ELSEVIER

Chemometrics and Intelligent Laboratory Systems 62 (2002) 199–216

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Computational analysis of microarray gene expression profiles: clustering, classification, and beyond

Jie Liang*, Seman Kachalo

*Department of Bioengineering, SEO, MC-063, University of Illinois at Chicago, 851 S. Morgan Street,
Room 218, Chicago, IL 60607-7052, USA*

Accepted 27 February 2002

Abstract

Gene array studies can assess the global expression patterns of thousands of genes under multiple conditions. This technology can provide important insights about the underlying genetic causes of many important biological questions, and can change our understanding of diseases, ultimately allowing the development of novel chemical entities as potential drug candidates. The informatics analysis and integration of gene expression pattern are critical for interpreting gene array studies. In this paper, we discuss the computational analysis of three important tasks: (1) the identification of differentially expressed genes, (2) the discovery of gene clusters, and (3) the classification of biological samples. In addition, we discuss how gene sequence and chemical structures can be profitably combined with microarray studies. Detailed examples are given throughout. Programs written in open source R language for achieving each of these tasks are freely available at `gila.engr.uic.edu/genex.`
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Microarray gene expression profiles; Computational analysis; Clustering

1. Introduction

Global gene expression profiling using microarrays is emerging as a key technology for understanding fundamental biology of gene function, development, and for discovering new classes of diseases such as cancer and for understanding their molecular pharmacology. [1–5] Microarray or microchip is a chip made of glass or other solid material, with an array of tiny DNA spots placed on it. Each spot contains fragments of DNA or RNA molecule whose sequence is predefined and corresponds to portions of a particular gene. The lengths of these fragments may vary from about 20 nucleotides in oligonucleotide microarrays to thousands of nucleotides in genome microarrays. Typical microarray contains several thousand spots on the surface of a quarter square inch, and a library of thousands of genes is placed on a single chip. To probe the global gene expression levels of many genes in biological samples such as cell lines, tissue extracts, or laser microdissected cells, messenger RNAs are first extracted. mRNAs are then reverse-transcribed into cDNA. The amount of cDNA produced are then amplified by polymerase chain reactions (PCR), a standard molecular biology technique that is based

* Corresponding author. Tel.: +1-312-355-1789; fax: +1-312-996-5921.
E-mail address: jliang@uic.edu (J. Liang).

PII: S0169-7439(02)00014-X

on primer extension reaction to amplify specific nucleic acid sequences in vitro. PCR allows a short stretch of DNA to be amplified up to a million folds, so that there is enough quantity of specific DNA molecules to be placed on the microarray chip. In the amplification process, radioactive or fluorescent nucleotides are incorporated. Under proper experimental conditions, the concentrations of the cDNA for a specific gene reflect the amount of expressed mRNA of this gene in the probe sample. Upon hybridization to the DNA fragments on a chip, the labeled probe produces a signal. The expression levels of individual genes present on the chip can then be measured quantitatively by laser scanning (fluorescent probes) or by phosphoimaging (radiolabelled probes) spots at different predefined locations on the chip. In another experiment scheme, the extracts from the sample tissue and a control tissue are marked with different dyes, and are hybridized simultaneously on the same chip. This approach provides information about the relative concentration of expressed RNA in sample and control tissues. As probe samples collected at different well-designed experimental conditions are applied, the relative expression levels of all genes on the chip can then be analyzed for changes in the expression patterns to obtain an integrated global picture about the underlying genetic networks [6–10].

Microarray studies often generate massive amounts of data, which are difficult to be exhaustively examined by hand. Bioinformatics analysis and interpretation to extract genetic patterns from these data are therefore essential for gaining biological insights from experiments. In a recent study [2], a cluster of genes associated with cell proliferation was identified from the expressions of 5000 genes in a mammary epithelial cell line under various experimental perturbations (i.e. varying conditions of TGF-β1, EGF, and IFN-α, γ). This "proliferation cluster" was recapitulated from analyzing gene expression profiles of human breast tumors, and was found highly expressed. This gene cluster contains human homologs of yeast CDC47 gene, cyclin B1, and antigen Ki-67. The utility of computational analysis such as clustering, classification and feature selection is demonstrated by another recent study, where subtypes of acute myeloid and acute lymphoblastic leukemia are successfully discovered without employing any prior biological knowledge. In addition, samples of leukemia tumor can be assigned to either of the two subtypes accurately based on the expression patterns of a selected subset of 50 genes [1].

In this article, we describe computational methods for several common tasks in microarray studies: (1) identifying genes that experience significant changes in expression under different experimental conditions; (2) clustering of genes to identify groups of genes that are likely to be co-regulated or participating in related metabolic and regulatory pathways, (3) predicting and classifying experimental samples whether they belong to a particular type of tissue, disease or phenotype classes, and (4) identification of candidate marker genes or marker gene clusters indicative of specific phenotypes. Finally, we discuss how microarray analysis can be combined with cheminformatic studies and high-throughput drug screening to draw interesting inferences on pharmacological mechanism of drugs.

2. Identifying differentially expressed genes

To identify genes differentially expressed under different conditions from cDNA microarray experiments, a heuristic approach frequently applied is to examine the ratio of fold increase/decrease of the expression levels of a gene. If the ratio is above a predefined cut-off threshold (e.g. three- or five-fold change), these genes are declared to be differentially expressed, and are selected for further experimental validation [10]. Although convenient, this approach is problematic, because the cut-off value is set rather arbitrarily, and it is difficult to assess the rate of false positives (unchanged genes declared differentially expressed) and rate of false negatives (missed differentially expressed genes). We discuss two statistical methods that can be used in conjunction with permutation tests to identify differentially expressed genes.

We begin with the lay-out of microarray data. Data from gene expression experiments can be organized as a matrix. Here, each row represents the hybridization results for a single gene across different conditions, and each column represents the measured expression levels of all genes for one condition. To draw statistical inference, it is essential to have replicated samples for each experimental condition [11]. For identification of differentially expressed genes, we can test against the following null hypothesis: the mean expression levels x_i of gene i under conditions 1 and 2 are the same. Here, we assume that there are r_1 replicate samples for condition 1 and r_2 replicate samples for condition 2.

t-Test. Student's *t*-test is a simple method for testing whether the distributions of two variables are identical. Provided that gene expression levels under two different experimental conditions have identical Gaussian distributions, the statistics

$$t_i = (\bar{x}_{2i} - \bar{x}_{1i}) \bigg/ \sqrt{\frac{r_1 S_{1i}^2 + r_2 S_{2i}^2}{r_1 + r_2 - 2}\left(\frac{1}{r_1} + \frac{1}{r_2}\right)}$$

follows a Student's *t*-distribution, with $r_1 + r_2 - 2$ degrees of freedom. Here, $\bar{x}_{1i}$ and $\bar{x}_{2i}$ are the mean expression levels of gene i in the r_1 replicated samples of condition 1 and r_2 replicated samples of condition 2, respectively; $S_{1i}^2$and $S_{2i}^2$are the sample variances of gene i under these two conditions:

$$S_i^2 = \sum (x_i - \bar{x}_i)^2 / r.$$

If t_i exceeds the threshold value for a specific confidence level (e.g. 95%), the expression levels of gene i at conditions 1 and 2 will then be considered to be different.

Although a large t_i value indicates that the expression levels of gene i are different under conditions 1 and 2, one cannot assume the distribution of gene expression level is Gaussian or the statistic t follows a *t*-distribution, and therefore, one cannot obtain direct estimates of statistical confidence intervals from standard tables of *t*-distributions.

With multiplicative samples, permutation tests can be applied to assess the statistical significance of the observed *t*-statistic. Suppose we lost the labels of conditions and do not know whether an observed value x_i for gene i comes from the samples of condition 1 or condition 2. We randomly divide the samples into group 1 with r_1 samples, and group 2 with r_2 samples. The statistic t can be calculated for this grouping. Altogether there are $\left(\dfrac{r_1 + r_2}{r_1}\right)$ such groupings, and when plausible, we can calculate the *t*-statistic, denoted as t^*, for each of them. An alternative approach is to sample a few thousands of such groupings. The distribution of calculated t^* values can provide an estimation of the p-value p_i^* for the observed value of t. If we let t to be the observed *t*-statistic for gene i, t_k^* the kth permuted sample, R to be the number of permuted samples, we have the estimated p-value for observing t:

$$p_i^* = 2 \times \frac{\min\left(\sum_{k=1}^{R} \#(t_k^* \geq t),\quad \sum_{k=1}^{R} \#(t_k^* \leq t)\right)}{R}$$

See Fig. 1 for an example of permutation *t*-test.

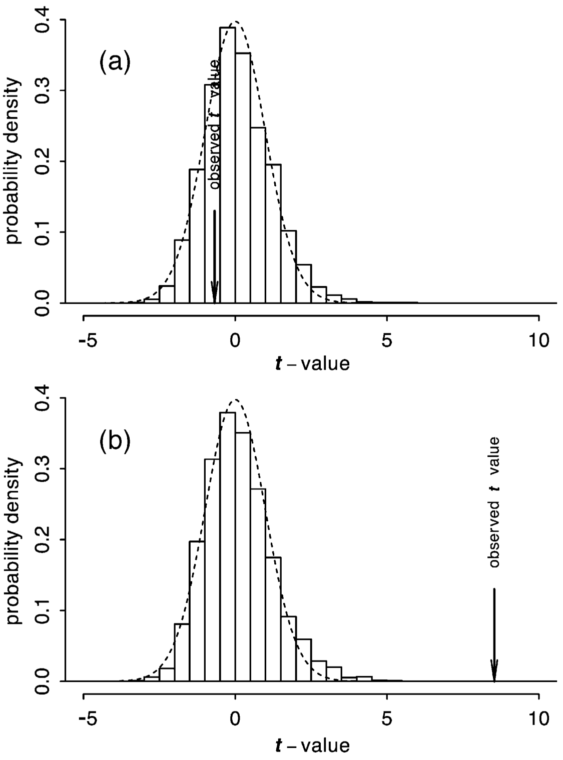

Fig. 1. Identification of differentially expressed gene by permutation *t*-test. This and all subsequent figures are based on expression profiles of 47 samples from acute lymphoblastic leukemia (ALL) patients and 25 samples from acute myeloid leukemia (AML) patients from publicly accessible data set at MIT [1]. The expression levels of 7129 genes are measured for each sample. Histogram shows the relative probability distribution of *t*-statistic values from 3000 permutations of 72 samples. Dotted line shows corresponding Student's *t*-distribution with 70 degrees of freedom. (a) HUMISGF3A Homo sapiens transcription factor ISGF-3 mRNA (an interferon-dependent positive-acting transcription factor that is cytoplasmically activated) does not show any significant differential expression among the two types of samples; (b) AB000449 Homo sapiens mRNA for VRK1 (vaccinia virus B1R kinase related kinase) shows differential expression with a very high statistical significance level.

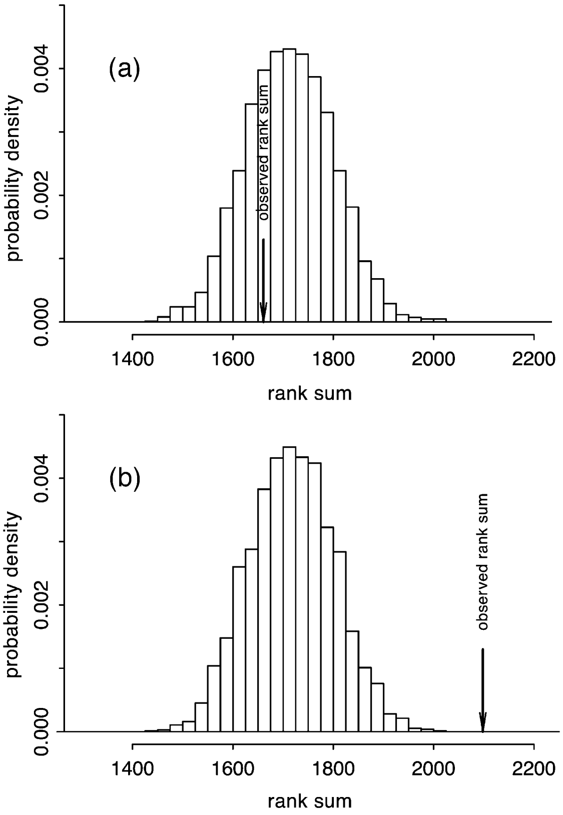

Fig. 2. Identification of differentially expressed gene by Wilcoxon test. The relative probability distribution of rank sum values from 3000 permutations of 47 ALL samples and 25 AML samples from the same leukemia data set as in Fig. 1. (a) HUMISGF3A Homo sapiens transcription factor ISGF-3 mRNA is an interferon-dependent positive-acting transcription factor that is cytoplasmically activated. It does not show any significant differential expression between ALL and AML samples; (b) AB000449 Homo sapiens mRNA for VRK1, a vaccinia virus B1R kinase related kinase, shows differential expression with a very high statistical significance level.

Wilcoxon test. Student's t-test is sensitive to extreme values. It is often safer to use the nonparametric Wilcoxon test when there may be skewness or contamination in the gene expression data. In this test, we assume that x_i is drawn from a symmetric distribution. We combine the r_1+r_2 samples, and rank them in ascending order by their magnitude, and assign each sample the ranks 1, 2,..., r_1+r_2. Next, we sum up the ranks of samples from condition 1, which will be our statistic w. To determine the significance of the p-value, the value of w can then be compared with the null model of the standard distribution of Wilcoxon rank sum values, which can be obtained by the moment generating function $M(t)=\Pi_{i=1}^{r_1+r_2}(e^{-it}+e^{it})/2$ [12], or more conveniently, it can be found in look-up tables in statistics textbooks [13].

With multiplicative samples, the permutation test again is more applicable to assess the statistical significance of the observed w statistic. With R permuted samples, we have the estimated p-value for observing w:

$$p_i^* = 2 \times \frac{\min\left(\sum_{k=1}^{R} \#(w_k \geq w), \quad \sum_{k=1}^{R} \#(w_k \leq w)\right)}{R}$$

See Fig. 2 for an example of Wilcoxon test. Fig. 3 shows the correlation of p-values obtained by the

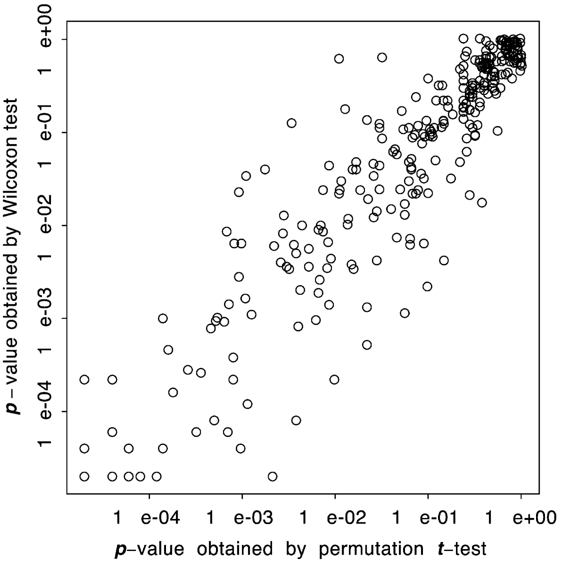

Fig. 3. Comparison of permutation t-test and Wilcoxon test. The plot represents p-values for the first 300 genes from leukemia data set, obtained by 100,000 permutations on 72 samples. p-Values obtained by Wilcoxon test are, in general, consistent with those obtained by permutation t-test, although sometimes they can differ by two orders of magnitude.

permutation *t*-test and by the Wilcoxon test. The *p*-values obtained by these two methods are in general agreement, although they can differ by two orders of magnitude.

3. Pattern discovery: clustering analysis

An important goal in interpreting the large amount of data from cDNA array experiments is to extract the fundamental patterns of gene expression, which are informative of the underlying biology of the samples. Genes with similar expression patterns under various conditions may participate in the same signal pathway or may be co-regulated. As a descriptive tool, clustering of expression patterns can reveal such relationships. The quantitative expression levels of n genes under d conditions can be thought as n points in d-dimensional space. Clustering methods group points together that are close-by in the d-dimensional space. Clustering has been shown to be very effective, in associating gene expression patterns with the ligand specificity of neurotransmitter receptors (Ach, GABA, glutamate, and 5HT) and their functional class (ion channel, G-protein-coupled receptor) [14]. In cancer studies [1,3,15–18], both gene expression "signatures" for cell types (e.g. T cell) and "signatures" for biological processes (e.g. proliferation) have been successfully identified by clustering [5].

3.1. Distance and similarity measure

The "closeness" between genes becomes concrete once a distance measure or similarity measure is defined to quantitatively describe how similar or dissimilar the expression profiles of two genes are. For n genes in the microarray experiment, each pair (x,y) of the $\binom{n}{2}$ pairs of genes can be assessed for their similarity in the expression levels under d condition. A widely used dissimilarity or distance measure is the Euclidean distance:

$$d_2(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

Another convenient measure is the correlation coefficients, which evaluates how correlated the expression levels of genes x and y under d different conditions:

$$R(x,y) = \frac{\sum_{i=1}^{d} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{d} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{d} (y_i - \bar{y})^2}}$$

The value $1 - R(x,y)$ can also be used as a dissimilarity measure.

When distances or correlations for all $\binom{n}{2}$ pairs of genes are calculated, we obtain a $n \times n$ distance or similarity matrix, which can then be used for cluster analysis.

3.2. Agglomerative hierarchical clustering

This method groups genes into a tree or dendrogram [19]. At the beginning, each individual gene forms its own cluster. Starting from n gene clusters, the two clusters with smallest distance are merged, and the clustering is updated. This process is repeated until all clusters are merged into one. The algorithm can be outlined as [20]:

```
Algorithm HierarchicalClustering
repeat
    find two clusters $C_i$ and $C_j$
        where $d(C_i,C_j) = \min_{r \neq s} d(C_r,C_s)$.
    merge $C_i,C_j$ into a single cluster $C_q$.
    replace clusters $C_i,C_j$ with $C_q$.
    update distance matrix of new clusters.
until all genes lie in the same cluster.
```

In order to update the distance matrix when two clusters C_i,C_j are merged into a new cluster C_q, the key question is how to define the distance between the new cluster C_q and all other existing clusters. In the *single linkage* approach, the distance of C_q to another existing cluster C_s is calculated as:

$$d(C_q,C_s) = \min(d(C_i,C_s),\ d(C_j,C_s))$$

where d is the distance or dissimilarity measure used.

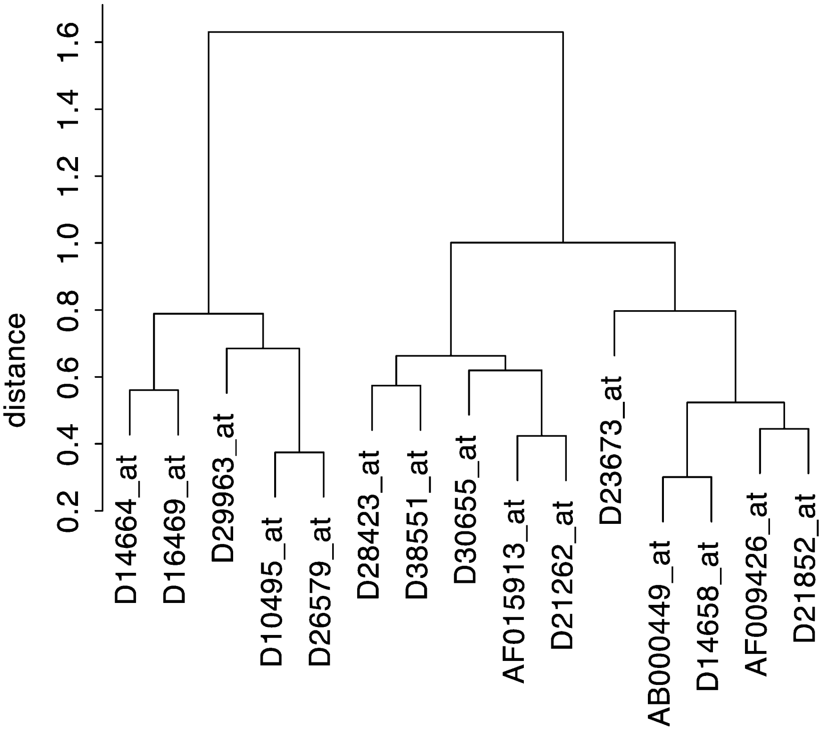

Fig. 4. An example of hierarchical clustering analysis where gene clusters are identified. Here, a hierarchical clustering tree is shown for 15 genes experiencing significant changes in mRNA expression level between the AML and ALL leukemia samples. Genes are grouped together in a tree structure based on their distance in their expression pattern. The distance between two genes is defined as $1 - R(x,y)$. Here, the 15 genes are chosen from the first 15 of the top 300 genes, which all have the same most significant p-values as estimated by Wilcoxon test for difference in mRNA expression levels. The hierarchical clustering is based on an agglomerative nesting algorithm using unweighted pair group average method [21].

In the *complete linkage* approach, the distance is calculated as:

$$d(C_q,C_s) = \max(d(C_i,C_s),\ d(C_j,C_s))$$

In the weighted pair group method average (WPGMA) approach:

$$d(C_q,C_s) = (d(C_i,C_s) + d(C_j,C_s))/2$$

In the unweighted pair group method average (UPGMA) approach:

$$d(C_q,C_s) = a_i \cdot d(C_i,C_s) + a_j \cdot d(C_j,C_s)$$

where $a_i = \frac{|C_i|}{|C_i|+|C_j|}$ and $a_j = \frac{|C_j|}{|C_i|+|C_j|}$. An example of hierarchical clustering using unweighted pair group average method is shown in Fig. 4.

3.3. k-Means clustering

Another widely used clustering method is k-means clustering [22]. It has the advantage that no strict phylogenetic relationship is enforced on every gene, as is in hierarchical clustering [15–17], which can be problematic because there is no absolute ancestral relationship in expression patterns.

In this method, genes are classified as belonging to one of the k clusters. Cluster membership is determined by calculating the centers $a_1,\ldots,a_k \in \mathbb{R}^d$ for each gene cluster, and assigning each gene i according to its expression profile x_i to the cluster with the closest centroid. The goal is to find empirically optimal cluster centers $a_1,\ldots,a_k$ such that the empirical error

$$E = \frac{1}{n}\sum_{i=1}^{n} \min_{1 \le j \le k} \|x_i - a_j\|^2$$

is minimized. This is achieved through an iterative approach:

```
Algorithm $k$-MeansClustering
$i := 0$
Assign $k$ initial centers $a_1^{(0)},\ldots, a_k^{(0)}$ arbitrarily;
Repeat
    cluster genes $x_1,\ldots,x_n$ to $k$ clusters
    for $x_j, j\in[1,\ldots,n]$
        if $\|x_j - a_m\|^2 \le \|x_j - a_l\|^2, l\neq m$
            Assign $x_j$ to the $m$-th cluster
    update cluster centers
    $a_m^{(i+1)} = \Sigma_{j:x_j\in C_m^{(i)}} x_j / |C_m^{(i)}|$
    $i := i+1$
Until no changes in the cluster centers.
```

Fig. 5 shows the results of clustering of 50 differentially expressed genes between AML and ALL samples into five clusters using *k*-means clustering.

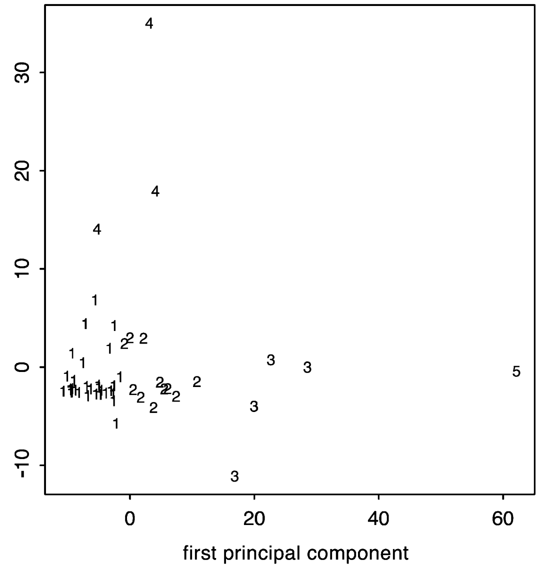

Fig. 5. Results of *k*-means clustering. Fifty differentially expressed genes from the leukemia data set are divided into five clusters. The data are plotted on two first principal components. Each point on the plot represents a single gene and bears the number of the cluster to which the gene belongs.

3.4. Quality control of clustering

An important issue in interpreting clustering results is to assess the quality of the classification of each gene. Here, we discuss two approaches: the *silhouette method* and the *resampling method.*

3.4.1. Silhouette method

This method calculates how well a gene lies within a cluster [23]. Let A denote the cluster to which gene i belongs, C any other cluster, $a(i)$ the average dissimilarity of i to all other genes in cluster A, and $d(i,C)$ the average dissimilarity of i to all genes in C. Dissimilarity can be measured for example by Euclidean distance, or by $1-R$ if the correlation coefficient R is used. After computing $d(i,C)$ for all clusters $C\neq A$, we select the smallest of those: $b(i)=\min_{C\neq A} d(i,C)$. The cluster B which this minimum is attained is the second-best choice for gene i. A silhouette value $s(i)$ for gene i can then be calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \text{ and } \qquad s(i)\in[-1, +1]$$

An $s(i)$ value close to $+1$ means gene i is well-classified, a value close to 0 means i lies between two clusters, and a value close to -1 means gene i is badly clustered. By assessing the silhouette value, the quality of clustering for each gene can be assessed [21,24]. An example of applying silhouette method for assessing the quality of clustering is shown in Fig. 6.

Silhouette method can also be applied to determine the optimal number of clusters k for the k-means method. By systematically changing the number of clusters, the one that maximizes the average $s(i)$ over the whole set of genes or the set of genes of interest can be chosen as the number of clusters.

3.4.2. Bootstrap resampling method

In this approach, we first generate a large number of resampled microarray data from the experimental $n \times d$ data matrix, where n is the number of genes and d the number of conditions. For example, we can generate a resampled data matrix by drawing with replacement n gene rows. The new data matrix may contain some genes multiple times, some genes one time, whereas some other genes may be missing.

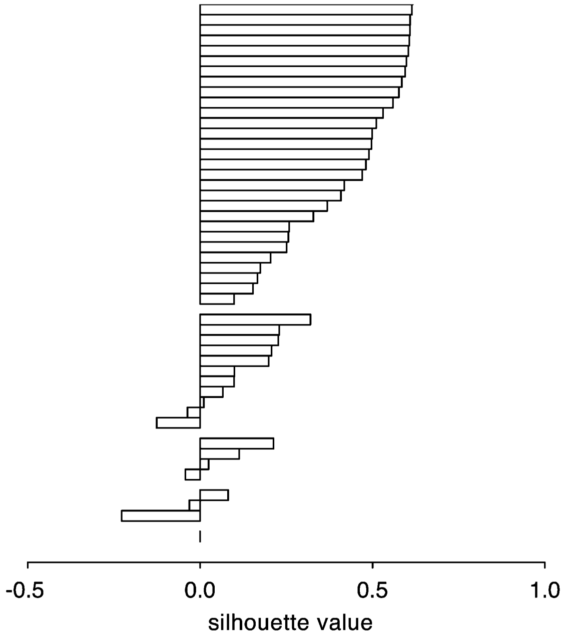

Fig. 6. Silhouette values of 50 genes clustered in Fig. 5. Some genes from clusters 2, 3 and 4 show low confidence of belonging to these clusters. There is no silhouette value for gene from cluster 5, as this cluster consists of only one gene.

Repeat this bootstrap procedure R times, we obtain R resampled data matrices. After applying a clustering method to each data matrix, we have a population of R clusterings of genes.

We analyze the bootstrapped samples as follows. For gene i and gene j, let c_{ij} be the number of times both i and j belong to the same cluster, and n_{ij} the times both i and j appear in the bootstrapped samples. The consistency ratio r is:

$$r_{ij} = \frac{c_{ij}}{n_{ij}}$$

When $r_{ij} = 1$, genes i and j are well clustered together. When $r_{ij} = 0$, these genes are never clustered together. Let A denote the cluster to which gene i belongs, $a(i)$ the average consistency ratio of gene i to all genes in cluster A:

$$a(i) = \sum_{j \in A} r_{ij} / |A|$$

where $|A|$ is the number of genes in cluster A. The average consistency ratio of gene i clustered to all genes in other cluster(s) is:

$$b(i) = \sum_{j \notin A} r_{ij} / (n - |A|)$$

The resampling quality index $q(i)$ for gene i is defined as:

$$q(i) = a(i) - b(i)$$

When $q(i) = 1$, gene i is clustered well. When $q(i)$ is around 0.5, its clustering is questionable. When $q(i) = 0$, the clustering of i is very poor.

An important advantage of the bootstrap quality index $q(i)$ is that it works well regardless of the metrics and clustering method. Fig. 7 shows the quality indices for the same 50 genes clustered in Fig. 5.

Bootstrap is a general method that has been applied in a variety of ways to study microarray expression data. Additional examples of assessing clustering qualities using bootstrap methods can be found in Refs. [25,26]. In Ref. [25], it is assumed that gene

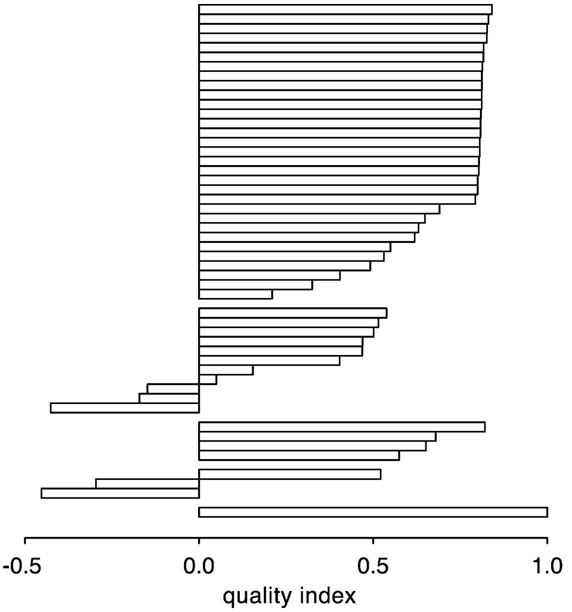

Fig. 7. Resampling quality index values of 50 genes clustered in Fig. 5. Some genes from clusters 2, 3 and 4 show low confidence of belonging to these clusters.

expression levels follows a Gaussian distribution, and resampled data for expression level x_{ij} of gene i under condition j is drawn from a Gaussian distribution $\mathcal{N}(x_{ij},s_{ij})$, where s_{ij} is the estimated variation of x_{ij} from replicated samples. Genes that belong to specific clusters at 95% confidence intervals are identified by clusterings of bootstrapped samples. In Ref. [26], an ANOVA (analysis of variance) model is used to generate the bootstrapped samples. Rather than assuming Gaussian or other parametric distribution models for gene expression levels, resampled data points are generated from independent draws from the studentized residues [27] from the original fit of an ANOVA model of the expression level of a gene.

4. Classifying biological samples: predictors and classifiers

An important application of microarray experiments is to classify biological samples into known classes of phenotypes, as exemplified by a large number of microarray studies in cancer research on tumor classification [1,5]. This is an important approach that may help tumor prognosis and diagnosis. For example, B-cell lymphoma can be classified into two new categories based on their expression patterns, each with marked differences in B-cell differentiation and in overall patient survival rate [5]. This new classification cannot be obtained using standard clinical parameters [5].

Recall that the expression data are organized as a $n \times d$ matrix containing the expression levels of n genes under d different conditions. Assuming that each sample belongs to strictly one of the P phenotypical classes $\mathscr{P}=\{1,\ldots,P\}$, we seek a function f: $\mathbb{R}^n \rightarrow \{1,\ldots,P\}$, that maps sample i to one of the P classes according to the global expression profile $x_i \in \mathbb{R}^n$ of this sample.

Although in some cases, natural classification of samples can be identified by unsupervised methods such as principal component analysis (see Fig. 8 for an example), in most cases, classification requires supervised learning where the examples of different classes are given in a training data set. Among the d samples, the class identification for a subset of size d_t is known, and these samples serve as the *training set* for developing classifier and predictor. The remaining $d - d_t$ samples are the *test set* whose class identifications will be predicted by the classifier. The development of classifier and predictor is an intensely studied area [22,28,29], and many techniques have been developed. We briefly describe a few selected classifiers that are commonly used in microarray analysis.

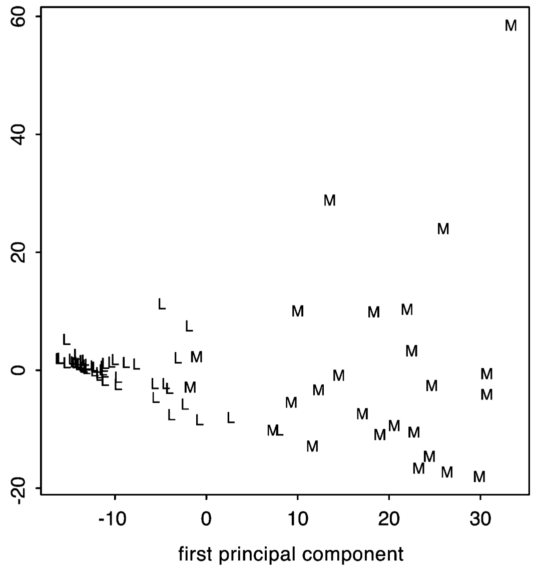

Fig. 8. Class of samples sometimes can be naturally revealed by unsupervised method such as principal component analysis. Here, 72 leukemia samples, 38 in training set and 34 in test set, are all plotted on the first two principal components found by the top 100 differentially expressed genes. These 100 genes are selected by Wilcoxon test from the training data. The AML samples are marked as M, and the ALL samples are marked as L. The AML and ALL samples are well separated in PCA space.

4.1. Classifiers based on Gaussian distribution

We begin with a simple parametric model for describing microarray data. Gaussian distribution is a convenient model for studying a wide variety of physical processes. The probability density function (pdf) of a univariate Gaussian distribution takes the following familiar form:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where μ is the mean and σ^2 is the variance. Its generalization is the multivariate Gaussian distribution $\mathcal{N}(\mu,\Sigma)$ where $\mu\in\mathbb{R}^d$ is the mean vector and Σ is the covariance matrix:

$$\Sigma = \mathbf{E}[(x-\mu)(x-\mu)^T]$$

and its probability density function is:

$$p(x) = \frac{1}{\sqrt{2\pi}\,|\,\Sigma\,|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$
$$x\in\mathbb{R}^d$$

When classifying biological samples where each sample belongs to exactly one of the P classes, we can evaluate the probability that sample j belongs to a specific class K:

$$P(y_j,K) = P(y_j \mid K)\pi(K)$$

here, $y_j\in\mathbb{R}^n$ is the vector representing the global expression profile of all n genes from sample j, i.e. it is a column vector in the $n \times d$ data matrix. $\pi(K)$ is the prior probability that any given sample belongs to class K[1], and $P(y_j\,|\,K)$ is the conditional probability of observing y_j from a sample of class K. Assume that the pdf of $P(y_j\,|\,K)$ is a Gaussian distribution $\mathcal{N}(\mu_K,\Sigma_K)$, several classifiers can be developed with different additional assumptions [20,30].

Quadratic classifier. If we can assess the joint probability $P(y_j,K)$ for the global expression profile of all n genes in condition j for every class $K\in\mathcal{P}$, we can simply classify sample j into the class with the highest probability $P(y_j,K)$. Technically, it is more convenient to work with the log transformed discriminant function g_K:

$$g_K = \ln[P(Y_j,K)\pi(K)]$$

When $P(y_j,K)$ follows a Gaussian distribution,

$$g_K = -\frac{1}{2}(x-\mu_K)^T\sum_K^{-1}(x-\mu_K) + \ln P(K)$$
$$+ \text{constant}.$$

The first term on the right hand side is quadratic. Standard techniques can be applied to calculate this term, for example, by using Moore–Penrose pseudo-inverse [31]. Fig. 9 shows an example of applying quadratic classifier for predicting clinical samples of AML and ALL leukemia from the MIT data set.

Linear classifier. When all classes have the same covariance matrix, i.e. $\Sigma_i=\Sigma$, the discriminant function is:

$$g_K = -\frac{1}{2}(x-\mu_K)^T\sum^{-1}(x-\mu_K) + \ln P(K)$$
$$+ \text{constant}.$$

Here, the quadratic term becomes the same for all classes, and the boundary between classes becomes linear [20]. Fig. 10 shows that linear classifier performs very well.

Diagonal classifier. When the covariance matrices for all classes are the same, and when the expression levels of all genes are uncorrelated, the covariance

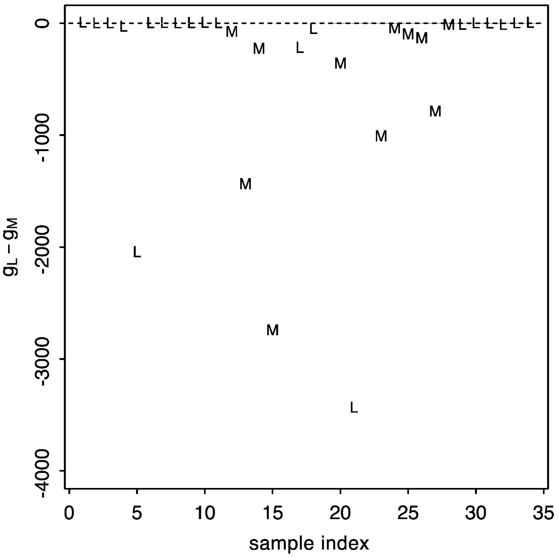

Fig. 9. Classification of clinical samples of AML and ALL leukemia by a quadratic classifier. After training with an independent set of 27 ALL and 11 AML samples from Ref. [1], the test samples of ALL (20, marked L) and AML (14, marked M) are classified. Test samples above the dashed line are predicted as L (ALL), and those below are predicted as M (AML). Quadratic classifier misclassifies many clinical samples. Here, we use only 5 informative genes instead of 50 genes as reported in Ref. [1]. These genes are selected following method described in Section 5.

[1] For example, the MIT Leukemia data set contains 38 bone marrow samples, 27 of them are acute lymphoblastic leukemia (ALL), and 11 are acute myeloid leukemia (AML). The prior probability for AML is estimated as: $\pi(1)=27/38=0.71$, and the prior probability for ALL is: $\pi(2)=11/28=0.39$.

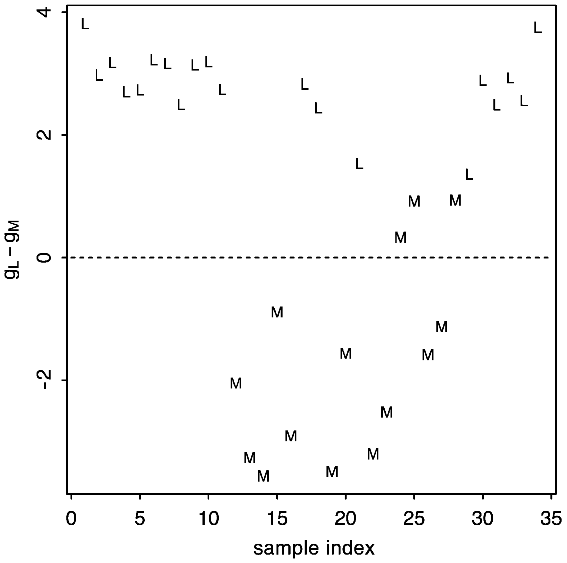

Fig. 10. Classification of clinical samples of AML and ALL leukemia by a linear classifier. The training set and test set are the same as in Fig. 9. Samples above the dashed line are predicted as L (ALL), and those below are predicted as M (AML). Here, we use only the same 5 informative genes as in Fig. 9 instead of 50 genes as reported in Ref. [1], and we achieve better performance (three instead of five misclassifications).

matrices are all diagonal $\Lambda = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$. The discriminant function in this case is simple:

$$g_K = -\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu_{ik})^2}{\sigma_i^2}$$

The "weighted voting scheme" developed in Ref. [1] for classification is a variation of this type of diagonal linear classifier [30]. Fig. 11 shows that diagonal linear classifier performs well.

Which of these classifiers is more preferable? Intuitively, the quadratic classifier is the most sophisticated among the three, and it is provable that it is the optimal classifier for Gaussian distributions. However, results shown in Figs. 9, 10 and 11 indicate that linear and diagonal linear classifiers often outperform the quadratic classifier. This observation has been noted before experimentally [30]. This empirical observation confirms Vapnik's principle of "avoiding solving a more general problem as an intermediate step for solving a problem with restricted amount of information" [32]. In this case, the construction of the quadratic classifier involves estimating mean vectors μ_1, μ_2 and covariance matrices Σ_1,Σ_2, altogether $n(n+3)/2$ parameters. Estimating these parameters with high accuracy is necessary for constructing a good discriminant rule, because the calculation of the inverse matrices Σ_1^{-1} and Σ_2^{-1} are often ill-conditioned. Estimating the high-dimensional covariance matrices requires a large amount of data. In contrast, simpler classifiers, such as the diagonal linear classifier, require only the estimation of order $O(n)$ number of parameters.

4.2. k-Nearest neighbor classifier

This is one of the simplest nonlinear classifiers that has found practical use in many applications. To classify a biological sample j of unknown phenotype, we calculate its distance based on its expression profile y_j to all of the d_t training set samples, where classifications are known. We then look for the k nearest neighbor samples to the d_t training set samples. The class for each of the k nearest neighbors is then identified, and the unknown sample is assigned

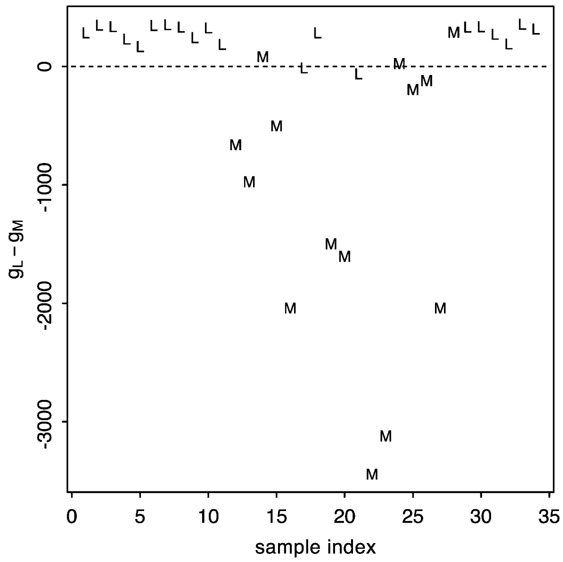

Fig. 11. Classification of clinical samples of AML and ALL leukemia by a diagonal linear classifier. The training set and test set are the same as in Fig. 9. Samples above the dashed line are predicted as L (ALL), and those below are predicted as M (AML). Using only five informative genes, we achieve good performance in classification (five misclassifications).

to the class where the majority of the k neighbors belong (see Fig. 12).

4.3. Support vector machine

When the expression levels of each gene are represented as a point $x_i \in \mathbb{R}^d$, the problem of classification can be viewed as finding the correct label $+1$ ("belonging to a class of interest") or -1 ("not belonging to this class") to each point:

$$x_i \mapsto y_i, \qquad y_i \in \{+1, -1\}$$

The goal of a classifier is to reproduce the mapping $x_i \mapsto y_i$. However, this mapping is unknown, and we can only use the prediction from a classifier $f(x,\alpha)$, such as the quadratic classifier described earlier. Here, α denotes the adjustable parameters of the classifier. If the prediction is wrong, the true label y_i and the predicted label $f(x_i,\alpha)$ will be different: $|y_i - f(x_i,\alpha)| \neq 0$. By adjusting the parameters α, we can minimize the rate of prediction error for a set of training data.

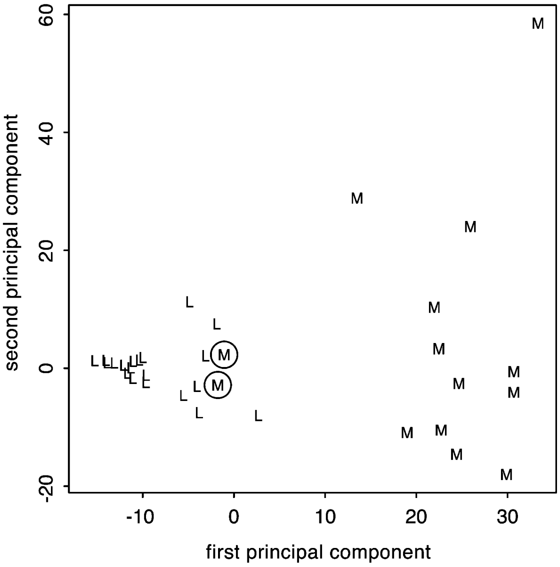

Fig. 12. Classification by 3-nearest neighbor method using the top 100 most differentially expressed genes selected as in Fig. 8. The training set contains 11 AML samples (marked as M) and 27 ALL samples (marked as L). All are classified correctly by the 3NN method. Results of classification of the test set are plotted here. There are 14 AML samples (marked as M) and 20 ALL samples (marked as L) in the test set. Two samples are misclassified by the 3NN method, and are marked by circles. Samples are plotted by the first two principle components.

A fundamental question for classification is the error rate of misclassification for a well-trained classifier when challenged in the future with unseen data. Even though a well-trained classifier can have perfect predictions on the training set, there is no guarantee that it will perform well and will not make many mistakes when future unseen samples are presented. The statistical learning theory developed by Vapnik and Chervonenkis (also called VC theory) [33,34] provides theoretical foundation to address this question. It also suggests the approach of Support Vector Machine, an important class of classifiers that generalize well for unseen data.

We examine the simplest situation, where two classes of samples (disease vs. non-disease) are completely separable by a hyperplane in $\mathbb{R}^d$. This hyperplane has the algebraic form: $w \cdot x + b = 0$, $x, w \in \mathbb{R}^d$. We may be able to find many such hyperplanes. Among these, VC theory shows that we need to look for the hyperplane that maintains the maximum distance (or "margin") to both the closest disease sample point x_i, where the corresponding label $y_i = +1$, and the closest non-disease sample point x_j, where the corresponding label $y_j = -1$. This hyperplane is the unique classifier that makes the least mistakes when future data are presented. Using the machinery of Lagrange multiplier, such a hyperplane can be found by solving the following linearly constrained convex quadratic programming problem [32]:

Maximize

$$L(\alpha) = \Sigma_i \alpha_i - \tfrac{1}{2} \Sigma_{i,j} \alpha_i \alpha_j y_i y_j \cdot x_i x_j$$

with constraints

$$\Sigma_i \alpha_i y_i = 0, \text{ and } \alpha_i \geq 0$$

Because this is an optimization problem on convex set, the solution found is automatically guaranteed to be the global solution. This offers an important advantage not shared by other classifiers such as the neural network, where one often encounters the problem of local optimum in the training phase.

Since the data points x_i and x_j only enter the optimization problem as the inner product, we can replace $x_i \cdot x_j$ with a kernel transformation: $K(x_i, y_j) = \phi(x_i) \cdot \phi(x_j)$.

This amounts to applying a nonlinear mapping ϕ to the input data and project it to a high-dimensional space. If kernels are chosen appropriately, non-separable data points in the original space will become separable in the space of ϕ [35]. Frequently used kernels include the inhomogeneous polynomial kernel, which takes the form of $K(x_i,x_j)=(x_i{\cdot}x_j+1)^p$, and Gaussian radial basis function kernel $K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/(2\sigma^2))$.

Because a large data set is involved, solving the quadratic programming efficiently is crucial in developing an effective SVM classifier. Recent development in subset selection [36] methods such as Sequential Minimal Optimization (SMO) [37] allows practical implementation of SVM for solving classification problems involving very large data sets such as gene array data. In a recent study, Ben-Dor et al. [38] described successful application of SVM with linear and quadratic kernels to classify tumor and normal tissues for colon, ovarian, and bone marrow samples. Brown et al. [39] tested several SVMs using different similarity metrics on gene expression data and found that SVMs have the best performance among several classifiers in successfully identification of sets of genes involved in a common biological function. SVM has also been used for identifying biologically active chemical compounds as drug candidates [40].

5. Identify individual signature genes and groups of signature genes

Recent studies indicate that there are large sets of genes displaying coordinated expression in samples of similar biological phenotype [5]. For example, different classes of malignancies can often be segregated based on the "signatures" of coordinated expressions of clusters of genes. These signatures reveal different biological features of the sample. How do we identify signature genes and gene clusters associated with specific phenotypes?

5.1. Signature genes

We discussed earlier two methods (*t*-test and Wilcoxon test methods) for identifying differentially expressed genes. If samples are divided into two groups of different phenotypes, these methods can also be applied to identify signature genes. Those genes with significant *p*-values can be identified as candidate marker genes. To explore the possibility for these genes as candidate marker genes for tasks such as diagnosis, further bioinformatics analysis and prediction of transmembrane helices [41], signal peptide [42], and subcellular location [43] will be necessary. Those genes located on membranes may be suitable as target proteins for antibodies, and those secreted into serum will be candidate targets for developing diagnostic markers.

5.2. Group of signature genes

If a gene cluster contains many more discriminating genes with *p*-values significantly different from the average, this gene cluster can be identified as a "signature gene cluster".

It is possible that several non-correlated genes collectively may provide strong discriminating signals. To identify such signature genes, we first choose a measure to describe how separable data from different groups are. To define the (class separability) [20], we calculate the within-class scatter matrix S_w, which is the weighted combination of the covariance matrices, S_1 and S_2, for data of groups 1 and 2:

$$S_\mathrm{w}=p_1\Sigma_1+p_2\Sigma_2$$

where $\Sigma_i=\mathbf{E}[(x-\mu_i)(x-\mu_i)^T]$, $i\in 1,2$, $x\in$group i, p_1 and p_2 are the *a priori* probability of groups 1 and 2. We then calculate the mixture scatter matrix S_m:

$$S_\mathrm{m}=\mathbf{E}[(x-\mu)(x-\mu)^T],\ i\in\text{groups 1 and 2}$$

where μ is the mean of the data after combining groups 1 and 2. Our class separability is defined as [20]:

$$J=\mathrm{trace}S_\mathrm{w}^{-1}S_\mathrm{m}$$

This is a generalization of the commonly used Fisher's discriminant ratio, which is in turn similar to the *t*-statistic. J is large when samples are well clustered around their mean within each class, and the clusters of the different classes are separated well.

We can use a simple heuristic *sequential forward seletion* method to select a subset of k genes. Our goal is to maximize class separability J. We initially calculate the J value for each individual gene. After selecting the best gene x_1 with the largest J value, we form all possible pairs that contain x_1 and another gene, and calculate their J values. We pick the winning pair (x_1,x_2) which has again the better J value. This process is repeated until we have found k genes. The five most informative genes for distinguishing ALL and AML samples in our study (Figs. 9, 10 and 11) are chosen using this approach.

6. Beyond gene expression profiles

Gene expression profiling experiments can provide additional biological insight when further integrated and interpreted with other bioinformatics analyses. In this section, we briefly discuss how expression profiling and genomic sequence analysis can help to identify transcription factor binding sites, a fundamental problem in developmental biology and cancer biology. We then discuss how gene expression profiling can help to suggest and clarify mechanisms of drug actions when analyzed in conjunction with drug activity profiling experiments and cheminformatics studies of chemical compounds.

Motif detection of upstream regulatory region. For genes clustered together that share the same expression patterns, it is important to further investigate whether they share the same control elements such as transcriptional regulatory sites in the upstream regions. Genes sharing control elements will likely respond similarly to developmental change and environmental stress. Knowledge of genes controlled by the same transcription factor therefore is critical in understanding the regulatory and metabolic genetic networks.

A basic approach to identify simple control elements in the upstream regions is to enumerate all possible candidate motifs and evaluate statistical significance of their occurrence against a null or random model. In a recent study of yeast genome [44], the observed frequencies of all possible oligonucleotides up to length of 9 were compared to the frequencies expected from a null model, and candidate motifs were identified based on estimated statistical significance. In this study, all sequences in the non-coding region were used to estimate the expected frequencies $F_{\rm e}(b)$ of the null model for each possible oligonucleotide b. The number of expected occurrences for each oligonucleotide b under this null model is: $F_{\rm e}(b)\cdot T$. Here, T is the total number of possible matching positions for an oligonucleotide of length w, across both strands of a set of upstream sequences. Assuming that there are S upstream sequences and all are of the same length L, then for each sequence, there are $L-w+1$ possible matching positions in one of the two directions, and therefore: $T=2\cdot S\cdot(L-w+1)$. The statistical significance can be evaluated using a binomial model. The probability of observing n occurrences of b is:

$$P(b,\text{occur}=n)=\frac{T!}{(T-n)!n!}F_{\rm e}(b)^n(1-F_{\rm e}(b))^{T-n}$$

and the probability to observe n or more occurrences of b is:

$$P(b,\text{occur}\geq n)=1-\sum_{j=0}^{n-1}P(b,\text{occur}=j)$$

Over-represented oligonucleotides in the yeast genome detected by this method are frequently found to be regulatory sequences previously confirmed by experiments. For example, the motif CGTTCC is found to be a control element of the YAP gene family, whose expression levels are induced more than two-fold by the controlled expression of the Yap1p gene in a cDNA microarray study [10]. This approach is very effective in discovering short motifs with a highly conserved core. It has been extended to detect motifs that consist of two trinucleotides separated by a gap [45].

The above approach does not perform well when control elements are long and have higher internal variation. Motif detection algorithms such as those based on Gibbs sampler can be applied to detect more complex control elements. Developed originally for multiple sequence alignment [46–49], different variations have been shown to be successful in detecting complex motifs of control elements of co-expressed genes studied in microarray experiments [50–52]. We refer to Refs. [53,54] for details of Gibbs sampler and general Markov Chain Monte Carlo methods applied in sequence analysis.

Microarray studies and drug discovery. Another important area of microarray research is to combine gene expression profiling with chemical and pharmacological studies such as drug response profiling studies pioneered by researchers at National Cancer Institute (NCI). The goal is to discover novel pharmacological mechanism, and to understand and suggest novel hypotheses of toxicity of chemical compounds, which are critical problems of drug development.

A useful resource is the National Cancer Institute (NCI) compound-cell line database. NCI has screened a large number of chemical compounds to identify potential anticancer drugs. The profiles of sensitivities of 60 cell lines to 70,000 different compounds have been collected and made available publicly. These 60 cell lines are derived from many different types of cancer, and the NCI data therefore provide important information about the mechanism of drug action, resistance and modulation [55].

To analyze the patterns of drug actions, clustering algorithms can be applied to the profiles of drug activities for the 60 cell lines. Strong groupings often were observed for cell lines from different tissue origins. This is different from clustering by gene expression profiles, where samples from the same tissues are usually clustered together. This difference reflects that only the activities of a subset of genes important to drug sensitivity and resistance are relevant. For example, cell lines from different tissues expressing the multi-drug resistance gene MDR1 all have similar drug activity profiles [55].

Clustering by drug activity profiles can reveal different drug mechanisms. For example, drugs inhibiting tubulin monomer polymerization and drugs inhibiting tubulin depolymerization are found to belong to two different clusters. It can also suggest new mechanisms of drug activity. Fore example, 5-fluorouracil (5-FU) is an antimetabolite to treat colorectal and breast cancer. It can act on both DNA and RNA, but the clustering with RNA synthesis inhibitors suggests that its dominant mechanism is likely to be inhibition of RNA synthesis.

The gene expression profiles of the 60 cell lines [55] can be analyzed together with profiles of drug activities, with the goal to relate changes in gene expression to drug sensitivities. For each pair of chemical compound and gene from the set of *n* genes and the set of *m* chemical compounds, the correlation of drug activity profile across the 60 cell lines and the gene expression profile across the same 60 cell lines can be assessed. For each compound, there are *n* such drug–gene correlation coefficients. These correlation coefficients can be used to cluster the *m* compounds. Such combined drug activity and gene expression profile analysis can suggest a causal relationship between gene and drug. In the case of 5-FU and dihydropyrimidine dehydrogenase (DPYD), strong causal relationships are suggested based on considerable experimental evidence and significant negative correlation between the expression of DPYD and the potency of 5-FU against the 60 cell lines, accumulated experimental knowledge [55]. Another causal relationship suggested is the lack of expression of asparagine synthetase (ASNS) and sensitivity of cell lines to exogenous L-asparaginase [55].

Microarray gene expression and drug response profiling can be further combined with cheminformatics studies. The molecular structures of low molecular weight compounds are used to generate molecular descriptors that are then used to predict the physicochemical properties and drug activities of compounds. The mix of gene expression studies, drug response studies, and the computation of molecular structures is a very useful approach of great promise. More details of research on this line can be found in Refs. [56–60].

Outlook. Microarray-based gene expression profiling experiments allow the monitoring of global changes of gene transcripts of cells. It already has had a profound impact in diverse fields of biomedical research such as pathology, cancer biology, diagnosis and developmental biology. With the gradual adoption of the practice of repeated experiments using multiplicative samples by practitioners, computational analysis of microarray data can provide more biological insights and generate more interesting hypotheses. To a large extent, the utility of microarray studies will depend on the experimental design: what biological question can be studied with what available biological samples using what biological techniques under what perturbation. As an example, a challenge in microarray studies is the heterogeneity of biological samples from different tissues. Although laser microdissection can provide a more homogeneous source of cells, the number of cells that can be harvested remains small, and experimental linear amplification techniques so far have not been rigorously validated. In this context, clever design strategy such as the application of various tissue specific inhibitors may prove to be useful.

Another major bioinformatics challenge of microarray analysis is the global integration of microarray studies of different tissues and cell lines under various different conditions from different investigators. Yet another challenge is to integrate microarry expression profiles with other bioinformatics analyses, for examples, the detection of membrane proteins as potential markers, the discovery of previously unknown biological roles by combining expression studies and the detection of sequence/structure function motifs, as well as integration with pharmacological and cheminformatics studies. Ultimately, the integration of gene expression under various conditions with the analysis of multiple bioinformatics tools will help to tease out various components of regulatory and metabolic genetic networks of cells.

Acknowledgements

We thank Whitaker Foundation, National Science Foundation for supporting this research. We thank Drs. Robert Danziger, Carol Westbrook, and Nienqing Xiao for helpful discussions, and Prof. W. Dum for editorial suggestions.

References

[1] 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. Calgiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[2] C.M. Perou, S.S. Jeffrey, M. van de Rijn, C.A. Rees, M.B. Eisen, D.T. Ross, A. Pergamenschikov, C.F. Williams, S.X. Zhu, J.C.F. Lee, D. Lashkari, D. Shalon, P.O. Brown, D. Botstein, Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 9212–9217.

[3] 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 and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 6745–6750.

[4] R. Anbazhagan, T. Tiban, D.M. Bornman, J.C. Johnston, J.H. Saltz, A. Weigering, S. Piantadosi, E. Gabrielson, Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles, Cancer Res. 59 (1999) 5119–5122.

[5] A.A Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.

[6] D. Lockart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobahashi, H. Horton, E. Brown, Expression monitoring by hybridization to high-density oligonucleotide arrays, Nat. Biotechnol. 14 (1996) 1675–1680.

[7] J. DeRisi, L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, J.M. Trent, Use of a cDNA microarray to analyse gene expression patterns in human cancer, Nat. Genet. 14 (1996) 457–460.

[8] G. Pietu, O. Alibert, V. Guichard, B. Lamy, F. Bois, E. Leroy, R. Mariage-Sampson, R. Houlgatte, P. Soularue, C. Auffray, Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array, Genome Res. 6 (1996) 492–503.

[9] L. Wodicka, H. Dong, M. Mittmann, M.H. Ho, D.J. Lockhart, Genome-wide expression monitoring in *Saccharomyces cerevisiae*, Nat. Biotechnol. (1997) 1359–1367.

[10] J.L. DeRisi, V.R. Iyer, P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scal, Science 278 (1997) 680–686.

[11] M.-L.T. Lee, F.C. Kuo, G.A. Whitmore, J. Sklar, Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 9834–9839.

[12] R.V. Hogg, A.T. Craig, Introduction to Mathematical Statistics, Prentice-Hall, Upper Saddle River, New Jersey, 1995.

[13] W.J. Conover, Practical Nonparametric Statistics, Wiley, 1999.

[14] G.S. Michaels, D.T. Carr, M. Askenazi, S. Fuhrman, X. Wen, R. Somogyi, Cluster analysis and data visualization of large-scale gene expression data, Pac. Symp. Biocomput. 3 (1998) 42–53.

[15] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, Mol. Biol. Cell 9 (1998) 3273–3297.

[16] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 14863–14868.

[17] X. Wen, S. Fuhrman, G.S. Michaels, D.B. Carr, S. Smith, J.L. Barker, R. Somogyi, Large-scale temporal gene expression mapping of central nervous system development, Proce. Natl. Acad. Sci. U. S. A. 95 (1998) 334–339.

[18] 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 and application to hematopoietic differentiation, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 2907–2912.

[19] J.A. Hartigan, Clustering Algorithms, Wiley, 1975.

[20] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Academic Press, 1999.

[21] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data, Wiley, 1990.

[22] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, 1973.

[23] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[24] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S-Plus, Springer, 1999.

[25] K. Zhang, H. Zhao, Assessing reliability of gene clusters from gene expression data, Funct. Integr. Genomics 1 (2000) 156–173.

[26] M.K. Kerr, G.A. Churchill, Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 8961–8965.

[27] A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Applications, Cambridge Univ. Press, Cambridge, UK, 1997.

[28] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, 1996.

[29] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge Univ. Press, New York, 1996.

[30] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Technical Report 576, Department of Statistics, UC Berkeley, June 2000.

[31] R. Penrose, A generalized inverse for matrices, Proc. Cambridge Philos. Soc. 51 (1955) 406–413.

[32] V.N. Vapnik, The Nature of Statistical Learning Theory, 2nd edn., Springer, New York, 2000.

[33] V. Vapnik, A. Chervonenkis, Theory of Pattern Recognition [in Russian], Nauka, Moscow, 1974. (German Translation: W. Wapnik and A. Tscherwonenkis, Theorie der Zeichenerkennung, Akademie-Verlag, Berlin, 1979).

[34] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[35] I. Steinwart. On the influence of the kernel on the generalization ability of support vector machines. Technical Report 01–01, 2001.

[36] E. Osuna, R. Freund, F. Girosi, An improved training algorithm for support vector machines, in: J. Principe, L. Gile, N. Morgan, E. Wilson (Eds.), Neural Networks for Signal Processing VII—Proceedings of the 1997 IEEE Workshop, IEEE, New York, 1997, pp. 276–285.

[37] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Schölkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp. 185–208.

[38] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, Z. Yakhini, Tissue classification with gene expression profiles, J. Comput. Biol. 7 (2000) 559–584.

[39] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T. Fury, M. Ares, D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 262–267.

[40] C. Hu, J. Liang, Classification of biologically active compounds by support vector machines, J. Chem. Inf. Comput. Sci., submitted for publication.

[41] European molecular biology network, TMpred. http://www.ch.embnet.org/software/ TMPRED_form.html, 2001.

[42] Center for Biological Sequence Analysis, Department of Biotechnology, The Technical University of Denmark, SignalP. http://www.cbs.dtu.dk/services/SignalP, 2001.

[43] K. Nakai, Protein sorting signals and prediction of subcellular localization, Adv. Protein Chem. 54 (2000) 277–344.

[44] J. van Helden, B. Andre, J. Collado-Vides, Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies, J. Mol. Biol. 281 (5) (1998) 827–842.

[45] J. van Helden, A.F. Rios, J. Collado-Vides, Discovering regulatory elements in non-coding sequences by analysis of spaced dyads, Nucleic Acids Res. 28 (2000) 1808–1818.

[46] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, Science 262 (1993) 208–214.

[47] C.E. Lawrence, A.F Neuwald, J.S. Liu, Gibbs motif sampling: detection of bacterial outer membrane protein repeats, Protein Sci. 4 (1995) 1618–1632.

[48] A.F. Neuwald, J.S. Liu, D.J. Lipman, C.E. Lawrence, Extracting protein alignment models from the sequence database, Nucleic Acids Res. 25 (1997) 1665–1677.

[49] J. Zhu, J.S. Liu, C.E. Lawrence, Bayesian adaptive alignment and inference, Proc. Int. Conf. Intell. Syst. Mol. Biol. 5 (1997) 358–368.

[50] F. Roth, J. Hughes, P. Estep, G. Church, Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation, Nat. Biotechnol. 16 (1998) 939–945.

[51] I. Holmes, W.G. Bryno, Finding regulatory elements using joint likelihoods for sequence and expression profile data, Proc. 8th Int. Conf. Intell. Syst. Mol. Biol. (2000) 202–208.

[52] X. Liu, D.L. Brutlag, J.S. Liu, BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes, Pac. Symp. Biocomput. (2001) 127–138.

[53] J.S. Liu, C.E. Lawrence, Bayesian inference on biopolymer models, Bioinformatics 15 (1999) 38–52.

[54] J.S. Liu, Monte Carlo Strategies in Scientific Computing, Springer-Verlag, New York, 2001.

[55] U. Scherf, D.T. Ross, M. Waltham, L.H. Smith, J.K. Lee, L. Tanabe, K.W. Kohn, W.C. Reinhold, T.G. Myers, D.T. Scudiero, D.A. Scudiero, M.B. Eisen, E.A. Sausville, Y. Pommier, D. Botstein, P.O. Brown, J.N. Weinstein, A gene expression database for the molecular pharmacology of cancer, Nat. Genet. 24 (3) (2000) 236–244.

[56] J.N. Weinstein, T.G. Myers, P.M. O'Connor, S.H. Friend, A.J. Fornace Jr., K.W. Kohn, T. Fojo, S.E. Bates, L.V. Rubinstein, N.L. Anderson, J.K. Buolamwini, W.W. van Osdol, A.P. Monks, D.A. Scudiero, E.A. Sausville, D.W. Zaharevitz, B. Bunow, V.N. Viswanadhan, G.S. Johnson, R.E. Wittes, K.D. Paull, An information-intensive approach to the molecular pharmacology of cancer, Science 275 (5298) (1997) 343–349.

[57] T.G. Myers, N.L. Anderson, M. Waltham, G. Li, J.K. Buolamwini, D.A. Scudiero, K.D. Paull, E.A. Sausville, J.N. Weinstein, A protein expression database for the molecular pharmacology of cancer, Electrophoresis 18 (3–4) (1997) 647–653.

[58] L.M. Shi, Y. Fan, T.G. Myers, P.M. O'Connor, K.D. Paull, S.H. Friend, J.N. Weinstein, Mining the NCI anticancer drug discovery databases: genetic function approximation for the qsar study of anticancer ellipticine analogues, J. Chem. Inf. Comput. Sci. 38 (2) (1998) 189–199.

[59] L.M. Shi, Y. Fan, J.K. Lee, M. Waltham, D.T. Andrews, U. Scherf, K.D. Paull, J.N. Weinstein, Mining and visualizing large anticancer drug discovery databases, J. Chem. Inf. Comput. Sci. 40 (2) (2000) 367–379.

[60] O. Keskin, I. Bahar, R.L. Jernigan, J.A. Beutler, R.H. Shoemaker, E.A. Sausville, D.G. Covell, Characterization of anticancer agents by their growth inhibitory activity and relationships to mechanism of action and structure, Anticancer Drug Dev. 5 (2000) 79–98.