32. Statistical Genetics for Genomic Data Analysis

In this chapter, we briefly summarize the emerging statistical concepts and approaches that have been recently developed and applied to the analysis of genomic data such as microarray gene expression data. In the first section we introduce the general background and critical issues in statistical sciences for genomic data analysis. The second section describes a novel concept of statistical significance, the so-called false discovery rate, the rate of false positives among all positive findings, which has been suggested to control the error rate of numerous false positives in large screening biological data analysis. In the next section we introduce two recent statistical testing methods: significance analysis of microarray (SAM) and local pooled error (LPE) tests. The latter in particular, which is significantly strengthened by pooling error information from adjacent genes at local intensity ranges, is useful to analyze microarray data with limited replication. The fourth section introduces analysis of variation (ANOVA) and heterogenous error modeling (HEM) approaches that have been suggested for analyzing microarray data obtained from multiple experimental and/or biological conditions. The last two sections describe data exploration and discovery tools largely termed *supervised learning* and *unsupervised learning*. The former approaches include several multivariate statistical methods for the investigation of coexpression patterns of multiple genes, and the latter approaches are used as classification methods to discover genetic markers for predicting important subclasses of human diseases. Most of the statistical software packages for the approaches introduced in this chapter are freely available at the open-source bioinformatics software web site (Bioconductor; http://www.bioconductor.org/).

Accelerated by the Human Genome Project, recent advances in high-throughput biotechnologies have dramatically changed the horizon of biological and biomedical sciences. Large screening expression profiling techniques such as DNA microarrays, mass spectrometry, and protein chips offer great promise for functional genomics and proteomics research, and have the potential to transform the diagnosis and treatment of human diseases [32.1]. In particular, DNA microarray and GeneChip^{TM}gene expression approaches are becoming increasingly important in current biomedical studies [32.2–4].

Analysis of such genome-wide data, however, has brought extreme challenges not only in the biological sciences but also in the statistical sciences. Fundamental difficulties exist in applying traditional statistical approaches to genome-wide expression data, namely the *multiple comparisons issue* and the *small n–large p problem* [32.5]. The former problem arises because classical statistical hypothesis testing, modeling, and inference strategies are designed for studying a small number of candidate targets at a time, whereas one often investigates tens of thousands of genes' differential expression in a single microarray study. For example,

when a two-sample t-test is applied for evaluating statistical significance of thousands of genes' differential expression patterns in a microarray study, the p-values obtained from this within-gene test must be adjusted to take into account the random chance of all the candidate genes in the array data.

The latter difficulty, the small n–large p problem, arises due to the fact that many biological and biomedical microarray studies are performed with a small number of replicated arrays, or without replication. Unlike DNA sequence information, gene expression data are context-dependent and offer different interpretations depending on (patient) sample condition, time point, and treatment for a single subject [32.6]. In addition to the high costs of microarray experiments, certain biological or human patient specimens are often limited, thereby necessitating that microarray studies be performed with limited replication. Consequently, one must perform statistical inference on a small number of observations (n) compared to a large number of potential predictor genes (p). The latter number of tens of thousands of genes is simply too large to be considered in standard statistical testing and modeling, whereas the sample size (or number of replicated arrays at each condition) of a microarray study is typically small, a few tens at most and often only one or two replicates. This presents great difficulty for the application of traditional statistical approaches, which generally require a reasonably large sample size for maximal performance. As microarray (and similar high-throughput) technology becomes an important tool in biological and biomedical investigation, the lack of appropriate statistical methods for large screening microarray data will undoubtedly become a great obstacle in the current biological sciences. In this chapter, we will briefly summarize the statistical approaches that have been applied to microarray gene expression data analysis by avoiding these pitfalls.

We also introduce several multivariate statistical methods that have been used to investigate the coregulation structures of multiple genes as *unsupervised learning* and to discover genetic markers for predicting important subclasses of human diseases and biological targets as *supervised learning*. In particular, clustering approaches have been widely applied to the analysis of gene expression microarray data. The method of visualizing gene expression data based on cluster order, so-called cluster-image map (CIM) analysis, is found to be very efficient in summarizing the thousands of gene expression values and aiding in the identification of some interesting patterns of gene expression [32.3,7,8]. Since a clustering algorithm provides an efficient dimension reduction for extremely high-dimensional data based on their association, it is much easier to simultaneously screen thousands of gene expression values and to identify interesting patterns on the image maps. The statistical software packages for most of these approaches are freely available at the open-source bioinformatics development web site (Bioconductor; http://www.bioconductor.org/).

We note that this kind of microarray data analysis is implemented based on certain standard preprocessing procedures. Suppose we have gene expression data with n genes and p arrays. A matrix of this gene expression data is defined by $Y_{n\times p}$ with n rows and p columns. The data are then typically log2-transformed to remedy the right-skewed distribution, to make error components additive, and to aplly other statistical procedures that are based on underlying Gaussian distributional assumptions. Each column of the matrix (or each array) is scaled or normalized to a common baseline by matching interquartile ranges or by nonparametric regression methods, e.g., lowess.

32.1 False Discovery Rate

In order to avoid a large number of false positive findings (or type I errors) in genomic data analysis, the classical family-wise error rate (FWER) has initially been used to control for the random chance of multiple candidates by evaluating the probability that at most one false positive is included at a cutoff level of a statistic [32.9]. However, FWER adjustment has been found to be very conservative in microarray studies, resulting in a high false-negative error rate [32.10]. To avoid pitfalls such as this, a novel statistical significance concept, the so-called *false discovery rate* (FDR) and its refinement, the *q-value*, have been suggested [32.11, 12] (QVALUE package at Bioconductor). FDR is defined as follows. Consider a family of m simultaneously tested null hypotheses of which m_0 are true. For each hypothesis H_i a test statistic is calculated along with the corresponding p-value, P_i. Let R denote the number of hypotheses rejected by a procedure, V the number of true null

hypotheses erroneously rejected, and S the number of false hypotheses rejected as summarized in Table 32.1. Now let Q denote V/R when $R > 0$ and 0 otherwise. Then the FDR is defined as the expectation of Q, i. e. FDR$= E(Q)$.

As shown in [32.11], the FDR of a multiple comparison procedure is always smaller than or equal to the FWER, where equality holds if all null hypotheses are true. Thus, control of the FDR implies control of the FWER only when all null hypotheses are true, but it generally controls such an error rate much less conservatively than FWER because there exist quite a few true positives in practical data analysis. In the context of gene expression analysis, this result means that, if FDR is controlled at some level q, then the probability of erroneously detecting any differentially expressed genes among all genes identified by a certain selection criterion is less than or equal to q. Intuitively, FDR controls the expected proportion of false positives among all candidate genes identified significantly by a testing criterion. Therefore, based on FDR, researchers can now assess their statistical confidence among the identified targets with a much smaller false-negative error rate. FDR evaluation has been rapidly adopted for microarray data analysis including the significance analysis of microarrays (SAM) and local pooled error (LPE) approaches [32.9, 10, 13].

Table 32.1 Outcomes when testing m hypotheses

Hypothesis	Accept	Reject	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

32.2 Statistical Tests for Genomic Data

Each gene's differential expression pattern in a microarray experiment is usually assessed by (typically pairwise) contrasts of mean expression values among experimental conditions. Such comparisons have been routinely assessed as fold changes whereby genes with greater than two or three fold changes are selected for further investigation. It has been frequently found that a gene showing a high fold-change between experimental conditions might also exhibit high variability and hence its differential expression may not be significant. Similarly, a modest change in gene expression may be significant if its differential expression pattern is highly reproducible. A number of authors have pointed out this fundamental flaw in the fold-change-based approach [32.14]. In order to assess differential expression in a way that controls both false positives and false negatives, a standard approach is emerging based on statistical significance and hypothesis testing, with careful attention paid to the reliability of variance estimates and multiple comparison issues.

The classical two-sample t-statistic has initially been used for testing each gene's differential expression; procedures such as the Westfall–Young step-down method have been suggested to control FWER [32.9]. These t-test procedures, however, rely on reasonable estimates of reproducibility or within-gene error to be constructed, requiring a large number of replicated arrays. When a small number of replicates are available per condition, e.g. duplicate or triplicate, the use of naive, within-gene estimates of variability does not provide a reliable hypothesis-testing framework. For example, a gene may have very similar differential expression values in duplicate experiments by chance alone. This can lead to inflated signal-to-noise ratios for genes with low but similar expression values. Furthermore, the comparison of means can be misled by outliers with dramatically smaller or bigger expression intensities than other replicates. As such, error estimates constructed solely within genes may result in underpowered tests for differential expression comparisons and also result in large numbers of false positives.

A number of approaches to improving estimates of variability and statistical tests of differential expression have thus recently emerged. Several variance function methods have been proposed. Reference [32.15] suggested a simple regression estimation of local variances; [32.16] used a smoothing-spline pooled-error method by regressing standard error estimates on the mean log intensities; and [32.17] estimates a two-parameter variance function of mean expression intensity. Reference [32.18] compared some of these variance-estimation methods. Recently, [32.19] suggested the use of data-adapted robust estimate of array error based on a smoothing spline and standardized local median absolute deviation (MAD). The variance function methods described above borrow strength across genes in order to improve reliability of variance estimates in differential expression tests. This is conceptually similar to the SAM method of [32.10] and the empirical Bayes methods of [32.20] and [32.21].

These methods also shrink the within-gene variance estimate towards an estimate including more genes, and construct signal-to-noise ratios using the shrunken variance in a similar fashion to the local-pooled-error (LPE) test described below.

The local-pooled-error (LPE) estimation strategy has also been introduced for within-gene expression error, whereby variance estimates for genes are formed by pooling variance estimates for genes with similar expression intensities from replicated arrays within experimental conditions [32.13]. The LPE approach leverages the observations that genes with similar expression intensity values often show similar array-experimental variability within experimental conditions; and that variance of individual gene expression measurements within experimental conditions typically decreases as a (nonlinear) function of intensity [32.5,22]. The LPE approach handles the situation where a gene with low expression may have very low variance by chance and the resulting signal-to-noise ratio is unrealistically large. The pooling of errors within local intensities shrinks such variances to the variance of genes with similar intensities. In this chapter, two recent statistical testing procedures – SAM and LPE – are described in more detail while many classical testing and p-value adjustment strategies can be found elsewhere [32.9].

32.2.1 Significance Analysis of Microarrays

The significance analysis of microarrays (SAM) approach is based on analysis of random fluctuations in the data [32.10] (SIGGENES package at Bioconductor). Based on the observation that the signal-to-noise ratio decreases with decreasing gene expression, as shown in [32.13], fluctuations are considered to be gene specific even for a given level of expression in [32.10]. To account for gene-specific fluctuations, a statistic is defined based on the ratio of change in gene expression to standard deviation in the data for that gene. The *relative difference* $d(i)$ in gene expression is:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0} , \qquad (32.1)$$

where $x_I(i)$ and $x_U(i)$ are defined as the average levels of expression for gene (i) in states I and U, respectively. The *gene-specific scatter* $s(i)$ is the standard deviation of repeated expression measurements:

$$s(i) = \sqrt{a\left\{\sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2\right\}} , \qquad (32.2)$$

where $\sum_m$ and $\sum_n$ are summations of the expression measurements in states I and U, respectively, $a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2)$, and n_1 and n_2 are the numbers of measurements in states I and U. To compare values of $d(i)$ across all genes, the distribution of $d(i)$ is assumed to be independent of the level of gene expression. At low expression levels, variance in $d(i)$ can be high because of small values of $s(i)$. To ensure that the variance of $d(i)$ is independent of gene expression, a small positive constant s_0 is added to the denominator of (32.1). The coefficient of variation of $d(i)$ is computed as a function of $s(i)$ in moving windows across the data. The value for s_0 was chosen to minimize the coefficient of variation.

32.2.2 The Local-Pooled-Error Test

The local-pooled-error (LPE) method has been introduced specifically for analysis of small-sample microarray data, whereby error variance estimates for genes are formed by pooling variance estimates for genes with similar expression intensities from replicated arrays within experimental conditions [32.13] (LPE package at Bioconductor). The LPE approach leverages the observations that genes with similar expression intensity values often show similar array-experimental variability within experimental conditions; and that variance of individual gene-expression measurements within experimental conditions typically decreases as a (nonlinear) function of intensity, as shown in Fig. 32.1. This is due, in part, to common background noise at each spot of the microarray. At high levels of expression intensity, this background noise is dominated by the expression intensity, while at low levels the background noise is a larger component of the observed expression intensity. The LPE approach controls the situation where a gene with low expression may have very low variance by chance and the resulting signal-to-noise ratio is unrealistically large. The LPE method borrows strength across genes in order to improve accuracy of error variance estimation in microarray data. This is conceptually similar to the SAM method above and the empirical Bayes methods of [32.20], which shrink the within-gene variance estimate towards an estimate including more genes in a similar fashion to LPE.

To take into account heterogeneous error variability across different intensity ranges in microarray data, the LPE method can be applied as follows (refer to [32.13] for a more detailed technical description). For oligo array data, let x_{ijk} be the observed expression intensity at gene j for array k and sample i. For duplicate

arrays, $k = 1, 2$, plots of $A = \log 2(x_{ij_1} x_{ij_2})/2$ versus $M = \log 2(x_{ij_1}/x_{ij_2})$, $j = 1, \ldots, J$, can facilitate the investigation of between-duplicate variability in terms of overall intensity. The A versus M (or AM) plot provides a very raw look at the data and is useful in detecting outliers and patterns of intensity variation as a function of mean intensity [32.9]. At each of the local intensity regions of the AM plot under a particular biological condition, the unbiased estimate of the local variance is obtained. A cubic spline is then fit to these local variance estimates to obtain a smoothing variance function. The optimal choice of the effective degree of freedom df_λ of the fitted smoothing spline is obtained by minimizing the expected squared prediction error. This two-stage error estimation approach – estimation of the error of M within quantiles and then nonparametric smoothing on these estimates – is used because direct nonparametric estimation often leads to unrealistic (small or large) estimates of error when only a small numbers of observations are available at a fixed-width intensity range.

Based on the LPE estimation above, statistical significance of the LPE-based test is evaluated as follows. First, each gene's medians under the two compared conditions are calculated to avoid artifacts from outliers. The approximate normality of medians can be assumed with a small number of replicates based on the fact that the individual log-intensity values within a local intensity range follow a normal distribution [32.13]. The LPE statistic for the median (log-intensity) difference z is then calculated as:

$$z = (m_1 - m_2)/s_{\text{pooled}} , \qquad (32.3)$$

where m_1 and m_2 are the median intensities in two the compared array-experimental conditions X and Y, and s_{pooled} is the pooled standard error, $\left[s_x^2(m_1)/n_1 + s_y^2(m_2)/n_2\right]^{1/2}$ from the LPE-estimated baseline variances of s_x^2 and s_y^2. The LPE approach shows a significantly better performance than two-sample t-test, SAM, and Westfall–Young permutation tests, especially when the number of replicates is smaller

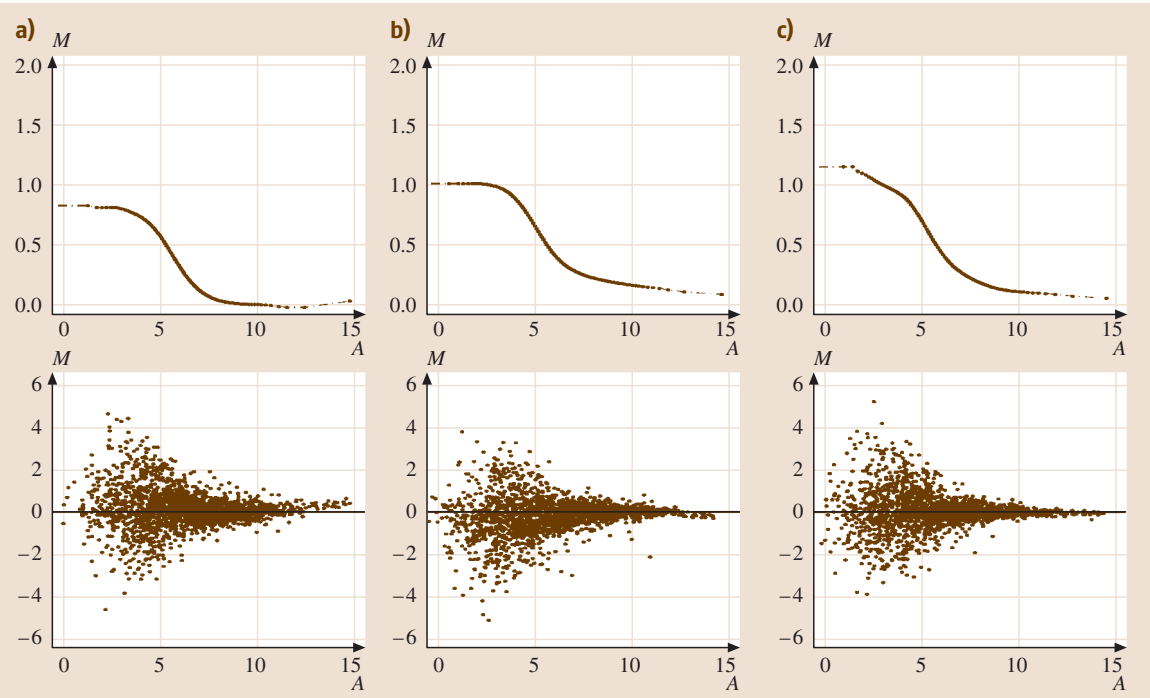

Fig. 32.1a–c Log intensity ratio $\log_2 \frac{X_{ij1}}{X_{ij2}}$ (M) as a function of average gene expression $\log_2 \sqrt{x_{ij1} x_{ij2}}$ (A). The *top row* of panels **(a)**, **(b)** and **(c)** represent local pooled error (LPE) for naive, 48 h activated, and T-cell clone D4 conditions, respectively, in the mouse immune-response microarray study in [32.13]. Variance estimates in percentile intervals are shown as *points*, and a *smoothed curve* superimposing these points is also shown. The *bottom row* of panels represent the corresponding M-versus-A graphs. The *horizontal line* represents identical expression between replicates

than ten. In a simulation study from a Gaussian distribution without extreme outliers, the LPE method showed a significant improvement of statistical power with three and five replicates (see Figure 2 in [32.13]).

32.3 Statistical Modeling for Genomic Data

Microarray gene-expression studies are also frequently performed for comparing complex, multiple biological conditions and pathways. Several linear modeling approaches have been introduced for analyzing microarray data with multiple conditions. Reference [32.23] considered an analysis of variance (ANOVA) model to capture the effects of dye, array, gene, condition, array–gene interaction, and condition–gene interaction separately on cDNA microarray data, and [32.24] proposed a two-stage mixed model that first models cDNA microarray data with the effects of array, condition, and condition–array interaction, and then fits the residuals with the effects of gene, gene–condition interaction, and gene–array interaction. Several approaches have also been developed under the Bayesian paradigm for analyzing microarray data including: the Bayesian parametric modeling [32.25], the Bayesian regularized t-test [32.21], the Bayesian hierarchical modeling with a multivariate normal prior [32.26], and the Bayesian heterogeneous error model (HEM) with two error components [32.27]. The ANOVA and HEM approaches are introduced in this chapter.

32.3.1 ANOVA Modeling

Reference [32.23] first suggested the use of analysis of variance (ANOVA) models to both estimate relative gene expression and to account for other sources of variation in microarray data. Even though the exact form of the ANOVA model depends on the particular data set, a typical ANOVA model for two-color-based cDNA microarray data can be defined as

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (AD)_{ij} + (AG)_{ig} + (DG)_{ig} + (VG)_{kg} + \epsilon_{ijkg} \,, \tag{32.4}$$

where y_{ijkg} is the measured intensity from array i, dye j, variety k, and gene g on an appropriate scale (typically the log scale). The generic term *variety* is often used to refer to the mRNA samples under study, such as treatment and control samples, cancer and normal cells, or time points of a biological process. The terms A, D, and AD account for all effects that are not gene-specific. The gene effects G_g capture the average levels of expression for genes and the array-by-gene interactions $(AG)_{ig}$ capture differences due to varying sizes of spots on arrays. The dye-by-gene interactions $(DG)_{jg}$ represent gene-specific dye effects. None of the above effects are of biological interest, but amount to a normalization of the data for ancillary sources of variation. The effects of primary interest are the interactions between genes and varieties, $(VG)_{kg}$. These terms capture differences from overall averages that are attributable to the specific combination of variety k and gene g. Differences among these variety-by-gene interactions provide the estimates for the relative expression of gene g in varieties 1 and 2 by

$$(VG)_{1g} - (VG)_{2g} \,.$$

Note that AV, DV, and other higher-order interaction terms are typically assumed to be negligible and considered together with the error terms. The error terms ϵ_{ijkg} are often assumed to be independent with mean zero and variance σ^2. However, such a *global* ANOVA model is difficult to implement in practice due to its computational restriction. Instead, one often considers gene-by-gene ANOVA models like:

$$y_{ijkg} = \mu_g + A_i + D_j + V_k + (AD)_{ij} + (VG)_{kg} + \epsilon_{ijkg} \,. \tag{32.5}$$

Alternatively, a two-stage ANOVA model is used [32.24]. The first layer is for the main effects that are not specific to the gene

$$y_{ijkg} = \mu + A_i + D_j + V_k + (AD)_{ij} + (AG)_{ig} + \epsilon_{ijkg} \,. \tag{32.6}$$

Let r_{ijkg} be the residuals from this first ANOVA fit. Then, the second-layer ANOVA model for gene-specific effects is considered as

$$r_{ijkg} = G_g + + (AG)_{ig} + (DG)_{ig} + (VG)_{kg} + \nu_{ijkg} \,. \tag{32.7}$$

Except the main effects of G and V and their interaction effects, the other terms A, D, (AD), (AG), and (DG) can be considered as random effects. These within-gene ANOVA models can be implemented using most standard statistical packages, such as R, SAS, or SPSS.

32.3.2 The Heterogeneous Error Model

Similarly to the statistical tests for comparing two sample conditions, the above within-gene ANOVA modeling methods are underpowered and have inaccurate error estimation in microarray data with limited replication. Using a Bayesian hierarchical approach with LPE-based (or error-pooling) empirical Bayes prior constructions, [32.27] have constructed a heterogeneous error model (HEM) with two layers of error to decompose the total error variability into the technical and biological error components in microarray data (HEM package at Bioconductor). Utilizing the LPE-estimated error-distribution information of microarray data for its empirical Bayes prior specifications, this modeling strategy provides separate estimates of the technical and biological error components in microarray data, especially the former error component, significantly more accurately. The first layer is constructed to capture the array technical variation due to many experimental error components, such as sample preparation, labeling, hybridization, and image processing

$$y_{ijkl} = x_{ijk} + \epsilon_{ijkl}\,, \quad \text{where}$$
$$\epsilon_{ijkl} \sim \text{iid Normal}\left[0, \sigma_\epsilon^2(x_{ijk})\right]\,, \tag{32.8}$$

where i = 1, 2, ..., G; j = 1, 2, ..., C; k = 1, 2, ..., m_{ij}; l = 1, 2, ..., n_{ijk}, and iid means independently and identically distributed. The second layer is then hierarchically constructed to capture the biological error component:

$$x_{ijk} = \mu + g_i + c_j + r_{ij} + b_{ijk}\,, \quad \text{where}$$
$$b_{ijk} \text{ iid Normal}\left[0, \sigma_b^2(ij)\right]\,. \tag{32.9}$$

Here, the genetic parameters are for the grand mean (shift or scaling) constant, gene, cell, interaction effects, and the biological error; the last error term varies and is heterogeneous for each combination of different genes and conditions. Note that the biological variability is individually assessed for discovery of biologically relevant expression patterns in this approach.

Bayesian posterior estimates and distributions are quite dependent on their prior specifications when the sample size is small in a microarray study. This difficulty in Bayesian applications to microarray data has been well-recognized and several authors have suggested the use of more-informative empirical Bayes priors [32.28, 29]. In these studies, empirical Bayes (EB) priors are used for defining distributions of genes with different expression patterns, e.g., distributions for equivalently and differentially expressed genes. Such specifications would be useful to determine each gene's expression pattern when the number of different expression patterns is small. However, as the number of conditions increases, the number of expression patterns increases exponentially, and these EB approaches quickly become impractical; many of these prior distributions also become unidentifiable. Conversely, the EB priors in HEM are used for specification of the two layers of error – technical and biological errors – which can be directly observed from each array data set, and can also be reliably estimated by the LPE method, pooling information from the genes with similar expression intensity. Thus, a nonparametric EB prior for the technical error $\sigma^2(x_{ijk})$ that is estimated by the LPE method and sampled by bootstrapping at each intensity x_{ijk}, whereas a parametric EB prior, Gamma (α, β_{ij}) is used because this error should freely vary to reflect the actual sampling variability of different biological subjects. Using these error-pooling-based prior specifications, HEM has demonstrated its improved performance in small sample microarray data analysis both in simulated and practical microarray data, see Fig. 32.2.

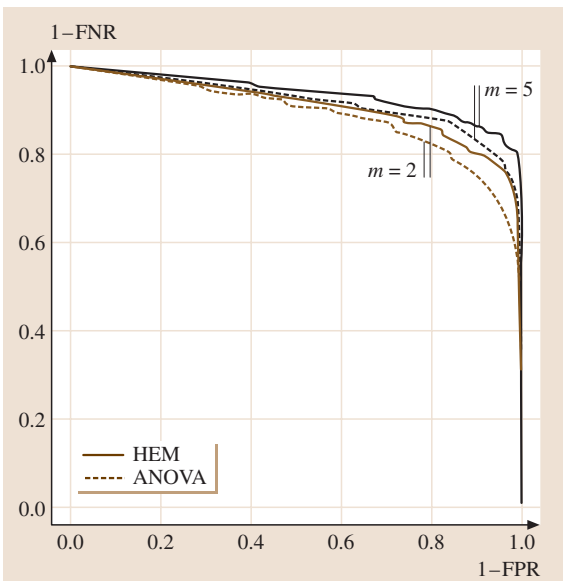

Fig. 32.2 ROC curves from HEM (*solid lines*) and ANOVA (*dotted lines*) models with two and five replicated arrays; The *horizontal axis* is 1 − FPR = 1 − Pr(positive|negative) and the *vertical axis* is 1 − FNR = 1 − Pr(negative|positive)

32.4 Unsupervised Learning: Clustering

Clustering analysis is widely applied to search for the groups (clusters) in microarray data because these techniques can effectively reduce the high-dimensional gene-expression data into a two-dimensional dendrogram organized by each gene's expression-association patterns. These clustering approaches first need to be defined by a measure or distance index of similarity or dissimilarity such as

- Euclidean: $d(x, y) = \sum_k (x_k - y_k)^2$;
- Manhattan: $d(x, y) = \sum |x_k - y_k|$;
- Correlation: $d(x, y) = 1 - r(x, y)$, where $r(x, y)$ is the Pearson or Spearman sample-correlation coefficient.

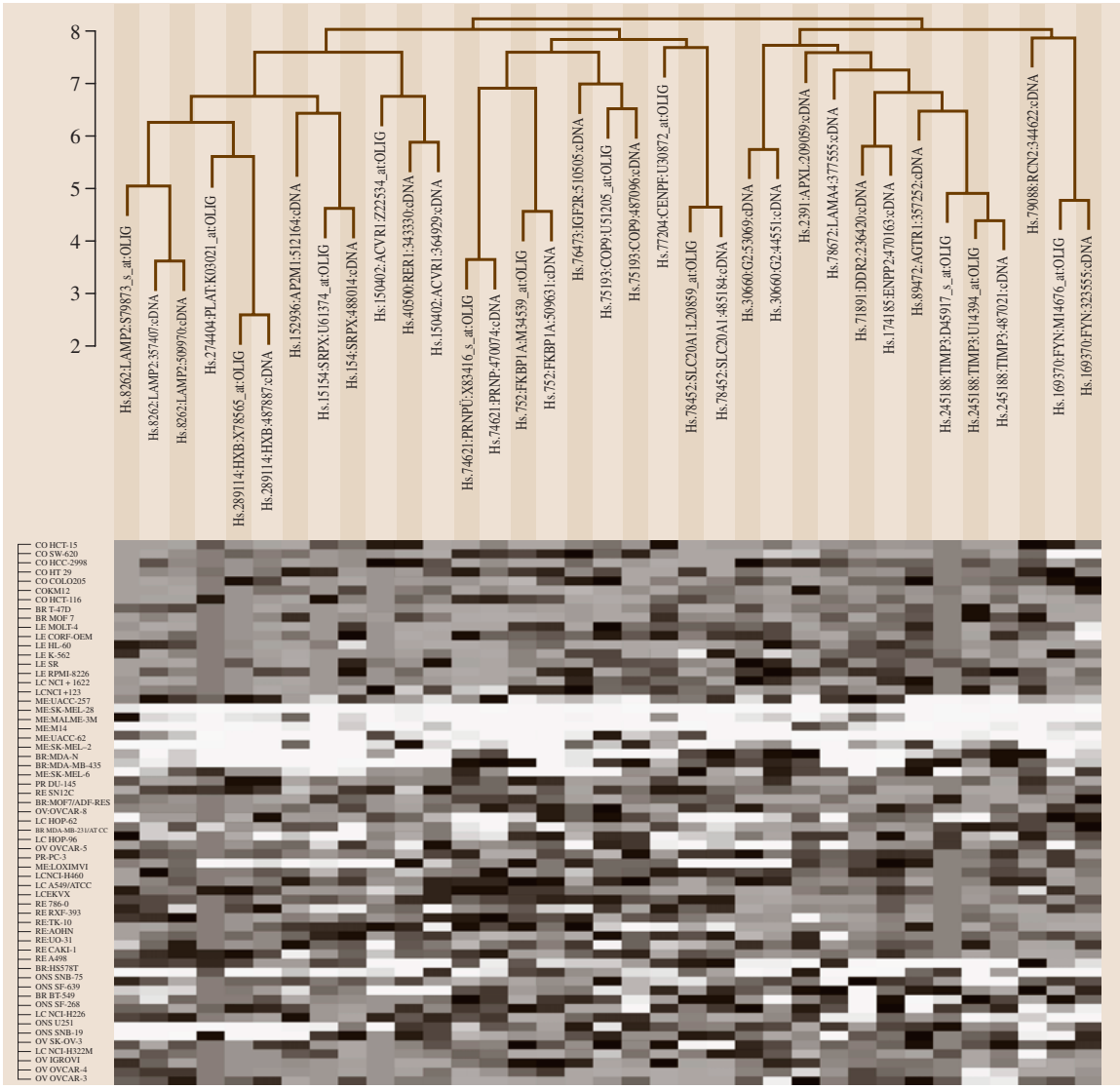

Fig. 32.3 Clustered image maps (CIMs) for hierarchical clustering of the cDNA and oligo array expression patterns. Each gene expression pattern is designated as coming from the cDNA or oligo array set. A region of CIM occupied by melanoma genes from the combined set of 3297 oligo and cDNA transcripts [32.3]

Note that if x and y are standardized, i. e., subtracted by each mean and divided by each standard deviation, then Euclidean and correlation distances can be easily shown to be mathematically equivalent:

$$\sum_k (x_k - y_k)^2 = \sum_k \left(x_k^2 + y_k^2 - 2x_k y_k\right)$$
$$= 2\left(1 - \sum_k x_k y_k\right)$$
$$= 2[1 - r(x, y)] .$$

Two classes of clustering algorithms have been used in genomic data analysis. The first class of clustering algorithms is based on hierarchical allocation including

1. Agglomerative methods:
   a) average linkage based on group average distance [32.3, 7]
   b) single linkage based on minimum nearest distance
   c) complete linkage based on maximum furthest distance.
2. Probability-based clustering: Bayes factor or posterior probability for choosing k clusters
3. Divisive methods: monothetic variable division, polythetic division

A cluster-image map is shown for the microarray data of the NCI 60 cancer cell lines in Fig. 32.3.

The second class is the partitioning algorithms that divide the data into a prespecified number of subsets including:

1. Self-organizing map: divides the data into a geometrically preset grid structure of subclusters [32.8];
2. Kmeans: iterative relocation clustering into a predefined number of subclusters;
3. Pam (partitioning around medoids) similar to, but more robust than Kmeans clustering;
4. Clara: clustering for applications to large data sets;
5. Fuzzy algorithm: provide fractions of membership, rather than deterministic allocations.

One of the most difficult aspects of using these clustering analyses is to interpret their heuristic, often unstable clustering results. To overcome this shortcoming, several refined clustering approaches have been suggested. For example, [32.23] suggest the use of bootstrapping to evaluate the consistency and confidence of each gene's membership to particular cluster groups. The *gene shaving* approach has been suggested to find the clusters directly relevant to major variance directions of an array data set [32.30]. Recently, *tight clustering*, a refined bootstrap-based hierarchical clustering is proposed to formally assess and identify the groups of genes that are most tightly clustered to each other [32.31].

32.5 Supervised Learning: Classification

Applications of microarray data have received considerable attention in many challenging classification problems in biomedical research [32.2, 32, 33]. In particular, such applications have been conducted in cancer research as alternative diagnostic techniques to the traditional ones such as classification by the origin of cancer tissues and/or the microscopic appearance; the latter are far from satisfaction for the prediction of many critical human disease subtypes [32.34]. Several different approaches to microarray classification modeling have been proposed, including gene voting [32.2], support vector machines (SVMs) [32.35, 36], Bayesian regression models [32.33], partial least squares [32.37], genetic-algorithm k-nearest-neighbor (GA/KNN) method [32.38], and between-group analysis [32.39].

Microarray data often have tens of thousands of genes on each chip whereas only a few tens of samples or replicated arrays are available in a microarray study. Therefore, it is desirable to avoid overfitting and to find a best subset of the thousands of genes for constructing classification rules and models that are robust to different choices of training samples and provide consistent prediction performance for future samples. In particular, to avoid inflated evaluation of prediction performance from a large screening search on many candidate models, feature selection must be simultaneously performed with classification model construction on a training set under a particular classification method. Evaluation of prediction performance should then be carefully conducted among the extremely large number of competing models, especially in using appropriate performance selection criteria and in utilizing the whole data for model training and evaluation.

32.5.1 Measures for Classification Model Performance

Several different measures are currently used to evaluate the performance of classification models: classification error rate, area under the receiver operating characteristics curve (AUC), and the product of posterior classification probabilities [32.40–42]. However, when a large number of candidate models, e.g., more than 10^8 two-gene models on 10 k array data, are compared in their performance, these measures are often quickly saturated; their maximum levels are achieved by many competing models, so that identification of the best prediction model among them is extremely difficult. Furthermore, these measures cannot capture an important aspect of classification model performance as follows. Suppose three samples are classified using two classification models (or rules). Suppose also that one model provides correct posterior classification probabilities 0.8, 0.9, and 0.4, and the other 0.8, 0.8, and 0.4 for the three samples. Assuming these were unbiased estimates of classification error probabilities (on future data), the former model would be preferred because this model will perform better in terms of the expected number of correctly classified samples in future data. Note that the two models provide the same misclassification error rate, 1/3. This aspect of classification performance cannot be captured by evaluating the commonly used error rate or AUC criteria, which simply add one count for each correctly classified sample ignoring its degree of classification error probability.

To overcome this limitation, the so-called *misclassification penalized posterior* (MiPP) criterion has recently been suggested [32.43]. This measure is the sum of the correct-classification (posterior) probabilities of correctly classified samples subtracted by the sum of the misclassification (posterior) probabilities of misclassified samples. Suppose there are m classes from populations π_i $(i = 1, \ldots, m)$ and a total of $N = \sum_{i=1}^{m} n_i$ samples. Let $\boldsymbol{X}_{ij}$, $j = 1, \ldots, n_i$, be the j-th sample vector from the i-th class under a particular prediction model (e.g., one-gene or two-gene model), denoted as R_{M} and a rule R, e.g., linear discriminant analysis (LDA) or SVMs. For sample vector $\boldsymbol{X}_{ij}$, the posterior classification probability to be assigned to class k (under R_{M}) is defined as $p_k(\boldsymbol{X}_{ij}) = P\left(\boldsymbol{X}_{ij} \in \pi_k | \boldsymbol{X}_{ij}\right)$. (We omit the notation R_{M} for simplicity.) For example, $p_k(\boldsymbol{X}_{kj})$ is thus the posterior probability of correct classification for the sample $\boldsymbol{X}_{kj}$. MiPP is then defined as:

$$\psi_p = \sum_{\text{correct}} p_k(\boldsymbol{X}_{kj}) - \sum_{\text{wrong}} \left[1 - p_k(\boldsymbol{X}_{kj})\right] . \tag{32.10}$$

Here correct and wrong correspond to the samples that are correctly and incorrectly classified, respectively. In the two-class problem, *correct* simply means $p_k(\boldsymbol{X}_{kj}) > 0.5$, but in general, it occurs when $p_k(\boldsymbol{X}_{kj}) = \max_{i=1,\ldots,m}[p_i(\boldsymbol{X}_{kj})]$.

It can be shown that MiPP is simply the sum of the posterior probabilities of correct classification penalized by the number of misclassified samples (N_{M})

$$\psi_p = \sum_{\text{correct}} p_k(\boldsymbol{X}_{kj}) + \sum_{\text{wrong}} p_k(\boldsymbol{X}_{kj}) - \sum_{\text{wrong}} 1 = \sum p_k(\boldsymbol{X}_{kj}) - N_{\mathrm{M}} . \tag{32.11}$$

That is, MiPP increases as the sum of correct-classification posterior probabilities increases, as the number of misclassified samples decreases, or both. Thus, MiPP is a continuous measure of classification performance that takes into account both the degree of classification certainty and the error rate, and is sensitive enough to distinguish subtle differences in prediction performance among many competing models. MiPP can be directly derived from the posterior classification probabilities of class membership in LDA, quadratic discriminant analysis (QDA), and logistic regression (LR), but it is slightly differently obtained for SVMs because they do not directly provide an estimate of posterior classification probability. In this case, a logit-link-based estimation can be used to derive a *pseudo* posterior classification probabilities as suggested by [32.44].

32.5.2 Classification Modeling

As described above, several classification modeling approaches are currently used in genomic data analysis. These approaches often adopt certain cross-validation techniques, such as leave-one-out or training-and-validation-set strategies for their modeling search and fitting.

Gene Voting

Adopting the idea of aggregating power by multiple predictors, the so-called *gene voting* classification method has been proposed for the prediction of subclasses of acute leukemia patients observed by microarray gene-expression data [32.2]. This method gains accuracy by

aggregating predictors built from a learning set and by casting their voting weights. For binary classification, each gene casts a vote for class 1 or 2 among p samples, and the votes are aggregated over genes. For gene g_j the vote is

$$v_j = a_j(g_j - b_j) ,$$

where $a_j = (\hat{\mu}_1 - \hat{\mu}_2)/(\hat{\sigma}_1 + \hat{\sigma}_2)$ and $b_j = (\hat{\mu}_1 + \hat{\mu}_2)/2$. Using this method based on 50 gene predictors, [32.2] has correctly classified 36 of 38 patients in an independent validation set between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

LDA and QDA

Linear discriminant analysis can be applied with leave-one-out classification as follow. Assume each of $f_k(x)$, $k = 1, \ldots, K$, follows a multivariate normal (μ_k, Σ) distribution with mean vector μ_k a common variance–covariance matrix Σ. Then,

$$\begin{aligned}
&\log Pr(G = k|X = x)/Pr(G = j|X = x) \\
&\quad = \log[f_k(x)/f_j(x)] + \log(\pi_k/\pi_j) \\
&\quad = \log(\pi_k/\pi_j) - \frac{1}{2}(\mu_k + \mu_j)^{\mathrm{T}} \Sigma^{-1}(\mu_k - \mu_j) \\
&\qquad + x^{\mathrm{T}} \Sigma^{-1}(\mu_k - \mu_j) .
\end{aligned}$$

A sample vector x_o will then be allocated to group k if the above equation is greater than zero or to group j otherwise.

The quadratic discriminant analysis can be similarly performed except that the variance–covariance matrix Σ is now considered differently for each subpopulation group. The differences between LDA and QDA are typically small, especially if polynomial factors are considered in LDA. In general, QDA requires more observations to estimate each variance–covariance matrix for each class. LDA and QDA have consistently shown high performance not because the data likely from Gaussian distributions, but more likely because simple boundaries such as linear or quadratic are sufficientto define the different classes in the data [32.42].

Logistic Regression (LR)

LR discriminant analysis requires less assumptions than the aforementioned LDA and QDA approaches. LR methods simply maximize the conditional likelihood $Pr(G = k|X)$, typically by a Newton–Raphson algorithm [32.45]. The allocation decision on a sample vector x_o by LR is based on the logit regression fit:

$$\mathrm{Logit}(p_i) = \log[p_i/(1 - p_i)] \sim \hat{\beta}^{\mathrm{T}} x ,$$

where $\hat{\beta}$ is the LR estimated coefficient vector for the microarray data. LR discriminate analysis is often used due to its flexible assumption about the underlying distribution, but if it is actually applied to a Gaussian distribution, LR shows a loss of 30% efficiency in the (misclassification) error rate compared to LDA.

Support Vector Machines (SVMs)

SVMs separate a given set of binary labeled training data with a hyperplane that is maximally distant from them; this is known as the maximal margin hyperplane [32.35]. Based on a kernel, such as a polynomial of dot products, the current data space will be embedded in a higher-dimensional space. The commonly used kernels are:

- Radial basis kernel: $K(x, y) = \exp\left(-\frac{|x-y|^2}{2\sigma^2}\right)$,
- Polynomial kernel: $K(x, y) = < x, y >^d$ or $K(x, y) = (< x, y > +c)^d$,

where $<, >$ denotes the inner-product operation. Note that the above polynomial kernel is of order d and is linear when $d = 1$. Using a training set, we derive a hyperplane with maximal separation and validate against a validation set.

SVMs often consider linear classifiers:

$$f_{w,b}(x) = < w, x > +b ,$$

which lead to linear prediction rules: $h_{w,b}(x) = \mathrm{sign}[f_{w,b}(x)]$ for the decision boundary of the hyperplane $f_{w,b}(x)$. SVMs maps each vector-valued example into a feature space:

$$x \to [\psi_1(x), \psi_2(x), \ldots, \psi_N(x)] .$$

32.5.3 Stepwise Cross-Validated Discriminant Analysis

Classification techniques must be carefully applied in prediction model training on genomic data. In particular, if all the samples are used both for model search/training and for model evaluation in a large screening search for classification models, a serious selection bias is inevitably introduced [32.46]. In order to avoid such a pitfall, a stepwise (leave-one-out) cross-validated discriminant procedure (SCVD) that gradually adds genes to the training set has been suggested [32.42, 47]. It is typically found that the prediction performance is continuously improved (or not decreased) by adding more features into the model. This is again due to a sequential search and selection

strategy against an astronomically large number of candidate models; some of them can show over-optimistic prediction performance for a particular training set by chance. Note also that even though a leave-one-out or similar cross-validation strategy is used in this search, the number of candidate models is too big to eliminate many random ones that survived from such a specific cross-validation strategy by chance. Thus, test data should be completely independent from the training data to obtain an unbiased estimate of each model's performance.

The SCVD Procedure

Using the MiPP criterion above, the SCVD classification model is constructed sequentially as follows. Given a classification rule R, the models are constructed on a training data set in a forward stepwise cross-validated discriminant fashion. Suppose we have a training data set consisting of N samples and g candidate features (genes). A schematic summary of the MiPP-based SCVD model construction is shown in Fig. 32.4.

The initial step begins by fitting each feature individually on the training set. For each of the G features, MiPP is calculated based on leave-one-out cross-validation (so MiPP for a gene is the average of MiPPs of the N leave-one-out fits for that particular gene). The gene with the maximal value of MiPP is then retained, and the optimal one-gene model O_1 is fit using all training samples. The second step adds each of the $G-1$ features and of these $G-1$ two-gene models, the one with the maximal value of MiPP is similarly retained and used to construct the optimal two-gene model O_2. This process continues adding features in this sequential fashion until the training model becomes saturated at the L-th step, i. e., MiPP converges to a certain maximum level and the L-gene MiPP is not bigger than the (L-1)-gene MiPP (note that MiPP has an upper bound of N).

Because of the sequential selection of features in this model construction, the performance of the prediction model improves when there a large number of features in a model and this cannot be used as an objective measure of classification performance. Therefore, the performance of each of the optimal models $O_1, \ldots, O_L$ is assessed on a completely independent test data set to determine the final robust prediction model. In this case, both MiPP and the error rate can be evaluated since the latter can be used among the small number of competing optimal models with different numbers of model features.

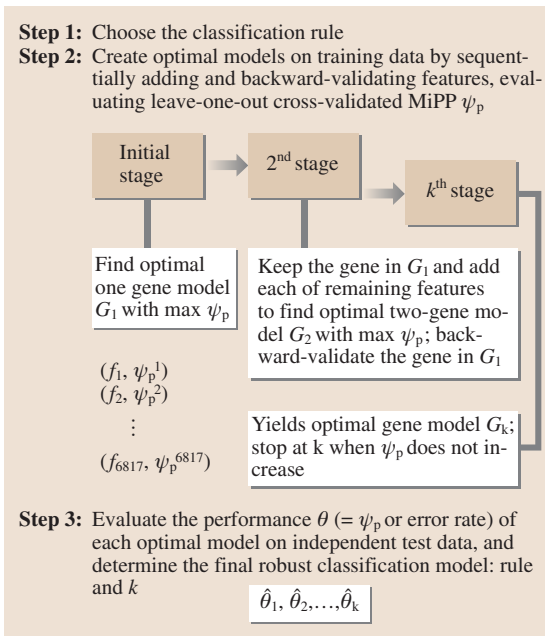

Fig. 32.4 A schematic diagram for classification modeling based on the stepwise cross-validated discriminant (SCVD) procedure

Comparison of Classification Methods

Using this SCVD strategy based on MiPP, several widely used classification rules such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression (LR), and support vector machines

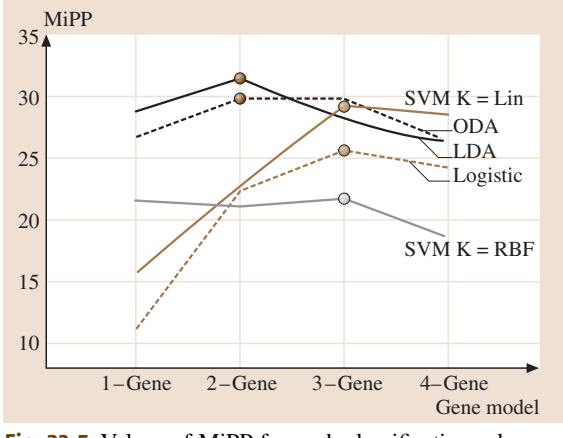

Fig. 32.5 Values of MiPP for each classification rule constructed for models with up to four genes. The best gene model of all the gene models for a given classification rule is denoted by a •

Table 32.2 Classification results of the classification rules and the corresponding gene model

Method	Model	Training data ER%	ψ_p	Test data ER %	ψ_p
LDA	1882+1144	0	37.91	2.94	31.46
QDA	4847+5062	0	37.96	5.88	29.81
Logistic	1807+4211+575	0	37.998	11.76	25.64
SVM K=Linear	2020+4377+1882	0	35.16	0	29.26
SVM K=RBF	4847+3867+6281	0	32.52	5.88	21.713

(SVMs) with linear or radial basis function (RBF) kernels have been compared. The leukemia microarray data in [32.2] had a training set of 27 ALL and 11 AML samples and an independent test set of 20 ALL and 14 AML samples. Since two distinct data sets exist, the model is constructed on the training data and evaluated on the test data set. Figure 32.5 shows the performance of each classification rule on the test data set. Each rule identified a different subset of features and the performance of the best subset for each classification rule along with its performance is shown in Table 32.2. This best subset is simply the point at which each line from Fig. 32.5 reaches its maximum value.

In terms of error rate, it appears as if the SVM with a linear kernel is the most accurate rule. However, LDA only misclassified one sample and the SVM with the RBF kernel and QDA misclassified two samples on the independent test data. Logistic regression does not seem to perform as well as the other rules, by misclassifying 4 out of 34 samples. Note again that comparing the rules on the basis of MiPP is somewhat tricky for SVMs since the estimated probabilities of correct classification from SVMs are based upon how far samples are from a decision boundary. As a result, unlike the LDA, QDA, and LR cases, these are not true posterior classification probabilities. In an application to a different microarray study on colon cancer, the RBF-kernel SVM model with three genes was found to perform best among these classification techniques.

Therefore, using the MiPP-based SCVS procedure, the most parsimonious classification models were derived with a very small number of features, only two or three genes from microarray data, outperforming many previous models with 50–100 features. This may imply that a set of a small number of genes may be sufficient to explain the discriminativeinformation of different types of a particular desease, even though it is often found that there exist multiple sets (of small numbers of genes) with similar classification prediction performance.

References

32.1 C. Sander: Genomic medicine and the future of health care, **287**, 1977–8 (2000)

32.2 T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science **286**, 5439 (1999)

32.3 J. K. Lee, U. Scherf, K. J. Bussey, F. G. Gwadry, W. Reinhold, G. Riddick, S. L. Pelletier, S. Nishizuka, G. Szakacs, J.-P. Annereau, U. Shankavaram, S. Lababidi, L. H. Smith, M. M. Gottesman, J. N. Weinstein: Comparing cDNA, oligonucleotide array data: Concordance of gene expression across platforms for the NCI-60 cancer cell lines, Genome Biol. **4**, R82 (2003)

32.4 D. Pinkel: Cancer cells, chemotherapy, gene clusters, Nat. Genet. **24**, 208–9 (2000)

32.5 J. K. Lee: Discovery, validation of microarray gene expression patterns, LabMedica Int. **19**, 8–10 (2002)

32.6 C. J. Stoeckert, H. C. Causton, C. A. Ball: Microarray databases: standards, ontologies, Nat. Genet. **32**, 469–473 (2002)

32.7 M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein: Cluster analysis, display of genome-wide expression patterns, Proc. Nat. Acad. Sci. **95**, 14863–8 (1998)

32.8 P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, T. R. Golub: Interpreting patterns of gene expression with self-organizing maps: Methods, application to hematopoietic differentiation, Proc. Nath. Acad. Sci. **96**, 2907–2912 (1999)

32.9 S. Dudoit, Y. H. Yang, M. J. Callow, T. P. Speed: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Stat. Sin. **12**, 111–139 (2002)

32.10 V. Tusher, R. Tibshirani, C. Chu: Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, Proc. Nat. Acad. Sci. **98**, 5116–21 (2001)

32.11 Y. Benjamini, Y. Hochberg: Controlling the false discovery rate: a practical, powerful approach to multiple testing, J. R. Stat. Soc., Ser. B, Methodological **57**, 289–300 (1995)

32.12 J. Storey, R. Tibshirani: SAM thresholding, false discovery rates for detecting differential gene expression in DNA microarrays. In: *The Analysis of Gene Expression Data: Methods and Software*, ed. by G. Parmigiani, E. S. Garrett, R. A. Irizarry, S. L. Zeger (Springer, Berlin Heidelberg New York 2003) Chap. 12

32.13 N. Jain, K. Ley, J. Thatte, M. O'Connell, J. K. Lee: Local pooled error test for identifying differentially expressed genes with asmall number of replicated microarrays, Bioinformatics **19**, 1945–51 (2003)

32.14 W. Jin, R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, G. Gibson: The contributions of sex, genotype, age to transcriptional variance in Drosophila melanogaster, Nat. Genet. **29**, 389–395 (2001)

32.15 A. Kamb, A. Ramaswami: A simple method for statistical analysis of intensity differences in microarray-derived gene expression data, BMC Biotechnol. **1**, 1–8 (2001)

32.16 R. Nadon, P. Shi, A. Skandalis, E. Woody, H. Hubschle, E. Susko, P. Ramm, N. Rghei: Statistical inference methods for gene expression arrays, BIOS **4266**, 46–55 (2001)

32.17 B. Durbin, J. Hardin, D. Hawkins, D. Rocke: A variance-stabilizing transformation for gene-expression microarray data, Bioinformatics **18**, 1105 (2002)

32.18 X. Huang, W. Pan: Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays, Funct. Integr. Genomics **2**, 126–133 (2002)

32.19 Y. Lin, S. T. Nadler, A. D. Attie, B. S. Yandell: Adaptive gene picking with microarray data: detecting important low abundance signals. In: *The Analysis of Gene Expression Data: Methods and Software*, ed. by G. Parmigiani, E. S. Garrett, R. A. Irizarry, S. L. Zeger (Springer, Berlin Heidelberg New York 2003) Chap. 13 (http://www.stat.wisc.edu/~yilin/)

32.20 I. Lönnstedt, T. P. Speed: Replicated microarray data, Stat. Sin. **12**, 31–46 (2002)

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

32.22 J. K. Lee, M. O'Connell: An S-PLUS library for the analysis of differential expression. In: *The Analysis of Gene Expression Data: Methods and Software*, ed. by G. Parmigiani, E. S. Garrett, R. A. Irizarry, S. L. Zeger (Springer, Berlin Heidelberg New York 2003) Chap. 7

32.23 M. K. Kerr, G. A. Churchill: Statistical design, the analysis of gene expression microarray data, Genetic Res. **77**, 123–128 (2001)

32.24 R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, R. S. Pales: Assessing gene significance from cDNA microarray expression data via mixed models, J. Comput. Biol. **8**, 37–52 (2001)

32.25 M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, K. W. Tsui: On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, J. Comp. Biol. **8**, 37–52 (2001)

32.26 J. G. Ibrahim, M.-H. Chen, R. J. Gray: Bayesian models for gene expression with DNA microarray data, J. Am. Stat. Assoc. **97**, 88–99 (2002)

32.27 H. J. Cho, J. K. Lee: Hierarchical error model for analyzing gene expression data, Bioinformatics **20**, 2016–2025 (2004)

32.28 B. Efron, R. Tibshirani, J. D. Storey, V. Tusher: Empirical bayes analysis of a microarray experiment, J. Am. Stat. Assoc. **96**, 1151–1160 (2001)

32.29 M. A. Newton, C. K. Kendziorski: Parametric empirical bayes methods for microarrays. In: *The Analysis of Gene Expression Data: Methods and Software*, ed. by G. Parmigiani, E. S. Garrett, R. A. Irizarry, S. L. Zeger (Springer, Berlin Heidelberg New York 2003)

32.30 T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, P. Brown: 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns, Genome Biol. **1**, Research03 (2000)

32.31 G. C. Tseng, W. H. Wong: Tight clustering: a resampling-based approach for identifying stable and tight patterns in data, Biometrics **61**(1), 10–16 (2004)

32.32 U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine: Broad patterns of gene expression revealed by clustering analysis of tumor, normal colon tissues probed by oligonucleotide arrays, Proc. Nath. Acid. Sci. **96**, 6745–6750 (1999)

32.33 M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. R. Marks, J. R. Nevins: Prediction the clinical status of human breast cancer by using gene expression profiles, Proc. Nath. Acad. Sci. **98**, 11462–11467 (2001)

32.34 J. Staunton, D. Slonim, P. Tanamo, M. Angelo, J. Park, U. Scherf, J. K. Lee, W. Reinhold, J. Weinstein, J. Mesirov, E. Lander, T. Golub: Chemosensitivity prediction by transcriptional profiling, Proc. Natl. Acad. Sci **11;98**(19), 10787–10792 (2001)

32.35 T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, D. Haussler: Support vector ma-

chine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics **16**, 906–914 (2000)

32.36 S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. P. Mesirov, T. Poggio: *Support Vector Machine Classification of Microarray Data* (MIT, Cambridge 1998)

32.37 D. V. Nguyen, D. M. Rocke: Tumor classification by partial least squares using microarray gene expression data, Bioinformatics **18**, 39–50 (2002)

32.38 L. Li, C. R. Weinberg, T. A. Darden, L. G. Pedersen: Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method, Bioinformatics **17**, 1131–1142 (2001)

32.39 A. C. Culhane, G. Perriere, E. C. Considine, T. G. Cotter, D. G. Higgins: Between-group analysis of microarray data, Bioinformatics **18**, 1600–1608 (2002)

32.40 A. P. Bradley: The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recog. **30**, 1145–1159 (1997)

32.41 D. J. Hand: *Construction and Assessment of Classification Rules* (Wiley, Chichester 1997)

32.42 M. Soukup, J. K. Lee: Developing optimal prediction models for cancer classification using gene expression data, J. Bioinf. Comp. Biol. **1**, 681–694 (2004)

32.43 M. Soukup: Robust optimization of classification model for predicting human disease subtypes using microarray gene expression data. Ph.D. Thesis (University of Virginia, Charlottesville 2004)

32.44 G. Wahba: Support vector machines, reporoducing Kenel Hilbert spaces, the randomized GACV. In: *Advances in Kernel Methods-Support Vector Learning*, ed. by B. Scholkopf, C. J. C. Burges, A. J. Smola (MIT Press, Cambridge 1999) pp. 69–88

32.45 F. C. Pampel: *Logistic Regression: A Primer.*, Sage Univ. Papers Ser. Quant. Appl. Social Sci. (Thousand Oaks, Sage 2000) pp. 07–132

32.46 C. Ambroise, G. J. McLachlan: Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Nath. Acid. Sci. **10**, 6562–6566 (2002)

32.47 M. Soukup, H. Cho, J. K. Lee: Robust classification modeling on microarary data using misclassification penalized posterior, Bioinformatics **21**(1), i423–i430 (2005)