarray data preprocessing, an important topic that has drawn as much attention from the research community as the data analysis itself, is discussed second. Finally, some of the most important methods for microarray data analysis are described and illustrated with examples and case studies.

4.1.1 Biology Background

Most cells within the same living system have identical copies of DNA that store inherited genetic traits. DNA and RNA are the carriers of the genetic information. They are both polymers of nucleotides. There are four different types of nucleotides: adenine (A), thymine/uracil (T/U), guanine (G), and cytosine (C). Thymine is present in DNA, while uracil replaces it in RNA. Genes are fundamental blocks of DNA that encode genetic information and are transcribed into messenger RNA, or mRNA (hereafter noted simply as “RNA”). RNA sequences are then translated into proteins,

Handbook of Applied Algorithms: Solving Scientific, Engineering and Practical Problems
Edited by Amiya Nayak and Ivan Stojmenović

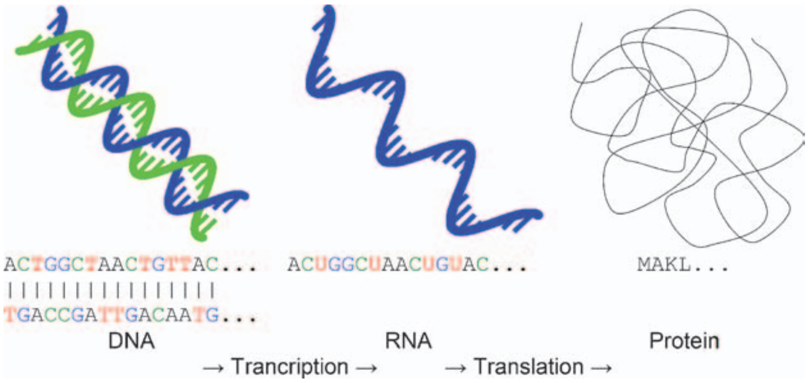

FIGURE 4.1 Central dogma of molecular biology: DNA–RNA–protein relationship.

which are the primary components of living systems and which regulate most of a cell's biological activities. Activities regulated and/or performed by a protein whose code is contained in the specific gene are also considered functions of that gene. For a gene, the abundance of the respective RNA in a cell (called the "expression level" for that gene) is assumed to correlate with the abundance of the protein into which the RNA translates. Therefore, the measurement of genes' expression levels elucidates the activities of the respective proteins. The relationship between DNA, RNA, and proteins is summarized in the *Central Dogma* of molecular biology as shown in Figure 4.1.

DNA consists of two helical strands; pairs of nucleotides from two strands are connected by hydrogen bonds, creating the so-called base pairs. Due to the chemical and steric properties of nucleotides, adenine can only form a base pair with thymine, while cytosine can only form a base pair with guanine. As a result, if one strand of DNA is identified, the other strand is completely determined. Similarly, the strand of RNA produced during the transcription of one strand of DNA is completely determined by that strand of DNA. The only difference is that uracil replaces thymine as a complement to adenine in RNA. Complementarity of nucleotide pairs is a very important biological feature. Preferential binding—the fact that nucleotide sequences only bind with their complementary nucleotide sequences—is the basis for the microarray technology.

4.1.2 Microarray Technology

Microarray technology evolved from older technologies that are used to measure the expression levels of a small number of genes at a time [1,2]. Microarrays contain a large number—hundreds or thousands—of small spots (hence the term "microarray"), each of them designed to measure the expression level of a single gene. Spots are made up of synthesized short nucleotide sequence segments called probes, which are attached to the chip surface (glass, plastic, or other material). Probes in each spot are designed to bind only to the RNA of a single gene through the

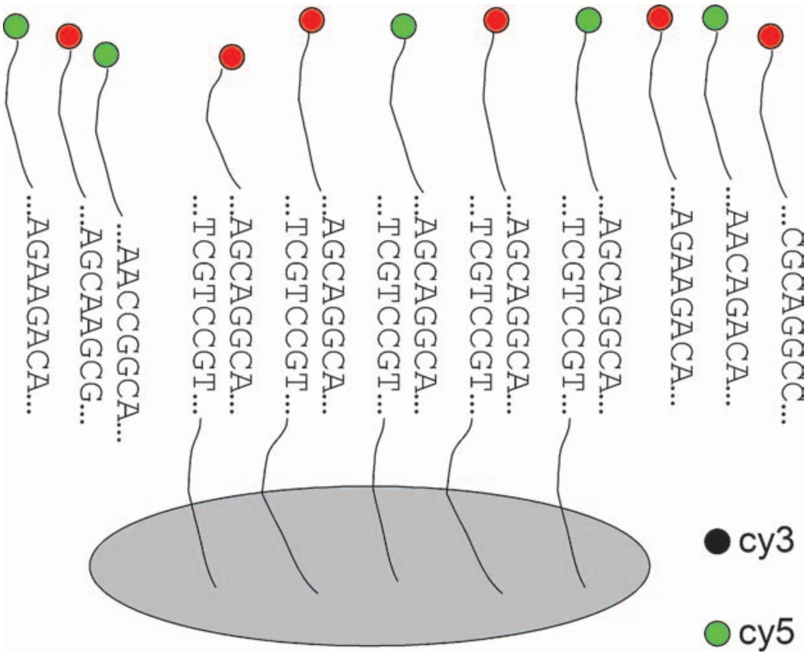

FIGURE 4.2 Binding of probes and nucleotide sequences. Probes in one spot are designed to bind only to one particular type of RNA sequences. This simplified drawing illustrates how only the complementary sequences bind to a probe, while other sequences do not bind to the probe.

principle of preferential binding of complementary nucleotide sequences, as illustrated in Figure 4.2. The higher the RNA expression level is for a particular gene, the more of its RNA will bind (or “hybridize”) to probes in the corresponding spot.

Single-channel and dual-channel microarrays are the two major types of gene expression microarrays. Single-channel microarrays measure the gene expression levels in a single sample and the readings are reported as absolute (positive) values. Dual-channel microarrays simultaneously measure the gene expression levels in two samples and the readings are reported as relative differences in the expression between the two samples. A sample (or two samples for dual-channel chips) and the microarray chip are processed with a specific laboratory procedure (the technical details of which are beyond the scope of this chapter). Part of the procedure is the attachment of a special fluorescent substrate to all RNA in a sample (this is called the “labeling”). When a finalized microarray chip is scanned with a laser, the substrate attached to sequences excites and emits light. For dual-channel chips, two types of substrates (cy3 and cy5) that emit light at two different wavelengths are used (Fig. 4.3). The intensity of light is proportional to the quantity of RNA bound to a spot, and this intensity correlates to the expression level of the corresponding gene.

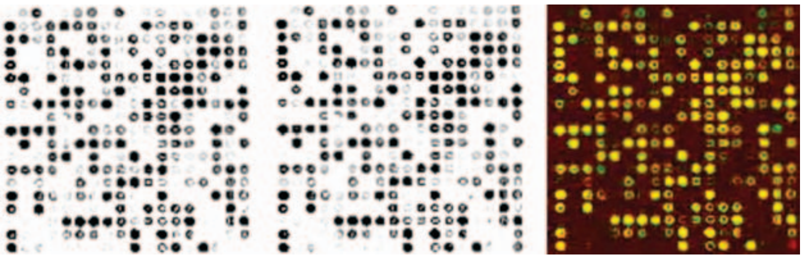

FIGURE 4.3 Dual-channel cDNA microarray. A sample of dual-channel microarray chip images, obtained from an image scanner. All images contain only a portion of the chip. From left to right: cy3 channel, cy5 channel, and the computer-generated joint image of cy3 and cy5 channels. A green spot in the joint image indicates that the intensity of the cy3 channel spot is higher than intensity of the cy5 channel spot, a red spot indicates a reverse situation, and a yellow spot indicates similar intensities.

Images obtained from scanning are processed with image processing software. This software transforms an image bitmap into a table of spot intensity levels accompanied by additional information such as estimated spot quality. The focus of this chapter is on the analysis of microarray data starting from this level. The next section describes methods for data preprocessing, including data cleaning, transformation, and normalization. Finally, the last section provides an overview of methods for microarray data analysis and illustrates how these methods are used for knowledge discovery. The overall process of microarray data acquisition and analysis is shown in Figure 4.4.

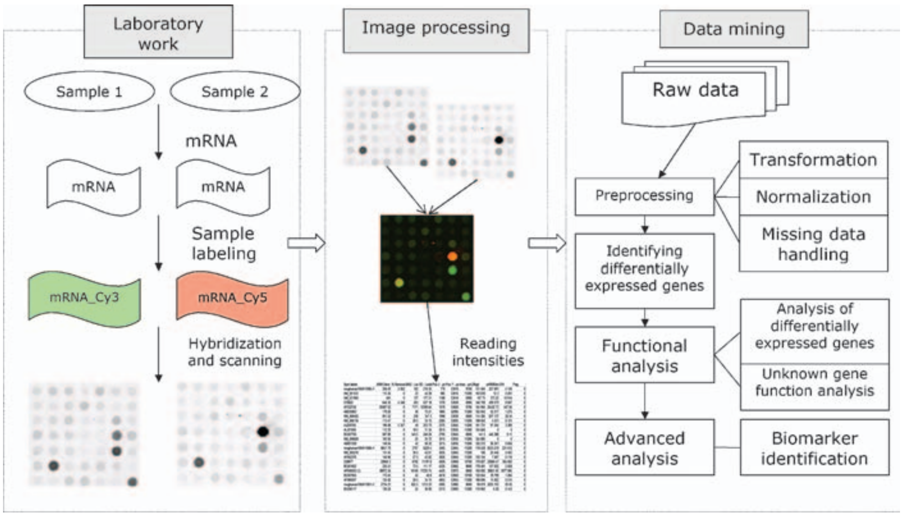

FIGURE 4.4 Data flow schema of microarray data analysis.

4.1.3 Microarray Data Sets

Microarray-based studies consider more than one sample and most often produce several replicates for each sample. The minimum requirement for a useful biological study is to have two samples that can be hybridized on a single dual-channel or on two single-channel microarray chips.

A data set for a single-channel microarray experiment can be described as an $M \times N$ matrix in which each column represents gene expression levels for one of the N chips (arrays), and each row is a vector containing expression levels of one of the M genes in different arrays (called "expression profile"). A data set for a dual-channel microarray experiment can be observed as a similar matrix in which each chip is represented by a single column of expression ratios between the two channels ($cy3$ and $cy5$), or by two columns of absolute expression values of the two channels. A typical microarray data table has a fairly small number of arrays and a large number of genes ($M \gg N$); for example, while microarrays can measure the expression of thousands of genes, the number of arrays is usually in the range from less than 10 (in small-scale studies) to several hundred (in large-scale studies).

Methods described in this chapter are demonstrated by case studies on acute leukemia, *Plasmodium falciparum* intraerythrocytic developmental cycle, and chronic fatigue syndrome microarray data sets. *Acute leukemia data set* [3] contains 7129 human genes with 47 arrays of acute lymphoblastic leukemia (ALL) samples and 25 arrays of acute myeloid leukemia (AML) samples. The data set is used to demonstrate a generic approach to separating two types of human acute leukemia (AML versus ALL) based on their gene expression patterns. This data set is available at http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43. *Plasmodium falciparum data set* [4] contains 46 arrays with samples taken during 48 h of intraerythrocytic developmental cycle of *Plasmodium falciparum* to provide the comprehensive overview of the timing of transcription throughout the cycle. Each array consists of 5080 spots, related to 3532 unique genes. This data set is available at http://biology.plosjournals.org/archive/1545-7885/1/1/supinfo/10.1371_journal.pbio.0000005.sd002.txt. *Chronic fatigue syndrome (CFS) data set* contains 79 arrays from 39 clinically identified CFS patients and 40 non-CFS (NF) patients [5]. Each chip measures expression levels of 20,160 genes. This data set was used as a benchmark at the 2006 Critical Assessment of Microarray Data Analysis (CAMDA) contest and is available at http://www.camda.duke.edu/camda06/datasets.

4.2 MICROARRAY DATA PREPROCESSING

Images obtained by scanning microarray chips are preprocessed to identify the spots, estimate their intensities, and flag the spots that cannot be read reliably. Data obtained from a scanner are usually very noisy; the use of raw unprocessed data would likely bias the study and possibly lead to false conclusions. In order to reduce these problems, several preprocessing steps are typically performed and are described in this section.

4.2.1 Data Cleaning and Transformation

4.2.1.1 Reduction of Background Noise in Microarray Images The background area outside of the spots in a scanned microarray image should ideally be dark (indicating no level of intensity), but in practice, the microarray image background has a certain level of intensity known as *background noise*. It is an indicator of the systematic error introduced by the laboratory procedure and microarray image scanning. This noise can often effectively be reduced by estimating and subtracting the mean background intensity from spot intensities. A straightforward approach that uses the mean background intensity of the whole chip is not appropriate when noise intensity is not uniform in all parts of the chip. In such situations, *local estimation* methods are used to estimate the background intensity individually for each spot from a small area surrounding the spot.

4.2.1.2 Identification of Low Quality Gene Spots Chip scratching, poor washing, bad hybridization, robot injection leaking, bad spot shape, and other reasons can result in microarray chips containing many damaged spots. Some of these gene spot problems are illustrated in Figure 4.5. Low quality gene spots are typically identified by comparing the spot signal and its background noise [6,7]. Although statistical techniques can provide a rough identification of problematic gene spots, it is important to carefully manually evaluate the microarray image to discover the source of the problem and to determine how to address problematic spots. The most simplistic method is to remove all data for the corresponding genes from further analysis. However, when the spots in question are the primary focus of the biological study, it is preferable to process microarray images using specialized procedures [8]. Unfortunately, such a process demands intensive manual and computational work. To reduce the data uncertainty due to damaged spots, it is sometimes necessary to repeat the hybridization of arrays with a large area or fraction of problematic spots.

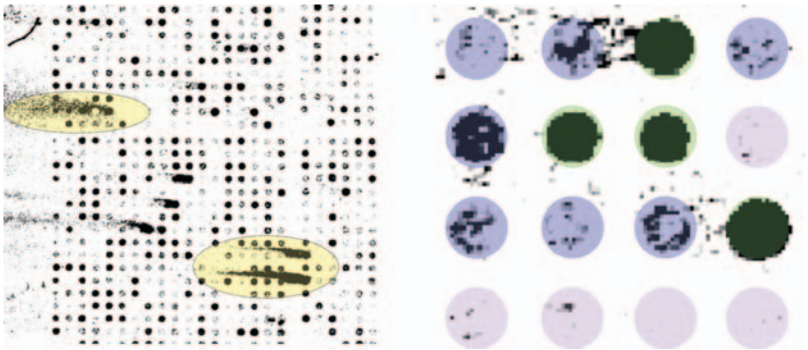

FIGURE 4.5 Examples of problematic spots. The yellow ovals in the left image are examples of poor washing and scratching. The green circle spots in the right image are good-quality spots. The pink circles indicate empty (missing) spots. The blue circles mark badly shaped spots.

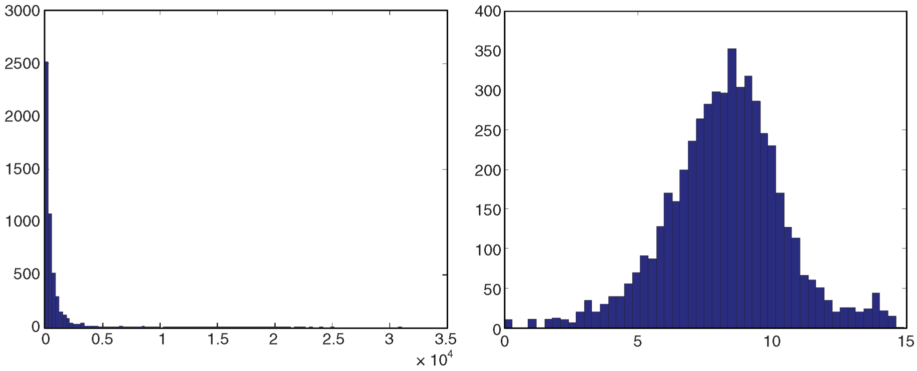

FIGURE 4.6 Data distribution before and after logarithmic transformation. Histograms show gene expression data distribution for patient sample #1 from acute lymphoblastic leukemia data set (X-axis represents the gene expression levels and Y-axis represents the amount of genes with given expression level). The distribution of raw data on the left is extremely skewed. The log-2 transformed data have a bell-shaped, approximately normal distribution, shown on the right.

4.2.1.3 Microarray Data Transformation After the numerical readings are obtained from the image, the objective of microarray data transformation is to identify outliers in the data and to adjust the data to meet the distribution assumptions implied by statistical analysis methods. A simple *logarithmic transformation* illustrated in Figure 4.6 is commonly used. It reshapes the data distribution into a bell shape that resembles normal distribution. This transformation is especially beneficial for data from dual-channel arrays, since data from these arrays are often expressed as ratios of signal intensities of pairs of samples. Alternative transformations used in practice include *arcsinh* function, *linlog* transformation, *curve-fitting* transformations, and *shift* transformation [9]; among them, the *linlog* transformation was demonstrated to be the most beneficial.

4.2.2 Handling Missing Values

Typical data sets generated by microarray experiments contain large fractions of missing values caused by low quality spots. Techniques for handling missing values have to be chosen carefully, since they involve certain assumptions. When these assumptions are not correct, artifacts can be added into the data set that may substantially bias the evaluation of biological hypotheses.

The straightforward approach is to completely discard genes with at least one missing value. However, if a large fraction of genes are eliminated because of missing values, then this approach is not appropriate.

A straightforward imputation method consists of replacing all missing values for a given gene with the mean of its valid expression values among all available arrays. This assumes that the data for estimating the most probable value of a missing gene expression were derived under similar biological conditions; for instance, they could be derived from replicate arrays. Most microarray experiments lack replicates due to the experimental costs. When there are no replicates available, a better choice for imputation is to replace all of the missing data in an array with the average of valid expression values within the array.

The *k-nearest-neighbor based method (KNN)* does not demand experimental replicates. Given a gene with missing gene expression readings, k genes with the most similar expression patterns (i.e., its k neighbors) are found. The given gene's missing values are imputed as the average expression values of its k neighbors [10], or predicted with the *local least squares (LLS)* method [11]. Recent research has demonstrated that the weighted nearest-neighbors imputation method (WeNNI), in which both spot quality and correlations between genes were used in the imputation, is more effective than the traditional KNN method [12].

Domain knowledge can help estimate missing values based on the assumption that genes with similar biological functions have similar expression patterns. Therefore, a missing value for a given gene can be estimated by evaluating the expression values of all genes that have the same or similar functions [13]. Although such an approach is reasonable in terms of biology, its applicability is limited when the function is unknown for a large number of the genes.

In addition to the problems that are related to poor sample preparation, such as chip scratching or poor washing, a major source of problematic gene spots is relatively low signal intensity compared to background noise. It is important to check the reasons for low signal intensity. Gene expression might be very low, for instance, if the biological condition successfully blocks the gene expression. In this case, the low gene expression signal intensity is correct and the imputation of values estimated by the above-mentioned methods would probably produce a value that is too high. An alternative is to replace such missing data with the lowest obtained intensity value within the same chip or with an arbitrary small number.

4.2.3 Normalization

Microarray experiments are prone to systematic errors that cause changes in the data distribution and make statistical inference unreliable. The objective of normalization is to eliminate the variation in data caused by errors of the experimental methods, making further analysis based only on the real variation in gene expression levels. All normalization methods may introduce artifacts and should be used with care. Most methods are sensitive to outliers, so outlier removal is crucial for the success of normalization.

There are two major types of normalization methods: *within-chip normalization* uses only the data within the same chip and is performed individually on each chip, while *between-chip normalization* involves microarray data from all chips simultaneously. Reviews on microarray data normalization methods are provided in [14–16].

4.2.3.1 Within-Chip Normalization Several within-chip normalization methods are based on linear transformations of the form $new\_value = (original\_value-$

a)/*b*, where parameters *a* and *b* are fixed for one chip. *Standardization normalization* assumes that the gene expression levels in one chip follow the standard normal distribution. Parameter *a* is set to the mean, while parameter *b* is set to the standard deviation of gene expression levels in a chip. This method can be applied to both dual-channel and single-channel microarray data.

Linear regression normalization [15] is another linear transformation that uses a different way to choose parameters *a* and *b*. The basic assumption for dual-channel arrays is that for a majority of genes, the intensity for the cy3 channel is similar to intensity for the cy5 channel. As a result, the two intensities should be highly correlated, and the fitted regression line should be very close to the main diagonal of the scatterplot. Parameters *a* and *b* in linear transformation are chosen so that the regression line for transformed data points aligns with the main diagonal.

A more advanced normalization alternative is the *loess transformation*. It uses a scatterplot of log ratio of two channel intensities ($\log(cy3/cy5)$) against average value of two channel intensities ($(cy3 + cy5)/2$). A locally weighted polynomial regression is used on this scatterplot to form a smooth regression curve. Original data are then transformed using the obtained regression curve. Loess normalization can also be used with single-channel microarrays where two arrays are observed as two channels and normalized together. For data from more than two arrays, loess normalization can be iteratively applied on all distinct pairs of arrays, but this process has larger computational cost. Some other forms of loess normalization are *local loess* [17], *global loess*, and *two-dimensional loess* [18].

Several normalization methods make use of domain knowledge. All organisms have a subset of genes—called housekeeping genes—that maintain necessary cell activities, and, as a result, their expression levels are nearly constant under most biological conditions. All the above-mentioned methods can be modified so that all transformation parameters are calculated based only on the expression levels of housekeeping genes.

4.2.3.2 Between-Chip Normalization

Row–column normalization [19] is applied to a data set comprised of several arrays, observed as a matrix with M rows (representing genes) and N columns (representing separate arrays and array channels). In one iteration, the mean value of a selected row (or column) is subtracted from all of the elements in that row (or column). This is iteratively repeated for all rows and columns of the matrix, until the mean values of all rows and columns approach zero. This method fixes variability among both genes and arrays. A major problem with this method is its sensitivity to outliers, a problem that can significantly increase computation time. Outlier removal is thus crucial for the performance of this method. The computation time can also be improved if standardization is first applied to all individual arrays.

Distribution (quantile) normalization [20] is based on the idea that a quantile–quantile plot is a straight diagonal line if two sample vectors come from the same distribution. Data samples can be forced to have the same distribution by projecting data points onto the diagonal line. For microarray data matrix with m rows and n columns, each column is separately sorted in descending order, and the mean values are calculated for all rows in the new matrix. Each value in the original matrix is then replaced with the mean value of the row in the sorted matrix where that value was placed during sorting. Distribution normalization may improve the reliability of statistical inference. However, it may also introduce artifacts; after normalization, low intensity genes may have the same (very low) intensity across all arrays.

Statistical model-fitting normalization involves the fitting of gene expression level data using a statistical model. The fitting residues can then be treated as bias-free transformation of expression data. For example, for a given microarray data set with genes g ($g = 1, \ldots, n$), biological conditions $T_i (i = 1, \ldots, m)$, and arrays $A_j (j = 1, \ldots, k)$, the intensity I of gene g at biological condition i and array j can be fitted using a model [21]

$$I_{gij} = u + T_i + A_j + (TA)_{ij} + \varepsilon_{gij}.$$

The fitting residues ε_{gij} for this model can be treated as bias-free data for gene g at biological condition i and array j after normalization.

In experiments with dual-channel arrays, it is possible to distribute (possibly multiple) samples representing m biological conditions over k arrays in many different ways. Many statistical models have recently been proposed for model-fitting normalization [22,23]. The normalization approaches of this type have been demonstrated to be very effective in many applications, especially in the identification of differentially expressed genes [21,24].

4.2.4 Data Summary Report

The data summary report is used to examine preprocessed data in order to find and correct inconsistencies in the data that can reduce the validity of statistical inference. Unlike other procedures, there are no golden standards for this step. It is a good practice to evaluate the data summary report before and after data preprocessing. Approaches used to inspect the data include the evaluation of a histogram to provide information about data distribution in one microarray, a boxplot of the whole data set to check the similarities of all data distributions, and the evaluation of correlation coefficient maps (see Fig. 4.7) to check consistency among arrays. Correlation coefficient heat maps plot the values of correlation coefficients between pairs of arrays. For a given pair of arrays, #i and #j, their expression profiles are observed as vectors and the correlation coefficient between the two vectors is plotted as two pixels—in symmetrical positions (*ij*) and (*ji*)—in the heat map (the magnitude of correlation coefficient is indicated by the color of the pixel). Correlation coefficients are normally expected to be high, since we assume that the majority of gene expression levels are similar in different arrays. A horizontal (and the corresponding vertical) line in a heat map represents all of the correlation coefficients between a given array and all other arrays. If a line has a near-constant color representing a very low value, we should suspect a problem with the corresponding array.

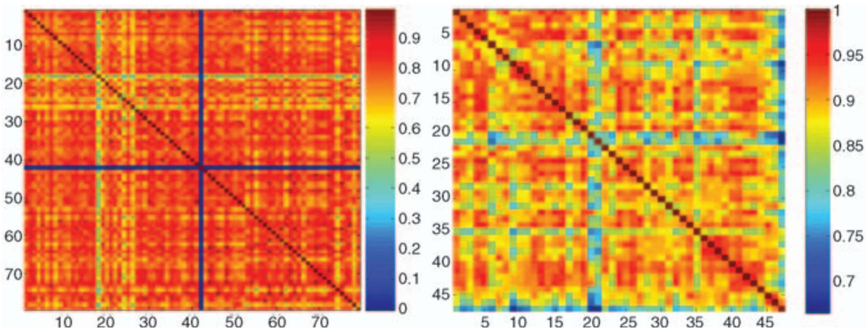

FIGURE 4.7 Correlation coefficient heat maps. The left heat map shows the correlation coefficients among the 79 samples of the CFS data set. The first 40 samples are from the nonfatigue (control) group. The remaining 39 samples are from the group of CFS patients. The color of a pixel represents the magnitude of the correlation coefficient (as shown in the color bar on the right). The correlation coefficients on the diagonal line are 1, since they compare each sample to itself. There are two clearly visible blue lines in the heat map on the left, corresponding to the sample #42. This indicates that this sample is different from the others; its correlation coefficients with all other samples are near zero. Therefore, we need to inspect this sample's chip image. Another sample that draws our attention is sample #18, which also has near-uniform correlation coefficients (around 0.5) with other samples. After inspecting the sample's chip image, we found that these correlation coefficients reflected sample variation and that we should not exclude sample #18 from our study. A similar heat map on the right shows the correlation coefficients among the 47 ALL samples from the acute leukemia data set. Overall data consistency is fairly high with an average correlation coefficient over 0.89.

4.3 MICROARRAY DATA ANALYSIS

This section provides a brief outline of methods for the analysis of preprocessed microarray data that include the identification of differentially expressed genes, discovery of gene expression patterns, characterization of gene functions, pathways analysis, and discovery of diagnostic biomarkers. All methods described in this section assume that the data have been preprocessed; see Section 4.2 for more details on microarray data preprocessing methods.

4.3.1 Identification of Differentially Expressed Genes

A gene is differentially expressed if its expression level differs significantly for two or more biological conditions. A straightforward approach for the identification of differentially expressed genes is based on the selection of genes with absolute values of log-2 ratio of expression levels larger than a prespecified threshold (such as 1). This simple approach does not require replicates, but is subject to high error rate (both false positive and false negative) due to the large variability in microarray data.

More reliable identification is possible by using statistical tests. However, these methods typically assume that the gene expression data follow a certain distribution, and require sufficiently large sample size that often cannot be achieved due to microarray experimental conditions or budget constraints. Alternative techniques, such as bootstrapping, impose less rigorous requirements on the sample size and distribution while still providing reliable identification of differentially expressed genes.

Given the data, a statistical test explores whether a *null hypothesis* is valid and calculates the *p*-value, which refers to the probability that the observed statistics are generated by the null model. If the *p*-value is smaller than some fixed threshold (e.g., 0.05), the null hypothesis is rejected. If the *p*-value is above the threshold, however, it should not be concluded that the original hypothesis is confirmed; the result of the test is that the observed events do not provide a reason to overturn it [25]. The most common null hypothesis in microarray data analysis is that there is no difference between two groups of expression values for a given gene. In this section, we briefly introduce the assumptions and requirements for several statistical tests that are often used for the identification of differentially expressed genes.

4.3.1.1 Parametric Statistical Approaches The *Student's t-test* examines the null hypothesis that the means of distributions from which two samples are obtained are equal. The assumptions required for *t*-test are that the two distributions are normal and that their variances are equal. The null hypothesis is rejected if the *p*-value for the *t*-statistics is below some fixed threshold (e.g., 0.05). The *t*-test is used in microarray data analysis to test—for each individual gene—the equality of the means of expression levels under two different biological conditions. Genes for which a *t*-test rejects the null hypothesis are considered differentially expressed.

The *t*-test has two forms: *dependent sample t-test* and *independent sample t-test*. *Dependent sample t-test* assumes that each member in one sample is related to a specific member of the other sample; for example, this test can be used to evaluate the drug effects by comparing the gene expression levels of a group of patients before and after they are given a certain type of drug. *Independent sample t-test* is used when the samples are independent of each other; for example, this test can be used to evaluate the drug effects by comparing gene expression levels for a group of patients treated with the drug to the gene expression levels of another group of patients treated with a placebo. The problem with using the *t*-test in microarray data analysis is that the distribution normality requirement is often violated in microarray data.

One-way analysis of variance (ANOVA) is a generalization of the *t*-test to samples from more than two distributions. ANOVA also requires that the observed distributions are normal and that their variances are approximately equal. ANOVA is used in microarray data analysis when gene expression levels are compared under two or more biological conditions, such as for a comparison of gene expression levels for a group of patients treated with drug A, a group of patients treated with drug B, and a group of patients treated with placebo.

The *volcano plot* (see Fig. 4.8) is often used in practice for the identification of differentially expressed genes; in this case, it is required that a gene both

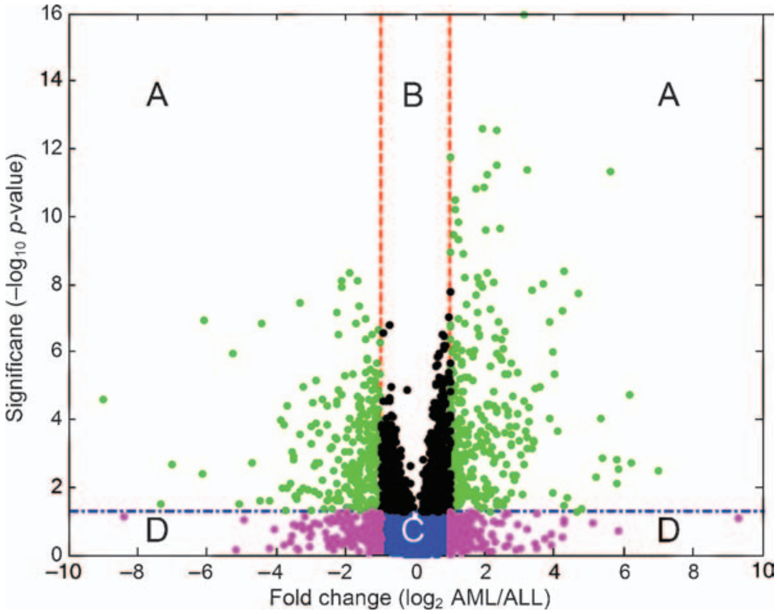

FIGURE 4.8 The volcano plot of significance versus fold change. This figure is a plot of the significance (*p*-value from ANOVA test, on a –log-10 scale) against fold change (log-2 ratio), for testing the hypothesis on the differences in gene expression levels between the AML group and the ALL group in the acute leukemia data set. The dark blue horizontal line represents a significance level threshold of 0.05. The two vertical lines represent the absolute fold-change threshold of 2. The genes plotted in the two "A" regions are detected as significant by both methods, while the genes plotted in region "C" are detected as insignificant by both methods. This type of plot demonstrates two types of errors that occur with the ratio-based method: false positive errors plotted in the two "D" regions, and false negative errors plotted in the "B" region. A common practice is to identify only the genes plotted in the two "A" regions as differentially expressed and discard the genes plotted in the "B" region.

passes the significance test and that its expression level log ratio is above the threshold.

4.3.1.2 *Nonparametric Statistical Approaches* Nonparametric tests relax the assumptions posed by the parametric tests. Two popular nonparametric tests are the Wilcoxon rank-sum test for equal median and the Kruskal–Wallis nonparametric one-way analysis of variance test.

The Wilcoxon rank-sum test (also known as Mann–Whitney *U*-test) tests the hypothesis that two independent samples come from distributions with equal medians. This is a nonparametric version of the *t*-test. It replaces real data values with their sorted ranks and uses the sum of ranks to obtain a *p*-value. Kruskal–Wallis test compares the medians of the samples. It is a nonparametric version of the one-way ANOVA, and an extension of the Wilcoxon rank-sum test to more than two groups.

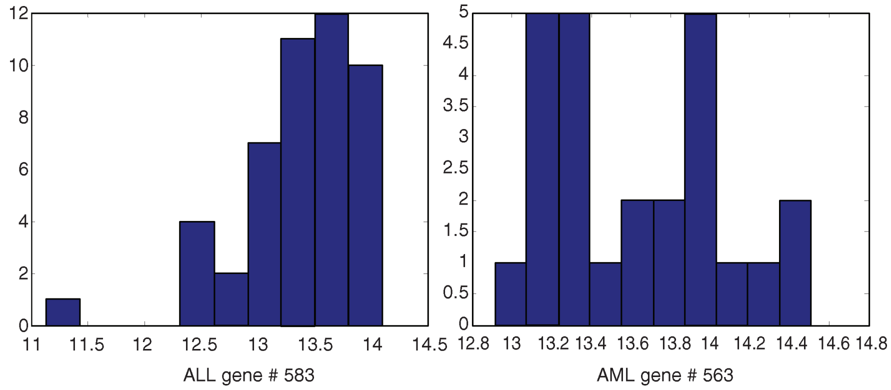

FIGURE 4.9 Importance of data distribution type for the choice of statistical test. Two histograms show the distribution of expression levels for gene #563 in two groups of samples in the acute leukemia data set: ALL on the left and AML on the right. The two distributions are clearly different. When testing the equality of means of two groups, the Kruskal–Wallis test gives us the p-value of 0.16, and the ANOVA test gives us the p-value of 0.05. Since the data distribution in the right panel has two major peaks, it is not close to normal distribution; therefore, it is preferable to choose the Kruskal–Wallis test.

Nonparametric tests tend to reject less null hypotheses than the related parametric tests and have lower sensitivity, which leads to an increased rate of false negative errors. They are more appropriate when the assumptions for parametric tests are not satisfied, as is often the case with microarray data (see Fig. 4.9). However, this does not imply that nonparametric tests will necessarily identify a smaller number of genes as differentially expressed than the parametric test, or that the sets of genes identified by one parametric test and one nonparametric test will necessarily be in a subset relationship. To illustrate the difference in results we used both ANOVA and the Kruskal–Wallis test to identify differentially expressed genes in the acute leukemia data set. Out of 7129 genes, 1030 genes were identified as differentially expressed by both methods. In addition to that, 155 genes were identified only by ANOVA, while 210 genes were identified only by the Kruskal–Wallis test.

4.3.1.3 Advanced Statistical Models Recently, more sophisticated models and methods for the identification of differentially expressed genes have been proposed [26,27]. For example, when considering the factors of array (A), gene (G), and biological condition (T), a *two-step mix-model* [21] first fits the variance of arrays, biological conditions, and interactions between arrays and biological conditions using one model, and then uses the residues from fitting the first model to fit the second model. An overview of mix-model methods is provided in the work by Wolfinger et al. [28]. Other advanced statistical approaches with demonstrated good results in identifying differentially expressed genes include the significance analysis of microarray (SAM) [29], regression model approaches [30], empirical Bayes analysis [31], and the bootstrap approach to gene selection (see the case study below).

Case Study 4.1: Bootstrapping Procedure for Identification of Differentially Expressed Genes

We illustrate the bootstrapping procedure for the identification of differentially expressed genes on an acute leukemia data set. The objective is to identify the genes that are differentially expressed between 47 ALL and 25 AML arrays. For each gene, we first calculate the p-value p_0 of two-sample t-test on the gene's expression levels in AML group versus ALL group. Next, the set of samples is randomly split into two subsets with 47 and 25 elements, and a similar t-test is performed with these random subsets and p-value p_1 is obtained. This step is repeated a large number of times ($n > 1000$), and as a result we obtain p-values $p_1, p_2, p_3, \ldots, p_n$. These p-values are then compared to the original p_0. We define the bootstrap p-value as $p_b = c/n$, where c is the number of times when values $p_i (i = 1, \ldots, n)$ are smaller than p_0. If p_b is smaller than some threshold (e.g., 0.05), then we consider the gene to be differentially expressed.
For the 88th gene in the data set, the expression levels are

ALL
759, 1656, 1130, 1062,
822, 1020, 1068, 1455,
1099, 1164, 662, 753,
728, 918, 943, 644,
2703, 916, 677, 1251,
138, 1557, 750, 814,
667, 616, 1187, 1214,
1080, 1053, 674, 708,
1260, 1051, 1747, 1320,
730, 825, 1072, 774,
690, 1119, 866, 564,
958, 1377, 1357

AML
1801, 1024, 3084, 1974,
1084, 1090, 908, 2474,
1635, 1591, 1323, 857,
1872, 1593, 1981, 2668,
1128, 3601, 2153, 1603,
769, 893, 2513, 2903,
2147

The p-value of the t-test for this gene is $p_0 = 3.4\text{E} - 007$, which is smaller than the threshold 0.05. The distribution of p-values obtained on randomly selected subsets $(p_1, \ldots, p_{1000})$ is shown in Figure 4.10. The bootstrap p-value is $p_b = 0$, so the bootstrapping procedure confirms the result of the t-test, that is, the 88th gene is differentially expressed.

4.3.1.4 False Discovery Rate (FDR) Control

Statistical procedures for the identification of differentially expressed genes can be treated as multiple hypothesis testing. A p-value threshold that is appropriate for a single test does not provide good control on false positive discovery for the overall procedure. For example, testing of 10,000 genes with p-value threshold of 0.05 is expected to identify $10,000 \times 0.05 = 500$ genes as differentially expressed even if none of the genes are actually differentially expressed. The false positive rate can be controlled by evaluating the expected proportion of true rejected null hypotheses out of the total number of rejected null hypothesis. An example of FDR control is shown in Figure 4.11.

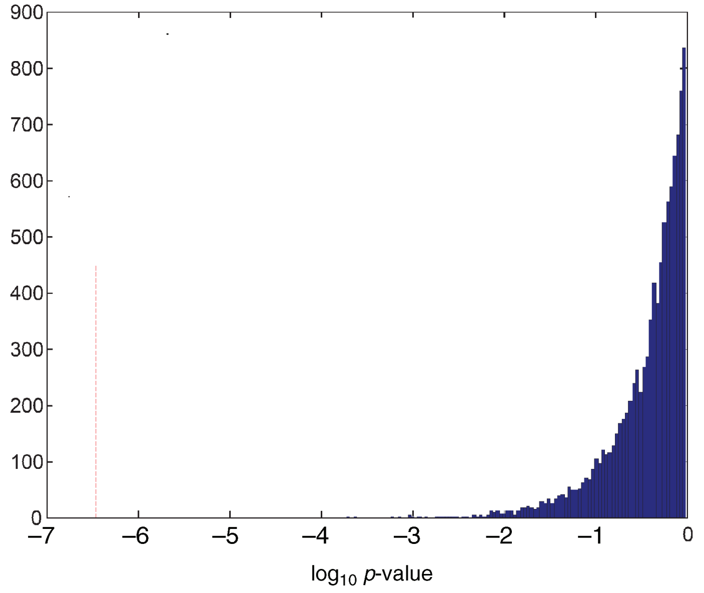

FIGURE 4.10 [Q1]

If N is the total number of genes, α_0 is the p-value threshold, and $p_i (i = 1, \ldots, N)$ are p-values in ascending order, then the ith ranked gene is selected if $p_i \leq \alpha_0 \cdot i/N$ [32]. A comprehensive review of this statistical FDR control is presented in the work by Qian and Huang [33]. It is worth noting that a bootstrap procedure for FDR control has also been introduced [29] and was shown to be suitable for gene selecting when data distribution deviates from normal distribution.

4.3.2 Functional Annotation of Genes

One of the goals of microarray data analysis is to aid in discovering biological functions of genes. One of the most important sources of domain knowledge on gene functions is Gene Ontology (GO), developed and maintained by the Gene Ontology Consortium [34,35]. Using a controlled and limited vocabulary of terms describing gene functions, each term in Gene Ontology consists of a unique identifier, a name, and a definition that describes its biological characteristic. GO terms are split into three major groups: biological processes, molecular functions, and cellular component categories. Within each category, GO terms are organized in a direct acyclic graph (DAG) structure, where each term is a node in the DAG, and each node can have several child and parent nodes. The GO hierarchy is organized with a general-to-specific relation between higher and lower level GO terms (see Fig. 4.12).

Sometimes, it is useful to compare several GO terms and determine if they are similar. Although there is no commonly accepted similarity measure between different GO terms, various distance measures were proposed for measuring the similarity between GO terms [36,37]. For example, the distance between nodes X and Y in a

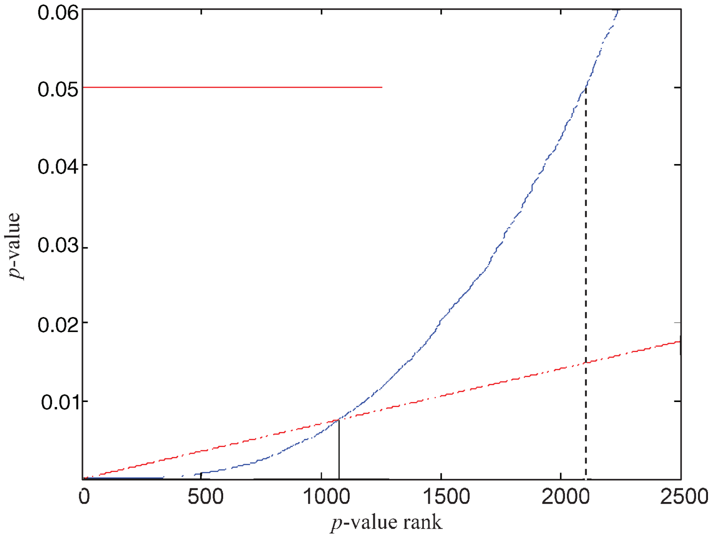

FIGURE 4.11 Benjamini–Hochberg FDR control. This figure compares the use of constant p-value threshold (in this case 0.05) and the use of Benjamini–Hochberg (BH) FDR control method for the two-sample t-test on acute leukemia data set. The blue curve is the plot of the original p-values obtained from the t-tests for individual genes, sorted in an increasing order. The horizontal red line represents the constant p-value threshold of 0.05. There are 2106 genes with a p-value smaller than this threshold. The slanted red line represents the p-value thresholds $p_i = \alpha_0 \cdot i/N$ that BH method uses to control the FDR at level of $\alpha_0 = 0.05$ (N is the total number of genes). It intersects with the blue curve at p-value 0.0075. Only the 1071 genes whose p-values are smaller than 0.0075 are considered to be significantly differentially expressed. The remaining 935 genes are considered to be false positive discoveries made by individual t-tests.

DAG can be measured as the length of the shortest path between X and Y within the GO hierarchy normalized by the length of maximal chain from the top to the bottom of the DAG [38]. One possible modification, illustrated in Figure 4.12, is to add a large penalty for paths that cross the root of a DAG to account for unrelated terms.

4.3.3 Characterizing Functions of Differentially Expressed Genes

After identifying differentially expressed genes, the next step in analysis is often to explore the functional properties of these genes. This information can be extremely useful to domain scientists for the understanding of biological properties of different sample groups. Commonly used methods for such analysis are described in this section. The *chi-square* and the *Fisher's exact* tests are used to test whether the selected genes are overannotated with a GO term F, as compared to the set of remaining genes spotted on a microarray [39,40]. For instance, the following 2 × 2 contingency table contains the data that can be used to test whether the frequency of genes annotated with a GO term F among the selected genes is different than the same frequency

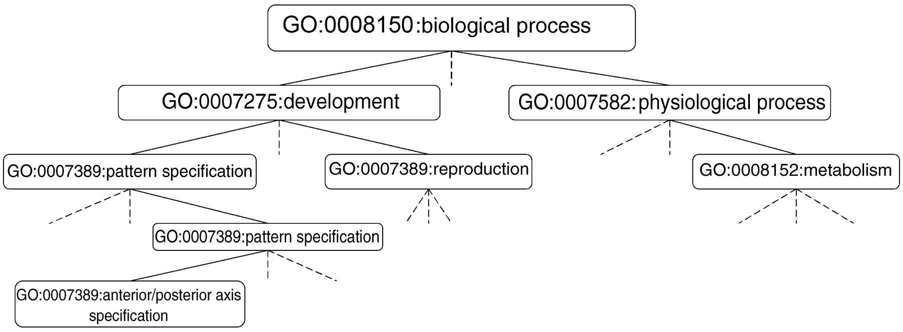
GO:0008150:biological process
GO:0007275:development
GO:0007582:physiological process
GO:0007389:pattern specification
GO:0007389:reproduction
GO:0008152:metabolism
GO:0007389:pattern specification
GO:0007389:anterior/posterior axis specification

FIGURE 4.12 Part of the Gene Ontology direct acyclic graph. The shortest path between GO:0007275:*development* and GO:0009948:*anterior/posterior axis specification* is 3 (the nearest common ancestor for the two terms is GO:0007275:*development*). The shortest path between the terms GO:0007275:*development* and GO:0008152:*metabolism* is 3 but the only ancestor for them is GO:0008150:*biological processes*, so the distance between them is 3 + 23, where 23 is the added penalty distance, which is the maximum distance in Biological Process part of Gene Ontology DAG.

among the remaining genes:

	Number of genes		
	Selected genes	Remaining genes	Total
Annotated with a GO term F	f_{11}	f_{12}	r_1
Not annotated with a GO term F	f_{21}	f_{22}	r_2
Total	c_1	c_2	S

Chi-square test uses a χ^2 statistic with formula

$$\chi^2 = \sum_{i=1}^{2}\sum_{i=1}^{2} \frac{(f_{ij} - r_i c_j/S)^2}{r_i c_j/S}.$$

The chi-square test is not suitable when any of the expected values $r_i c_j/S$ are smaller than 10. *Fisher's exact* test is more appropriate in such cases. In practice, all genes annotated with term F and all terms in the subtree of term F are considered to be annotated with F.

4.3.4 Functional Annotation of Uncharacterized Genes

The functional characterization of genes involves a considerable amount of biological laboratory work. Therefore, only a small fraction of known genes and proteins is functionally characterized. An important microarray application is the prediction of gene functions in a cost-effective manner. Numerous approaches use microarray gene expression patterns to identify unknown gene functions [41–43]. In the following section, we outline some of the most promising ones.

4.3.4.1 Unsupervised Methods for Functional Annotation

Gene expression profiles can be used to measure distances among genes. The basic assumption in functional annotation is that genes with similar biological functions are likely to have similar expression profiles. The functions of a given gene could be inferred by considering the known functions of genes with similar expression profiles. A similar approach is to group all gene expression profiles using clustering methods and to find the overrepresented functions within each cluster [44,45]. Then, all genes within a cluster are annotated with the overrepresented functions of that cluster. An alternative is to first cluster only the genes with known functions. An averaged expression profile of all genes within the cluster can then be used as the representative of a cluster [4]. The gene with the unknown function can be assigned functions based on its distance to the representative expression profiles. Conclusions from these procedures are often unreliable: a gene may have multiple functions that may be quite distinctive; also, genes with the same function can have quite different expression profiles. Therefore, it is often very difficult to select representative functions from a cluster of genes.

Many unsupervised methods for functional annotation face the issue of model selection in clustering, such as choosing the proper number of clusters, so that the genes within the cluster have similar functions. Domain knowledge is often very helpful in the model selection [46].

As we already mentioned, nearest-neighbor and clustering methods for assigning functions to genes are based on assumptions that genes with similar functions will have similar expression profiles [47]. However, this assumption is violated for more than half of the GO terms [48]. A more appropriate approach, therefore, is to first determine a subset of GO terms for which the assumption is valid, and use only these GO terms in gene function annotation.

Case Study 4.2: Identification of GO Terms with Conserved Expression Profiles

We applied a bootstrapping procedure to identify GO terms that have conserved gene expression profiles in the *Plasmodium* data set that contains 46 arrays. Each of the 46 arrays in the *Plasmodium* data set measures expression levels of 3532 genes at a specific time point over the 48-h *Plasmodium falciparum* intraerythrocytic developmental cycle (IDC). The bootstrap procedure was applied to 884 GO terms that are associated with at least two genes. For a given GO term with l associated genes, we collected their expression profiles and calculated the average pairwise correlation coefficients ρ_0. We compared ρ_0 to average expression profile correlation coefficients of randomly selected pairs of genes. In each step of the bootstrap procedure, we randomly selected l genes and computed their average correlation coefficient ρ_i. This was repeated 10,000 times to obtain $\rho_1, \rho_2, \ldots, \rho_{10,000}$. We counted the number c of ρ_i that are greater than ρ_0 and calculated the bootstrap p-value as $p_b = c/n$. If p_b is smaller than 0.05, the expression profiles of the GO term are considered to be conserved.

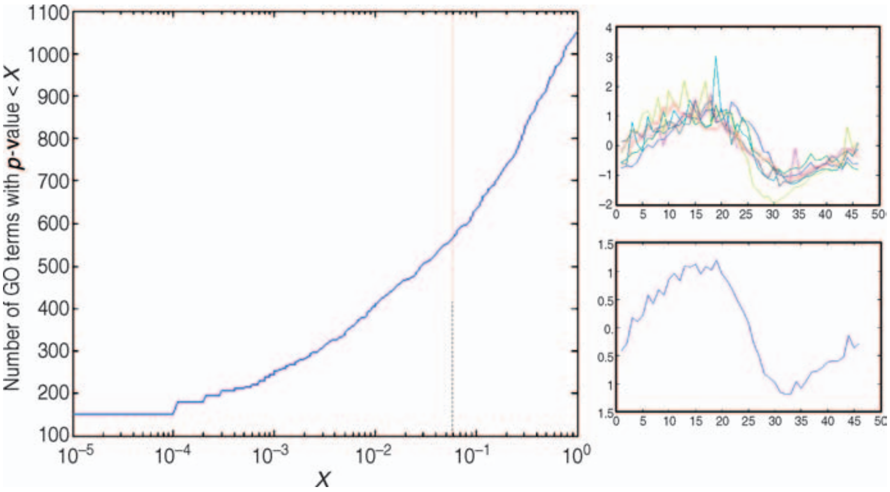

FIGURE 4.13 [Q2]

The plot in the left part of Figure 4.13 shows the cumulative number of GO terms with p-value smaller than $\mathbf{x}$. Four hundred and twenty-eight (48.4 percent) of the 884 GO terms have p-value smaller than 0.05; 199 of these are molecular function and 229 are biological process GO terms. This result validates to a large extent the hypothesis that genes with identical functions have similar expression profiles. However, it also reveals that for a given microarray experiment, a large fraction of functions do not follow this hypothesis.

Figure 4.13 also contains expression profiles of genes annotated with GO term GO:0006206 (pyrimidine base metabolism; bootstrap p-value 0) and its representative expression profile.

4.3.4.2 Supervised Methods for Functional Annotation Supervised methods for functional characterization involve building classification models that predict gene functions based on gene expression profiles. A predictor for a given function is trained to predict whether a given gene has that function or not [49]. Such a predictor is trained and tested on a collection of genes with known functions. If testing shows that the accuracy of the predictor is significantly higher than that for a trivial predictor, the predictor can then be used on the uncharacterized genes to annotate them. Previous research shows that the support-vector machines (SVM) model achieves the best overall accuracy when compared to other competing prediction methods [50]. The SVM-based predictor can overcome some of the difficulties that are present with the unsupervised methods. It can flexibly select the expression profile similarity measure and handle a large feature space. The unresolved problem of the supervised approach is the presence of multiple classes and class imbalance; a function can be associated with only a few genes, and there are several thousand functions describing genes in a given microarray data set.

4.3.5 Correlations Among Gene Expression Profiles

A major challenge in biological research is to understand the metabolic pathways and mechanisms of biological systems. The identification of correlated gene expressions in a microarray experiment is aimed at facilitating this objective. Several methods for this task are described in this section.

4.3.5.1 Main Methods for Clustering of Gene Expression Profiles Hierarchical clustering and K-means clustering are two of the most popular approaches for the clustering of microarray data. The *hierarchical clustering* approach used with microarray data is the *bottom-up approach*. This approach begins with single-member clusters, and small clusters are iteratively grouped together to form larger clusters, until a single cluster containing the whole set is obtained. In each iteration, the two clusters that are chosen for joining are two clusters with the closest distance to each other. The result of hierarchical clustering is a binary tree; descendants of each cluster in that tree are the two subclusters of which the cluster consists. The distance between two clusters in the tree reflects their correlation distance. Hierarchical clustering provides a visualization of the relationships between gene expression profiles (see Fig. 4.14).

K-means clustering groups genes into a prespecified number of clusters by minimizing the distances within each cluster and maximizing the distances between clusters. The K-means clustering method first chooses k genes called centroids (which can be done randomly or by making sure that their expression profiles are very different). It then examines all gene expression profiles and assigns each of these to the cluster with the closest centroid. The position of a centroid is recalculated each time a gene expression profile is added to the cluster by averaging all profiles within the cluster. This procedure is iteratively repeated until stable clusters are obtained, and no gene expression profiles switch clusters between iterations. The K-means method is computationally less demanding than hierarchical clustering. However, an obvious disadvantage is the need for the selection of parameter k, which is generally not a trivial task.

4.3.5.2 Alternative Clustering Methods for Gene Expression Profiles
Alternative clustering methods that are used with gene expression data include the self-organizing map (SOM) and random forest (RF) clustering.

An SOM is a clustering method implemented with a neural network and a special training procedure. The comparison of SOM with hierarchical clustering methods shows that an SOM is superior in both robustness and accuracy [51]. However, as K-means clusters, an SOM requires the value of parameter k to be prespecified.

RF clustering is based on an RF predictor that is a collection of individual classification trees. After an RF is constructed, the similarity measure between two samples can be defined as the number of times a tree predictor places the two samples in the same terminal node. This similarity measure can be used to cluster gene expression data [52]. It was demonstrated that the RF-based clustering of gene profiles is superior compared to the standard Euclidean distance measure [53].

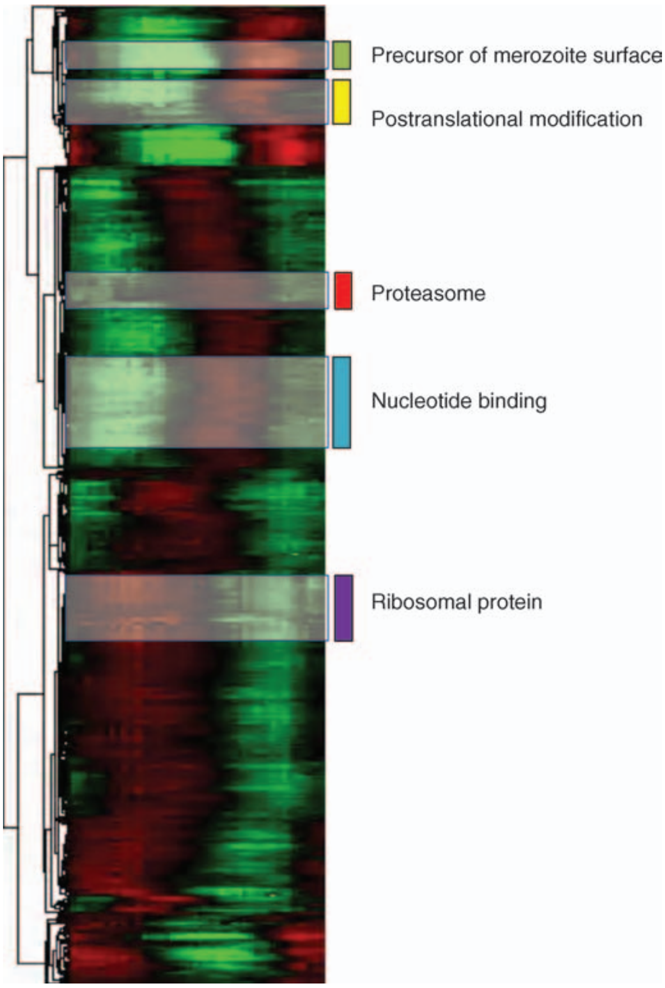

FIGURE 4.14 Visualization of hierarchically clustered data with identified functional correlation. The *Plasmodium* data set was clustered using hierarchical clustering. Rows of pixels represent genes' expression levels at different time points. Columns of pixels represent the expression level of all genes in one chip at one given time point in the IDC process, and their order corresponds to the order of points in time. The cluster hierarchy tree is on the left side. The image contains clearly visible patterns of red and green pixels that correspond to upregulated and downregulated expression levels, respectively. A domain expert investigated the higher level nodes in the clustering tree, examining the similarity of functions in each cluster for genes with known functions. Five examples of clusters for which the majority of genes are annotated with a common function are marked using the color bars and the names of the common functions. These clusters can be used to infer the functions of the genes within the same cluster whose function is unknown or unclear.

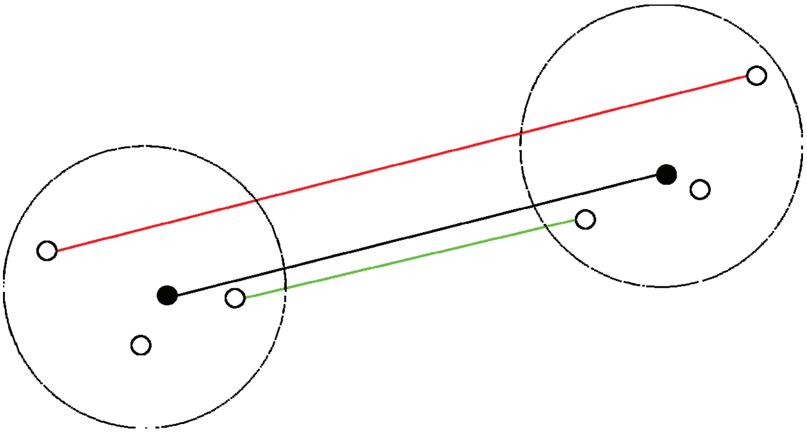

FIGURE 4.15 Cluster distance definitions. Hollow dots represent data points, and the two circles represent two distinct clusters of data points, while black dots are weighted centers of data points in each cluster. The green line illustrates the single linkage method of cluster distance, the red line illustrates the complete linkage method, and the black line represents the average linkage method.

Other advanced techniques proposed for clustering gene expression data include the mixture model approach [54], the shrinkage-based similarity procedure [55], the kernel method [56], and bootstrapping analysis [57].

4.3.5.3 Distance of Gene Expression Profile Clusters There are many ways to measure the distance between gene expression profiles and clusters of gene expression profiles. The Pearson correlation coefficient and the Euclidean distance are often used for well-normalized microarray data sets. However, microarray gene expression profiles contain noise and outliers. Nonparametric distance measures provide a way to avoid these problems. For instance, the Spearman correlation replaces gene expression values with their ranks before measuring the distance.

Average linkage, single linkage, and complete linkage are commonly used to measure the distances between clusters of gene expression profiles. Average linkage computes the distances between all pairs of gene expression profiles from two clusters and the average of these distances becomes the distance between the clusters. Single linkage defines the distance between two clusters as the distance between the two closest representatives of these clusters. Complete linkage defines the distance between two clusters as the distance between the two farthest representatives. The difference between these three definitions is illustrated in Figure 4.15.

4.3.5.4 Cluster Validation Regardless of the type of clustering, all obtained clusters need to be evaluated for biological validity before proceeding to further analysis. Visual validation is aimed at determining whether there are outliers in clusters or whether the gene expression profiles within each cluster are correlated to each other. If a problem is detected by validation, clusters are often refined by adjusting the

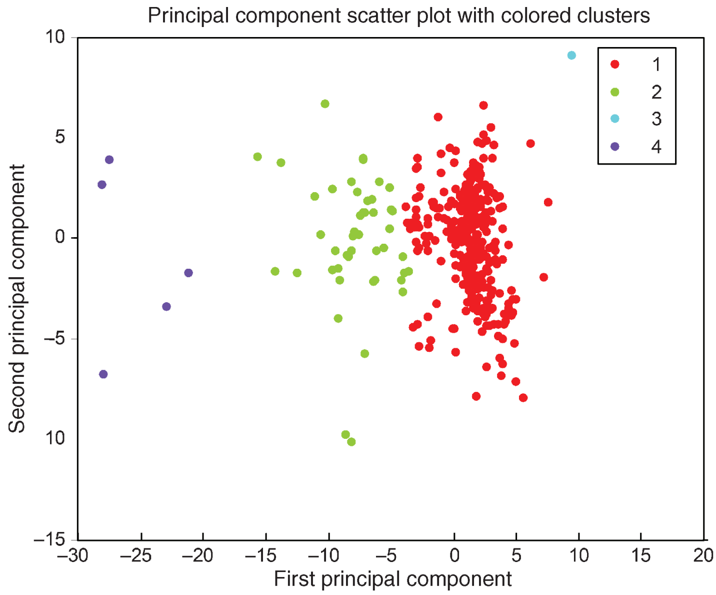

FIGURE 4.16 Principal component analysis. This scatterplot was obtained by plotting the first and the second principal component of the first 100 genes in an acute leukemia data set. It illustrates the benefit of PCA for visualizing data. There are apparently two to four clusters (depending on the criteria of separation of clusters), which is valuable information for the choice of parameter *k* in many clustering algorithms. A possible clustering to two groups of genes is shown as green and red points, while blue and cyan points can be discarded as outliers.

number of clusters (parameter *k*), the distance measuring method, or even by repeating the clustering with a different clustering method. Microarray data sets are highly dimensional. It is often difficult to provide a clear view of gene expression profile types within each cluster. By reducing the dimension of the microarray data set to two or three dimensions, analysis can be simplified and a visual overview of the data can be generated, which may provide useful information on gene expression profile clustering. Such a dimensionality reduction is typically achieved with principal component analysis (PCA). This technique finds the orthogonal components (also called *principal components*) of the input vectors and retains two or three orthogonal components with the highest variance. A visual examination of the projected clusters can help determine an appropriate number of distinct clusters for clustering as illustrated in Figure 4.16.

4.3.6 Biomarker Identification

One major challenge of microarray data analysis is sample classification. Examples of classification include the separation of people with and without CFS, or the classification of cancer patients into prespecified subcategories. Classifier construction includes the selection of the appropriate prediction model and the selection of features. Feature selection is a technique whereby genes with the most useful expression levels for classification are selected. Such genes can also be useful as *biomarkers* that in turn can be used for practical and cost-effective classification systems.

4.3.6.1 Classical Feature Selection Methods

Forward feature selection is an iterative process. It starts with an empty set of genes and at each iteration step adds the most informative of the remaining genes based on their ability to discriminate different classes of samples. This process is repeated until no further significant improvement of classification accuracy can be achieved. A reverse procedure, *backward feature elimination*, is also widely applied. It begins by using all the available genes and continues by dropping the least important genes until no significant improvement can be achieved.

In the *filter feature selection methods*, various statistical measures are used to rank genes by their discriminative powers. Successful measures include using the *t*-test, the chi-square test, information gain, and the Kruskal–Wallis test.

A recently proposed biomarker identification approach involves clustering gene expression profiles [58]. In such an approach, genes are clustered based on their microarray expression profiles. Then, within each cluster, the most representative gene is selected (the representative gene could be the gene closest to the mean or median expression value within the cluster). The representative genes are collected and used as selected features to build a predictor for classification of unknown samples. However, selected sets of genes often lack biological justification and their size is usually too large for experimental validation.

4.3.6.2 Domain Knowledge-Based Feature Selection

A recently proposed feature selection approach exploits the biological knowledge of gene functions as a criterion for selection [59]. The underlying hypothesis for this approach is that the difference between samples lies in a few key gene functions. Genes annotated with those key functions are likely to be very useful for classification. To use this observation, a statistical test is applied to microarray data in order to rank genes by their *p*-values and generate a subset of significant genes. Selected genes are compared to the overall population in order to identify the most significant function. Only genes associated with the most significant function are selected for classification. This approach results in a small set of genes that provide high accuracy (see the case study below).

Case Study 4.3: Feature Selection for Classification The CFS data set contains 39 test samples from patients clinically diagnosed with CFS and 40 control samples from subjects without CFS (nonfatigue, NF). The objective is to develop a predictor that classifies new subjects either as CFS or NF based on their gene expressions. Each microarray measures 20,160 genes.

We first used the Kruskal–Wallis test with *p*-value threshold of 0.05 for the initial gene selection. For each GO term, we count how many genes in the original set of 20,160 genes, as well as how many of the selected, are annotated with it. We then use the hypergeometric test to evaluate whether the representation of this GO term in the selected subset of genes is significantly greater than that in the original set of genes. We rank GO terms by their p-values and find the most overrepresented (those with smallest p-value) GO term. We narrow the selection of genes to include only the genes that are the most overrepresented GO term. We then select these genes as features for classification. Feature selection methods were tested using a leave-one-out cross-validation procedure. The prediction model used in all experiments was an SVM with quadratic kernel $k(x, y) = (C + x^{\mathrm{T}}y)^2$.

The Kruskal–Wallis test with a threshold of 0.05 produced the initial selection of 1296 genes. The overall accuracy of prediction with this feature selection method was 53 percent, which is barely better than the 50 percent accuracy of a random predictor. The proposed procedure narrowed the selection down to 17 genes. Although the number of features was reduced by almost two orders of magnitude, the overall accuracy of prediction with this smaller feature set improved to 72 percent. The GO term that was most often selected was GO:0006397 (mRNA processing). Interestingly, mRNA processing was verified by unrelated biological research as very important for CFS diagnosis [60]. We can compare the accuracy of the obtained predictor (72 percent) to the accuracy of a predictor with 17 features with the smallest p-values selected by the Kruskal–Wallis test, which was close to 50 percent; in other words, the predictor was not better than a trivial random predictor.

4.3.7 Conclusions

Microarray data analysis is a significant and broad field with many unresolved problems. This chapter briefly introduces some of the most commonly used methods for the analysis of microarray data, but many topics still remain. For example, microarray data can be used to construct gene networks, which are made up of links that represent relationships between genes, such as coregulation. Computational models for gene networks include Bayesian networks [61], Boolean networks [62], Petri nets [63], graphical Gaussian models [64], and stochastic process calculi [65].

Microarrays can also be studied in conjunction with other topics, such as microarray-related text mining, microarray resources and database construction, drug discovery, drug response study, and design clinical trials.

Several other types of microarrays are used in addition to gene expression microarrays: protein microarrays (including antibody microarrays), single-nucleotide polymorphism (SNP) microarrays, and chemical compound microarrays. Other experimental technologies, such as mass spectrometry, also produce results at a high throughput rate. Methods for the analysis of these various types of biological data have a certain degree of similarity with microarray data analysis. For example, methods used for the identification of differentially expressed genes are similar to the methods used for the identification of biomarkers in mass spectrometry data. Overall, there are many challenging open topics on analyzing high throughput biological data that can provide research opportunities for the data mining and machine learning community. Progress toward solving these challenges and the future directions of research in this area are discussed at various bioinformatics meetings; these include a specialized *International Conference for the Critical Assessment of Microarray Data Analysis (CAMDA)* that was established in 2000, and that was aimed at the assessment of the state-of-the-art methods in large-scale biological data mining. CAMDA provided standard data sets and put an emphasis on various challenges of analyzing large-scale biological data: time series cell cycle data analysis [45] and cancer sample classification using microarray data [3], functional discovery [42] and drug response [66], microarray data sample variance [67], integration of information from different microarray lung cancer data sets [68–71], the malaria transcriptome monitored by microarray data [4], and integration of different types of high throughput biological data related to CFS.

ACKNOWLEDGMENTS

This project is funded in part under a grant with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. We thank Samidh Chatterjee, Omkarnath Prabhu, Vladan Radosavljević, Lining Yu, and Jingting Zeng at our laboratory for carefully reading and reviewing this text. In addition, we would like to express special thanks to the external reviewers for their valuable comments on a preliminary manuscript.

REFERENCES

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270: 467–470. [Q3]
2. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.
3. Golub TR, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–537.
4. Bozdech Z, et al. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol 1, E5 (2003).
5. Vernon SD, Reeves WC. The challenge of integrating disparate high-content data: epidemiological, clinical and laboratory data collected during an in-hospital study of chronic fatigue syndrome. Pharmacogenomics 2006;7:345–354.
6. Yang YH, Buckley MJ, Speed TP. Analysis of cDNA microarray images. Brief Bioinform. 2001;2:341–349.
7. Yap G. Affymetrix, Inc. Pharmacogenomics 2002;3:709–711.
9. Kooperberg C, Fazzio TG, Delrow JJ, Tsukiyama T. Improved background correction for spotted DNA microarrays. J Comput Biol 2002;9:55–66.
9. Cui X, KM, Churchill GA. Transformations for cDNA microarray data. Stat Appl Genet Mol Biol 2003;2:article 4.
10. Troyanskaya O, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17:520–525.
11. Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 2005;21:187–198.

12. Johansson P, Hakkinen J. Improving missing value imputation of microarray data by using spot quality weights. BMC Bioinform 2006;7:306.
13. Tuikkala J, Elo L, Nevalainen OS, Aittokallio T. Improving missing value estimation in microarray data with gene ontology. Bioinformatics 2006;22:566–572.
14. Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002; 32(Suppl): 496–501.
15. Yang YH, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30:e15.
16. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods 2003;31: 265–273.
17. Berger JA, et al. Optimized LOWESS normalization parameter selection for DNA microarray data. BMC Bioinform 2004;5:194.
18. Colantuoni CHG, Zeger S, Pevsner J. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. Biotechniques 2002;32:1316–1320.
19. Holter NS, et al. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci USA 2000;97:8409–8414.
20. Bolstad BM, Irizarry RA, Astrand M, Speed TP, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185–193.
21. Wolfinger RD, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 2001;8:625–637.
22. Schadt EE, Li C, Ellis B, Wong WH, Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem Suppl 2001; 37:120–125.
23. Irizarry RA, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4:249–264.
24. Yu X, Chu TM, Gibson G, Wolfinger RD, A mixed model approach to identify yeast transcriptional regulatory motifs via microarray experiments. Stat Appl Genet Mol Biol 2004;3:Article22.
25. Ramsey FL, Shafer DW. The statistical sleuth: a course in methods of data analysis; 1997. [Q4]
26. Kerr MK, Martin M, Churchill GA, Analysis of variance for gene expression microarray data. J Comput Biol 2000; 7:819–837.
27. Pan WA comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 2002; 18:546–554.
28. Singer JD. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. journal of educational and behavioral statistics 1998; 24: 323–355.
29. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001; 98:5116–5121.
30. Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res 2001; 11:1227–1236.
31. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol 2002; 23:70–86.

32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. JRSSB 1995; 57:289–300. [Q5]

33. Qian HR, Huang S. Comparison of false discovery rate methods in identifying genes with differential expression. Genomics 2005; 86:495–503.

34. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25:25–29.

35. Consortium GO, Creating the gene ontology resource: design and implementation. Genome Res 2001; 11:1425–1433.

36. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003; 19:1275–1283.

37. Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinforma 2006; 7:302.

38. Rada R, Mili H, Bicknell E, Blettner M. development and application of a metric on semantic nets. IEEE Trans Syst Man Cybernet 1989; 19:17–30.

39. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004; 20:1464–1465.

40. Dennis G, Jr, et al. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol 2003; 4:P3.

41. Chu S, et al. The transcriptional program of sporulation in budding yeast. Science 1998; 282:699–705.

42. Hughes TR, et al. Functional discovery via a compendium of expression profiles. Cell 2000; 102:109–126.

43. Karaoz U, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA 2004;101:2888–2893.

44. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–14868.

45. Spellman PT, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 1998;9: 3273–3297.

46. Whitfield ML, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 2002;13:1977–2000.

47. Zhou X, Kao MC, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci USA 2002;99:12783–12788.

48. Xie H, Vucetic S, Sun H, Hedge P, Obradovic Z. Characterization of gene functional expression profiles of *Plasmodium falciparum*. Proceedings of the 5th Conference on Critical Assessment of Microarray Data Analysis; 2004.

49. Barutcuoglu Z, RES, Troyanskaya OG. Hierarchical multi-label prediction of gene function. Bioinformatics 2006;22:830–836.

50. Brown MP, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;97:262–267.

51. Mangiameli P, Chen SK, West D. A comparison of SOM of neural network and hierarchical methods. Eur J Operat Res 1996;93:402–417.

52. Breiman L. Random forests. Mach Learning 2001;45:5–32.

53. Shi T, S D, Belldegrun AS, Palotie A, Horvath S. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol 2005;18:547–557.

54. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 2002;18:413–422.

55. Cherepinsky V, Feng J, Rejali M, Mishra B. Shrinkage-based similarity metric for cluster analysis of microarray data. Proc Natl Acad Sci USA 2003;100:9668–9673.

56. Verri A. A novel kernel method for clustering. IEEE Trans Pattern Anal Mach Intell 2005;27:801–805.

57. Kerr K, Churchill GA. Bootstrapping cluster analysis: access the reliable of conclusions from microarray experiments. Proc Natl Acad Sci 2001;98:8961–8965.

58. Au W, Chan K, Wong A, Wang Y. Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Trans Comput Biol Bioinform 2005;2:83–101.

59. Xie H, Obradovic Z, Vucetic S. Mining of microarray, proteomics, and clinical data for improved identification of chronic fatigue syndrome. In: Proceedings of the Sixth International Conference for the Critical Assessment of Microarray Data Analysis; 2006.

60. Whistler T, Unger ER, Nisenbaum R, Vernon SD. Integration of gene expression, clinical, and epidemiologic data to characterize chronic fatigue syndrome. J Transl Med 2003;1:10.

61. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Combining location and expression data for principled discovery of genetic regulatory network models. Pac Symp Biocomput 2002;437–449. [Q6]

62. Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput 1999;17–28.

63. Gambin A, Lasota S, Rutkowski M. Analyzing stationary states of gene regulatory network using Petri nets. In Silico Biol 2006;6:0010.

64. Toh H, Horimoto K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics 2002;18:287–297.

65. Golightly A, Wilkinson DJ. Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics 2005;61:781–788.

66. Scherf U, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24:236–244.

67. Pritchard CC, Hsu L, Delrow J, Nelson PS. Project normal: defining normal variance in mouse gene expression. Proc Natl Acad Sci USA 2001;98:13266–13271.

68. Wigle DA, et al. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. Cancer Res 2002;62:3005–3008.

69. Beer DG, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–824.

70. Garber ME, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci USA 2001;98:13784–13789.

71. Bhattacharjee A, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001;98:13790–13795.