BIOINFORMATICS

Vol. 18 no. 9 2002
Pages 1184–1193

Determination of minimum sample size and discriminatory expression patterns in microarray data

Daehee Hwang, William A. Schmitt, George Stephanopoulos and Gregory Stephanopoulos*

Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received on August 3, 2001; revised on December 20, 2001; accepted on March 26, 2002

ABSTRACT

Motivation: Transcriptional profiling using microarrays can reveal important information about cellular and tissue expression phenotypes, but these measurements are costly and time consuming. Additionally, tissue sample availability poses further constraints on the number of arrays that can be analyzed in connection with a particular disease or state of interest. It is therefore important to provide a method for the determination of the minimum number of microarrays required to separate, with statistical reliability, distinct disease states or other physiological differences.

Results: Power analysis was applied to estimate the minimum sample size required for two-class and multi-class discrimination. The power analysis algorithm calculates the appropriate sample size for discrimination of phenotypic subtypes in a reduced dimensional space obtained by Fisher discriminant analysis (FDA). This approach was tested by applying the algorithm to existing data sets for estimation of the minimum sample size required for drawing certain conclusions on multi-class distinction with statistical reliability. It was confirmed that when the minimum number of samples estimated from power analysis is used, group means in the FDA discrimination space are statistically different.

Contact: gregstep@mit.edu

INTRODUCTION

DNA and protein arrays are becoming standard tools for probing the cellular state and determining important cellular behavior at the genomic and protein levels. Oligo-nucleotide and cDNA arrays (Schena *et al.*, 1995; Lockhart *et al.*, 1996) are being employed increasingly for determining discriminatory genes and discovering new classes of disease subtypes that are differentiated at the level of transcription (Golub *et al.*, 1999; Stephanopoulos *et al.*, 2002). Data-driven hypotheses are developed from these types of measurements that suggest, in turn, novel experiments furthering biomedical research.

There are certain issues of statistical reliability that need to be addressed in the implementation of array technologies. Microarray data are typically subjected to analyses such as hypothesis testing, classification, clustering, and network modeling that rely on statistical parameters in order to draw conclusions (Alizadeh *et al.*, 2000; Stephanopoulos *et al.*, 2002; Golub *et al.*, 1999). However, these parameters cannot be reliably estimated with only a small number of array samples and poor sample distributions of gene expression levels. Since the statistical reliability of conclusions largely depends on the accuracy of the parameters used, a certain minimum number of arrays is required to ensure confidence in the sample distribution and accurate parameter values.

This study is concerned with the determination of the minimum number of gene expression arrays required to ensure statistical reliability in disease classification and identification of distinguishing expression patterns. This is an important issue considering the scarcity of tissue samples that can be used for transcriptional profiling and the fact that microarray measurements are rather costly in terms of time and reagents required. As a result, there is a tendency to carry out only a small number of microarray measurements that in many cases are inadequate for the intended purpose. Conclusions based on an inadequate number of arrays will not be statistically sound.

The method proposed here first identifies differentially expressed genes across disease subtypes, hereafter called discriminatory genes, using Wilks' lambda score (Johnson and Wichern, 1992; Stephanopoulos *et al.*, 2002; Dillon and Goldstein, 1984) and leave one out cross-validation (LOOCV) (Lachenbruch and Mickey, 1968). Then, Fisher discriminant analysis (FDA) (Stephanopoulos *et al.*, 2002; Dillon and Goldstein, 1984; Zhao and Maclean, 2000) is invoked to define linear combinations of

*To whom correspondence should be addressed.

these discriminatory genes that form a lower dimensional discrimination space where disease subtypes (classes) are maximally separated. Finally, the minimum number of array samples necessary is estimated to ensure satisfactory separation of the linear combinations (i.e. the projections) of the discriminatory genes in the discrimination space. It should be noted that the minimum number of array samples is estimated only in the reduced-dimensional space, and therefore the composite expressions of the genes are well characterized and not necessarily the individual genes themselves.

SYSTEMS AND METHODS

The data-set

The previously published set of leukemia gene expression data (Golub *et al.*, 1999) was used in this work. The data set comprises 72 samples of which 47 were classified as of acute lymphoid leukemia (ALL) and 25 as acute myeloid leukemia (AML). A further division of ALL samples into 38 B-lineage acute lymphoid leukemia (B-ALL) and 9 T-lineage acute lymphoid leukemia (T-ALL) was considered in extending the sample determination approach to the multi-class case of three disease subtypes (B-ALL, T-ALL, and AML). The sample classification among the three subtypes given in Golub *et al.* was also used here, as it was based on both clinical information and validation through their pattern discovery technique.

Selection of discriminatory genes

Several statistical measures have been proposed to identify discriminatory genes for two conditions (e.g. cancerous and normal tissues). Parametric tests such as P-value (Golub *et al.*, 1999) and t-test (Thomas *et al.*, 2001) are based on differences of group means, while non-parametric tests such as Wilcoxon rank sum (Mann–Whitney) test are based on differences of rank sums in groups (Thomas *et al.*, 2001). Parametric tests may perform poorly due to violation of their underlying assumptions, such as normality and equal variance in the various groups. A non-parametric test does not rely on these assumptions and works well with a small sample size, but the results may be more critically sensitive on the nature of the samples used for the training of the classifier than those in parametric tests. No method is unanimously optimal for all kinds of data. Selection of a method for application to a certain data set should depend on the characteristics of the data, the extent of violation of the underlying assumptions, and the sample size. We propose a well-characterized alternative measure, called Wilks' lambda score (Johnson and Wichern, 1992; Dillon and Goldstein, 1984) to assess discriminatory powers of the individual genes. Wilks' lambda, which originated from ANOVA, is not limited only to two-class comparisons but can also be used for multi-class cases. It produces more robust test results than multiple two-class comparisons using t-test because the Wilks' lambda is based on group-variance instead differences between group means and rank sums.

Genes whose expression distribution has high between-group variance (the groups are well separated) and small within-group variance (the samples inside each group are relatively similar) are deemed to be discriminatory for the sample classes (Dillon and Goldstein, 1984; Dudoit *et al.*, 2001). The between-group variance (B_i) of the expression of a certain gene i is proportional to the sum of the differences between group means of expression levels. The within-group variance of the expression of gene i (W_i) is the sum of group variances of the expression levels of the gene in a single class. With the total variance of expression levels of gene i, $T_i = (\mathbf{x}_i - \mathbf{1}\bar{\mathbf{x}}_i)^T(\mathbf{x}_i - \mathbf{1}\bar{\mathbf{x}}_i)$, the within- and the between-group variances are defined respectively as follows:

$$W_i = \sum_{j=1}^{c} W_i^j = \sum_{j=1}^{c} (\mathbf{x}_i^j - \mathbf{1}\bar{x}_i^j)^T(\mathbf{x}_i^j - \mathbf{1}\bar{x}_i^j) \qquad (1)$$

$$B_i = T_i - W_i. \qquad (2)$$

The vector, $\mathbf{x}_i$ ($N \times 1$), contains the expression level of gene i in N samples and $\bar{x}_i$ is the mean expression of gene i in all N samples. The superscript j represents class j among the c classes. For the two genes shown schematically in Figure 1, gene 1 has a large between-group variance and a small within-group variance while gene 2 has a small between-group variance (overlapping distributions across the classes) and a large within-group variance. For gene 1, the large ratio of the between-group variance to within-group variance indicates a gene with a discriminatory expression pattern. Without loss of information, the above procedure is implemented through a statistical test based on Wilks' lambda (Λ_i) that allows one to establish a formal boundary between discriminatory genes and non-discriminatory genes:

$$\Lambda_i = \frac{W_i}{T_i} \qquad (3)$$

In order to compare the Wilks' lambda (Λ_i) score to a distribution with known parameters, it is transformed to the F distribution as follows (Dillon and Goldstein, 1984; SAS, 1989):

$$F_i = \frac{(1 - \Lambda_i)}{\Lambda_i}\frac{(N - c)}{(c - 1)} \sim F_{\alpha(c-1,N-c)} \qquad (4)$$

where N is the total number of samples and c is the number of classes. In this form, discriminatory genes are selected by applying a statistical cutoff determined from the F distribution using some level of significance (in this

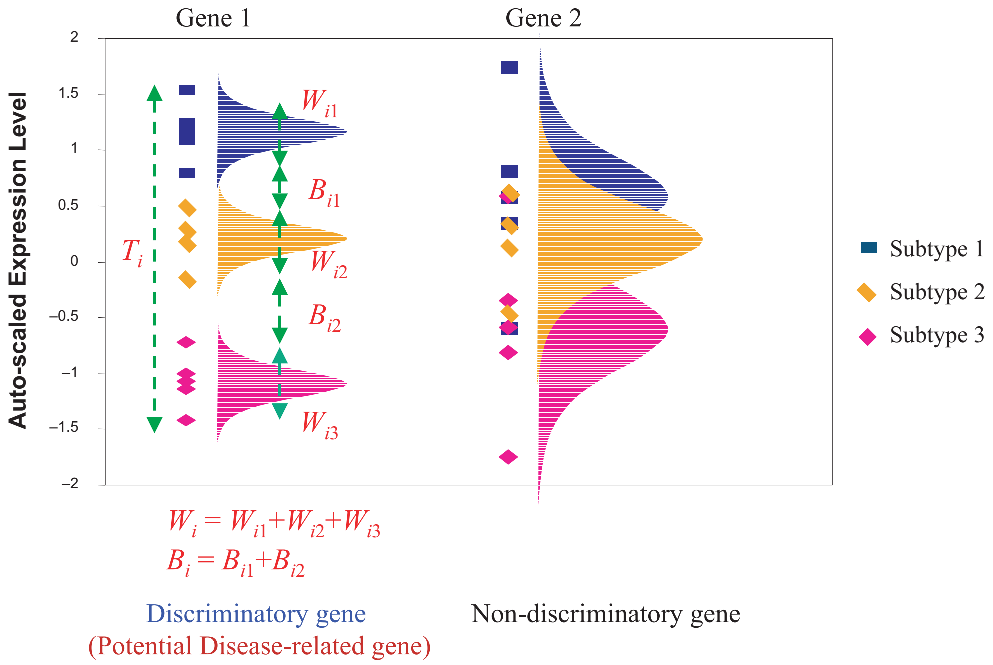

Fig. 1. Discriminatory genes (potential disease related genes) and non-discriminatory genes.

case $\alpha = 0.01$). Note that a high F value signifies a more discriminatory gene relative to one with a low F value.

As a parametric measure, Wilks' lambda score might produce a high false positive error due to violation of the underlying assumptions, especially normality. This is especially true for genes that have small difference in their expressions between groups. In order to improve the false-positive rate, we incorporate an error rate calculation through LOOCV procedure into the discriminatory analysis (Dillon and Goldstein, 1984; Lachenbruch and Mickey, 1968). In this procedure, a series of many LOOCVs are performed to get a good error estimate (see Figure 2a). The first step in this iterative procedure consists of randomly dividing the data set being considered into c test samples (i.e. one test sample for each class) and $N - c$ training samples. The training samples are used to generate an initial set of discriminatory genes using Wilks' lambda score (F statistic values). Using the gene with highest F value, a FDA classifier is constructed and the error rate calculated for the c test samples (see next section). A second classifier is then constructed using the top two discriminating genes, which is again applied to the test samples. The number of genes included in the classifier is thus sequentially increased to form more complex classifiers until all genes selected with the Wilks' lambda score have been included. At each step, the number of misclassified samples is determined for calculation of the misclassification error rate (see the next paragraph). A new division of the samples into training and test sets is then considered, and the procedure is repeated.

For the estimation of error rates, the entire LOOCV procedure is repeated g times using different test and training sets, until all samples have been withheld in the test set at least once. If we denote by m_p the number of misclassified samples in the g cross-validations for a given number of discriminatory genes (p) used in the classifiers, the averaged error rate is given by $e(p) = m_p/(c \times g)$. Then, the error rates from the g cross-validation iterations can be computed as function of the number of discriminatory genes. Using the error rate curve, the number of discriminatory genes can be determined at the point where the averaged error rates show an asymptotic behavior (see Figure 3d and 4d). Then, a final set of 45 discriminatory genes is determined based on the frequencies by which they appeared as discriminatory genes during the g LOOCVs. (see Figure 2a). The final list of genes (Figure 3d) is shorter than the original list of discriminatory genes (Figure 3c), thus enabling us to reduce the false positive error by identifying a small set of genes robust to sample variation. If a gene with small expression difference between the two classes of samples shows up consistently in the LOOCV procedure, it indicates that the observed difference, even though small, is statistically reliable. Other methods have also introduced to reduce false positives in identifying discriminatory genes (Storey and Tibshirani, 2001; Tusher

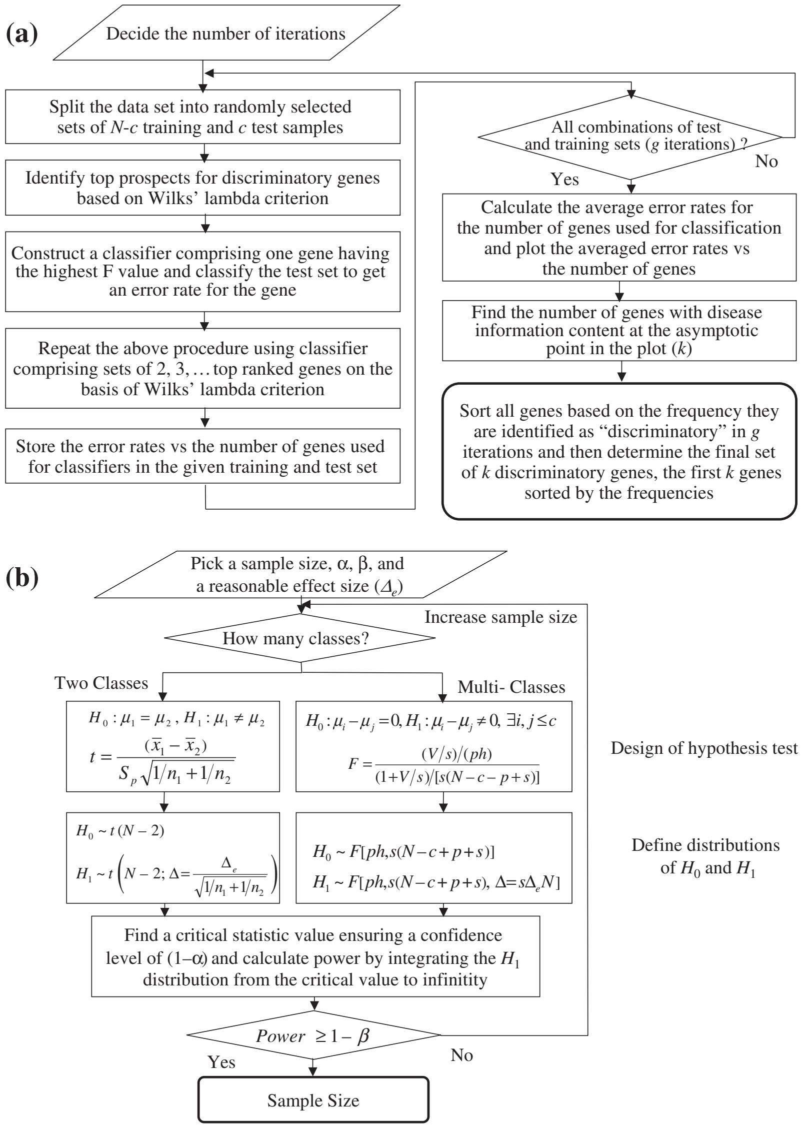
(a)
Decide the number of iterations
Split the data set into randomly selected sets of N-c training and c test samples
Identify top prospects for discriminatory genes based on Wilks' lambda criterion
Construct a classifier comprising one gene having the highest F value and classify the test set to get an error rate for the gene
Repeat the above procedure using classifier comprising sets of 2, 3, … top ranked genes on the basis of Wilks' lambda criterion
Store the error rates vs the number of genes used for classifiers in the given training and test set
All combinations of test and training sets (g iterations) ?
No
Yes
Calculate the average error rates for the number of genes used for classification and plot the averaged error rates vs the number of genes
Find the number of genes with disease information content at the asymptotic point in the plot (k)
Sort all genes based on the frequency they are identified as "discriminatory" in g iterations and then determine the final set of k discriminatory genes, the first k genes sorted by the frequencies
(b)
Pick a sample size, α, β, and a reasonable effect size (Δ_e)
Increase sample size
How many classes?
Two Classes
Multi- Classes
$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$
$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S_p\sqrt{1/n_1 + 1/n_2}}$
$H_0: \mu_i - \mu_j = 0, H_1: \mu_i - \mu_j \neq 0, \exists i, j \leq c$
$F = \frac{(V/s)/(ph)}{(1+V/s)/[s(N-c-p+s)]}$
Design of hypothesis test
$H_0 \sim t(N-2)$
$H_1 \sim t\left(N-2; \Delta = \frac{\Delta_e}{\sqrt{1/n_1 + 1/n_2}}\right)$
$H_0 \sim F[ph, s(N-c+p+s)]$
$H_1 \sim F[ph, s(N-c+p+s), \Delta = s\Delta_e N]$
Define distributions of H_0 and H_1
Find a critical statistic value ensuring a confidence level of (1–α) and calculate power by integrating the H_1 distribution from the critical value to infinitity
$Power \geq 1 - \beta$
Yes
No
Sample Size

Fig. 2. (a) Leave one out cross-validation (LOOCV) algorithm, where N is the total number of samples and c is the number of classes, so that one sample from each class is included in the test. (b) Power analysis algorithm for determination of the minimum sample size.

et al., 2001), and may be used instead of LOOCV at the user's discretion.

Dimensional reduction by discriminant analysis

FDA is a linear method of dimensionality reduction from the expression space comprising all selected discriminatory genes to just a few dimensions where the separation of sample classes is maximized. FDA is similar to principal component analysis (PCA) (Alter *et al.*, 2000; Holter *et al.*, 2000) in the linear reduction of data (Johnson and Wichern, 1992; Dillon and Goldstein, 1984). The major difference is that the discriminant axes of the FDA space are selected such as to maximize class separation in the reduced FDA space, instead of variability as in the case of PCA. The discriminant axes of FDA, termed as discriminant weights (**V**), maximizing the separation of sample classes in their projection space can be shown to be equivalent to the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$, the ratio of between-group variance (**B**) to within-group variance (**W**), as shown in Equation (5).

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{V} = \mathbf{V}\Lambda \tag{5}$$

where

$$\mathbf{B} = \mathbf{T} - \mathbf{W}, \mathbf{W} = \sum_{j=1}^{c} (\mathbf{X}_j - \mathbf{1}\bar{\mathbf{x}}_j^T)^T (\mathbf{X}_j - \mathbf{1}\bar{\mathbf{x}}_j^T),$$

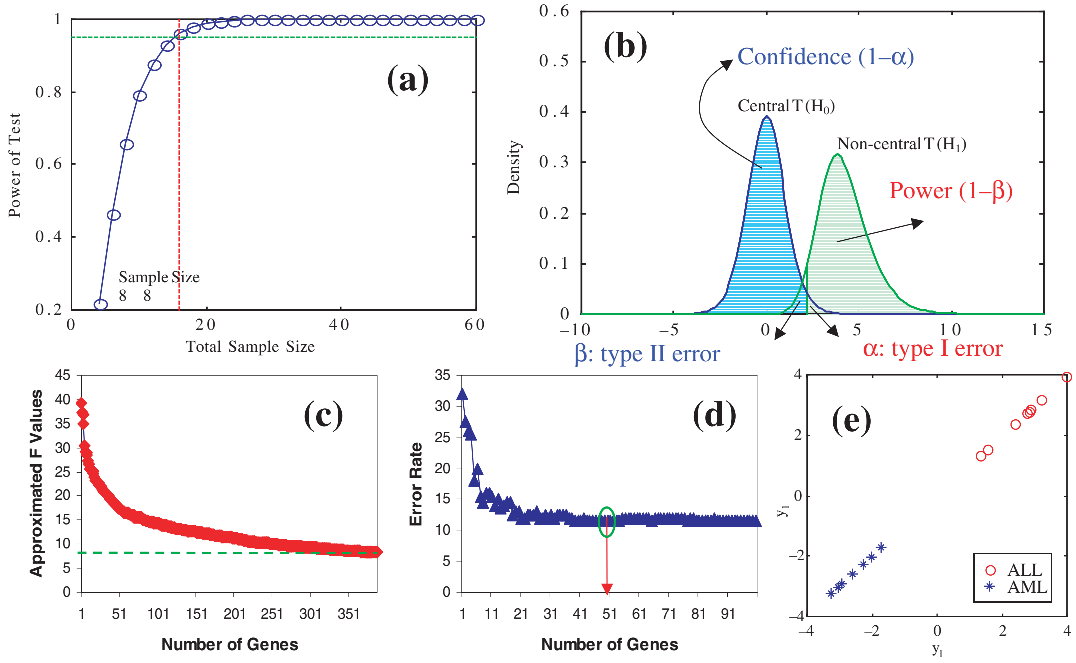

Fig. 3. Determination of minimum sample size for two-class (ALL, AML) distinction, selection of discriminatory genes with the estimated sample sizes of two classes, and FDA projection. (a) Power plot versus sample size showing how to determine the sample size required for two-class distinction (eight from each class). (b) The distributions of H_0 and H_1 for the determined sample size. (c) Univariate F statistic values of the initial 388 discriminatory genes with a threshold ($F_{0.01(1,\ 18)} = 8.2854$) in randomly selected eight ALL and eight AML samples out of the entire data set. (d) Leave-one-out cross-validation applied to estimate the classification error rates and then to select the 50 most discriminatory genes with the same samples. (e) Separation of eight ALL and eight AML samples in the two-dimensional FDA projection space defined discriminant axes of 50 discriminatory genes.

and

$$\mathbf{T} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T)^T(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T).$$

The eigenvalues (Λ) indicate the discrimination power for the corresponding discriminant axes. Further details of FDA, and its application in classification of microarray data are described in Stephanopoulos *et al.* (2002). Figures 3e and 4e show the projection of the expression data in the 2-class (AML and AML) or the 3-class (B-ALL, T-ALL and AML, respectively).

A classification rule can be built in the FDA space. A new sample is projected into the FDA space using the discriminant weights (**V**). Then, the new sample will be assigned to the predefined class whose mean is closest to the projection of the new sample (Johnson and Wichern, 1992): a new sample (**x**) will be allocated to class j if

$$\begin{aligned}\|\hat{\mathbf{y}} - \bar{\mathbf{y}}_j\|^2 &= \|(\hat{\mathbf{x}} - \bar{\mathbf{x}}_j)\mathbf{V}\|^2 \\ &\leqslant \|(\hat{\mathbf{x}} - \bar{\mathbf{x}}_k)\mathbf{V}\|^2 \quad \text{for all } k \neq j \qquad (6)\end{aligned}$$

where $\hat{\mathbf{y}}$ is a projection of the new sample into the discriminant axes (**V**). It has been shown (Johnson and Wichern, 1992) that FDA is an optimal classification procedure in the sense of the error rates under two assumptions: (1) multivariate normality of the p discriminatory genes, and (2) equal $p \times p$ covariance matrices for each of the c classes. Violation of the assumptions affects several aspects of FDA. For instance, with unequal covariance matrices, a quadratic classification rule in the FDA projection space performs better than the linear classification rule in Equation (6). Agreement between the quadratic rule and the linear one will decline as the sample sizes decreases, the differences in class covariance matrices increase, the class means become closer, or the number of discriminatory genes increases. In this case study, we employed the linear FDA classifier because of simplicity and appropriateness in our gene selection procedure. However, we tried to minimize the effect of violations of the assumptions: false positives have been minimized by two-step selection of a small sub-set of genes, and we ensured sufficient mean difference among classes using power analysis (see next section). Other classifiers have also been introduced which can be applied to various microarray data (Dudoit *et al.*, 2000; West *et al.*, 2001), and in general these may be substituted for the FDA classifier.

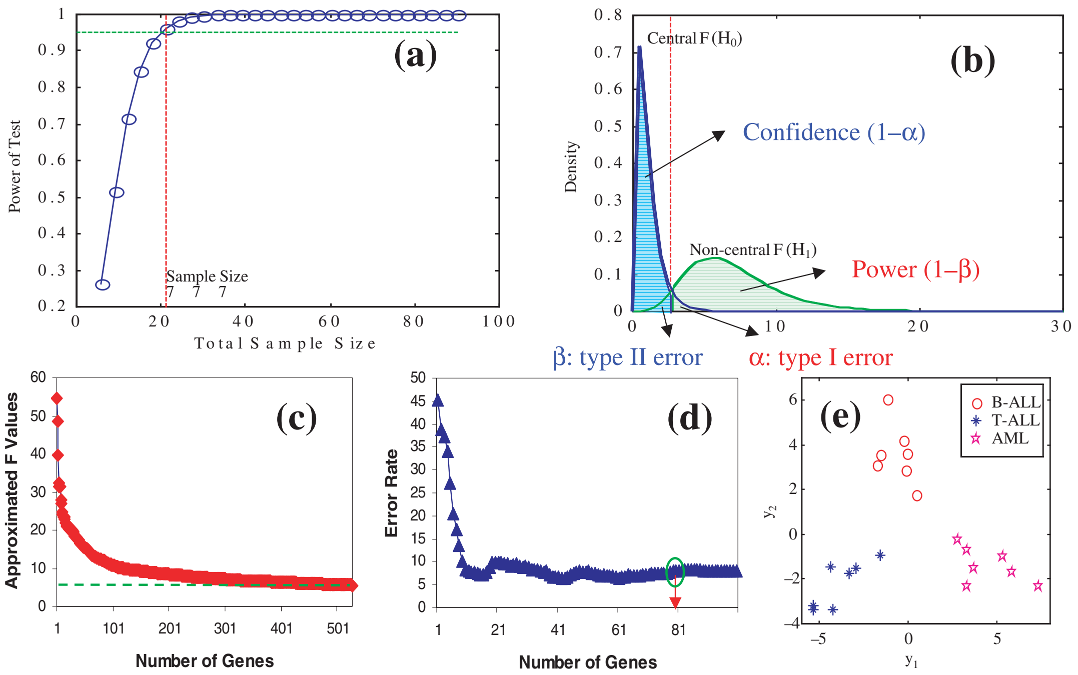

Fig. 4. Determination of minimum sample size for the three-class (B-ALL, T-ALL, AML) distinction, selection of discriminatory genes with the estimated sample sizes of three classes, and FDA projection. (a) Power plot versus sample size showing how to determine the sample size (seven from each class). (b) The distributions of H_0 and H_1 for the determined sample size. (c) Univariate F statistic values of the initial 527 discriminatory genes with a threshold ($F_{0.01(2,\ 26)} = 5.5263$) in randomly selected seven B-ALL, seven T-ALL and seven AML samples out of the entire data set. (d) Leave-one-out cross-validation applied to estimate the classification error rates and then to select the 80 most discriminatory genes with the same samples. (e) Separation of seven B-ALL, seven T-ALL and seven AML samples in the two-dimensional FDA projection space defined discriminant axes of the discriminatory 80 genes.

Determination of the minimum sample size using power analysis

Determining the number of microarray samples has been presented as an important issue previously (Pan *et al.*, 2001; Zien *et al.*, 2001) and is one of the first things to be considered when attempting classification of samples through microarrays. We present power analysis for determination of the minimum sample size required for accurate classification. Instead of using individual genes, we used the c-1 dimensional FDA projections (**y** in Equation (6)) in our analysis, because the FDA classification is based on those projection variables. Then, we validated the estimated minimum sample size by testing the entire methodology presented in this paper: selecting discriminatory genes, building a FDA classifier, and finally calculating the actual power (see Results).

Power analysis (Cohen, 1988; Kraemer and Thiemann, 1987; Mace, 1974) has been used in many applications and is based on two measures of statistical reliability in the hypothesis test, the confidence level (1-α) and power (1-β). The test compares the null hypothesis (H_0) that the means of classes are the same against the alternative hypothesis (H_1) that the means of classes are not same. While the confidence level of a test is the probability of accepting the null hypothesis, when the means of classes are in fact same, the power of a test is the probability of accepting the alternative hypothesis, when the means of classes are in fact different (see 'G*Power reference material'). Alternatively, the type I error (false positives, α) is the probability of accepting the alternative hypothesis, when the means of the classes are in fact the same, while the type II error (false negatives, β) is the probability of accepting the null hypothesis, when the means of the classes are in fact different (see 'G*Power reference material'). The estimation of the sample size in power analysis is done in such a way that the two statistical reliability measures, the confidence and the power, in the hypothesis test can reach predefined values. Typical analyses may require as 95% confidence and 95% power, for example.

The confidence level and the power are calculated from the distributions of the null hypothesis (H_0) and the alternative hypothesis (H_1). Defining these distributions depends on the statistical measure being used in the hypothesis test. In the case of two-class distinction having a one-dimensional FDA projection, the normalized mean difference follows the t distribution in the FDA space. This t statistical measure for the hypothesis test is defined below (Kraemer and Thiemann, 1987; Mace, 1974):

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 \neq \mu_2 \tag{7}$$

$$t = \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \tag{8}$$

where μ_i and $\bar{\mathbf{y}}_i$ are the actual mean and the sample mean of the one-dimensional projection variable ($\mathbf{y}_i$) in class i. S_p is the pooled standard deviation of the projection variable of the two classes, n_i is the number of samples in class i, N is the total number of samples and N-2 is the degrees of freedom in the t distribution. While the distribution of H_0 with all classes having the same mean is defined as a central distribution, the distribution of H_1 with all classes having different means is a non-central. The *effect size* (Δ_e) should be set in advance, before power analysis is conducted. The effect size is a critical mean difference that can be considered important enough to warrant attention. Power analysis estimates the minimum sample size to ensure the power in the test for the effect size. The non-central distribution H_1 is defined by the non-centrality parameter (Δ), which is defined by the effect size (see below). For the case of two-class distinction, the effect size (Δ_e) is the critical mean difference normalized by the pooled standard deviation (S_p). Thus, the distributions of H_0 and H_1 are defined as follows:

$$H_0 : t = \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(N-2) \tag{9}$$

$$H_1 : t = \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t\left(N-2; \Delta = \frac{\Delta_e}{\sqrt{1/n_1 + 1/n_2}}\right) \tag{10}$$

$$\text{with } \Delta_e = \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)_{crit}}{S_p}.$$

The confidence level and the power are calculated using the defined distributions of H_0 and H_1 for a given sample size and an initial guess for the effect size determined on the basis of engineering judgement or prior knowledge of the system. The critical value of the inverse t distribution at the probability of 1-$\alpha/2$, shown by the dotted line in Figure 3b, is first identified for the distribution of H_0 (here, $\alpha = 0.05$ to give a 95% confidence level). For this confidence level, the power is determined next using the distribution of H_1 in the region from this critical t value to positive infinity (indicated in Figure 3b by the area under the H_1 distribution after the critical value). If the power calculated is below the predefined value 1-β (here, 95%), the sample size is increased until the power reaches this threshold, as shown in Figure 3a. Figure 3b shows the confidence level, power, type I error and type II error in the distributions of H_0 and H_1 defined by the determined sample size. The sample size estimated from this power analysis is the total number of samples, so that the number of samples required in each class is obtained by dividing the total sample size by the number of classes (c). This assumes that the standard deviation matrix is approximately similar for each class, implying that equal numbers of samples are needed for each class.

In the case of distinguishing $c > 2$ classes, instead of the t statistic (see Equation (8)), the F statistic measure derived from Pillai's V is used for the estimation of the sample size (Olson, 1974). Pillai's V is the trace of the matrix defined by the ratio of between-group variance (**B**) to total variance (**T**), and is a statistical measure often used in multivariate analysis of variance (MANOVA) (SAS, 1989; Olson, 1974):

$$V = trace\left(\mathbf{B}\mathbf{T}^{-1}\right) = \sum_{i=1}^{h} \frac{\lambda_i}{1+\lambda_i} \tag{11}$$

where λ_i is the ith eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ and h is the number of factors being considered in MANOVA, defined by $h = c$-1. When **W** and **B** are computed in Equation (5), the c-1 dimensional FDA projections are used, because they are the test variables for this analysis. A high Pillai's V means a high amount of separation between the samples of classes, with the between-group variance being relatively large compared to the total variance. The hypothesis test can be designed as shown below using the F statistic transformed from Philai's V (see 'Other F-tests').

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_c \text{ and } H_1 : \mu_i - \mu_j \neq 0 \; \exists \; i, j \tag{12}$$

$$H_0 : F = \frac{(V/s)/(ph)}{(1 - V/s)/[s(N-c-p+s)]} \sim F[ph, s(N-c-p+s)] \tag{13}$$

$$H_1 : F = \frac{(V/s)/(ph)}{(1 - V/s)/[s(N-c-p+s)]} \sim F[ph, s(N-c-p+s), \Delta = s\Delta_e N]$$

$$\text{with } \Delta_e = \frac{V_{crit}}{(s - V_{crit})} \tag{14}$$

where p and c are the number of variables and the number of classes, respectively. s is defined by $\min(p, h)$. The confidence level and the power can be calculated using these defined distributions of H_0 and H_1 for a given sample size and an effect size. The same procedure used in the case of two-class distinction is used here to estimate the minimum sample size for statistical reliability whereby the sample size is increased until the calculated power reaches the predefined threshold value of 1-β (95% for the cases shown here). Figure 4a and 4b show the calculated sample size and the distributions of H_0 and H_1 for the case of leukemia samples from three classes.

The above approach is applicable only to FDA projection variables and not to the expression data from a large number of individual genes, because the denominator in the F statistic (Equation (13) and (14)), which generally has a positive value, becomes negative due to a large number of genes (p). PCA can be used to reduce the number of variables (p) to resolve such a problem. There is, however, a limitation that the number of PCs (p) cannot be larger than $N - c + s = N - 1$ and in most array cases the maximum number of PCs ($p = N - c + s$) does not capture enough discriminating characteristics among the classes. Thus, we use only the projections through FDA in our analysis. This analysis may produce a misleading sample size estimate when the real gene expression data are not consistent with the assumptions (normality and equal variance) underlying the statistics used in power analysis. To check the effect of possible violations of the assumptions on the estimated sample size, the actual power and mean differences between classes are compared to the predefined values (see Results). The actual values in both cases studied were sufficiently large that we need not be worried about the impact of data which does not perfectly match the normality or equal variance assumptions.

Algorithm

Figure 2 provides a schematic of the leave one out cross-validation (Figure 2a) for error rate estimation and power analysis algorithms (Figure 2b). Each LOOCV first splits the data set into randomly selected sets of $N - c$ training and c test sets. Then, discriminatory genes are selected on the basis of their Wilks' lambda. The number of genes included in the FDA classifier is increased by one in order of decreasing magnitude of their F value. The number of misclassified c test samples is counted as a function of the number of genes. The above procedure is repeated g times for different randomly selected training and test sets. The average error rates calculated by $e(p) = m_p/(c \times g)$ are then plotted versus the number of genes included in the classifier and the discriminatory genes are selected based on the number of times they are identified as "discriminatory" in all the iterations.

For power analysis, it is first necessary that the type I and type II errors, an initial sample size, and a reasonable effect size are selected for the initiation of the algorithm. Then, after the test is designed in terms of the null hypothesis, the alternative hypothesis, and an appropriate statistic measure (t-test or F test), the distributions of H_0 and H_1 are determined using the degrees of freedom and the non-centrality parameter. Next, the inverse of the F distribution at the value of 1-α in probability is identified and the power is calculated using the distribution of H_1. If the calculated power is less than the predefined power, 1-β, then the sample size is increased and the power is recalculated using the same α, β, and effect size but a new sample size until it reaches the preset power value. Following determination of the number of samples from power analysis, the actual effect size and power are computed and their values compared to the initial guesses. The actual effect size and power should be larger than those used/calculated in the original analysis so as to not underestimate the sample size.

Implementations and Results

Power analysis was applied to two-class distinction between ALL and AML subtypes of leukemia. The null hypothesis (H_0) was that the two group means (i.e. the group averages in the FDA space) were the same, with the alternative hypothesis (H_1) that the two group means were not the same. The mean difference normalized by the pooled standard deviation was used as the t statistic measure. The effect size was preset to 2, which corresponds to a mean difference two times larger than the pooled standard deviation and the predefined confidence and power were set to 95% (equivalent to $\alpha = 0.05$ and $\beta = 0.05$). Figure 3a shows the dependence of the power calculated from the H_1 distribution on the sample size. Eight samples from each class (16 total) are required for the FDA projection to establish a sufficient base for the H_1 to be accepted. This indicates that with these eight samples from each class, there is mean difference between ALL and AML so that an accurate classifier can be constructed in the FDA projection space with statistical reliability.

In order to validate this minimum sample size, the proposed procedures for discriminatory gene selection and FDA classification were applied to eight randomly chosen samples from each class and then the actual effect size and the actual power were calculated. The procedure of discriminatory gene selection identified 50 discriminatory genes (Figure 3d). This final list of 50 genes is shorter than the 388 discriminatory genes obtained by using a simple Wilks' lambda score metric without the error rate calculation (Figure 3c), thus enabling us to reduce the errors due to false positives. Then, using the 50 discriminatory genes, FDA classification was performed (Figure 3e). In the FDA projection space, the actual normalized sampled mean difference was computed to be equal to 7.2453. This is more than three times larger than the effect size used for power analysis, confirming that the effect size chosen was reasonable enough not to underestimate the sample size. There are two potential explanations for the difference between the sampled mean difference and the effect size: (1) only the most discriminatory genes were selected with a stringent level of significance in Wilks' lambda and by the LOOCV, and (2) the FDA further screens out the maximal discriminating information from the most discriminatory genes. The actual confidence and power were also close to 100%.

As a multi-class case study, a distinction of three subtypes, B-ALL, T-ALL, and AML was considered. The H_0 is that the three group means are same, while H_1 states that at least one of the group means is different from the rest. The F statistic measure was used for power analysis and the effect size was chosen to be 0.538. This effect size is equivalent to 0.7 critical Pillai's V for three classes, meaning that the between-group variance is 0.7 of the total variance. The predefined confidence and power were set to be 95%, equivalent to $\alpha = 0.05$ and $\beta = 0.05$. The minimum sample size was computed to be seven samples from each class from the power curve shown in Figure 4a. The distributions of H_0 and H_1 are shown in Figure 4b. After the gene selection procedure was applied to seven randomly selected samples from each class, the final set of 80 discriminatory genes was identified. With those genes, the FDA was done as shown in Figure 4e. In the FDA space, the actual measure of effect size defined by $V/(s - V)$ was computed to be 1.7552, which is about three times larger than the one used for power analysis for the same reasons given in the previous case. The actual confidence and power were also close to 100%.

DISCUSSION AND CONCLUSIONS

This study has addressed the issue of statistical reliability for the classification of disease subtypes on the basis of the sample size. The appropriate statistical measures have been defined for two-class and multi-class problems, and these statistics have been applied in a power-analysis framework to determine the minimal sample size based on the distributions of the statistic measures. This framework has been applied in earlier studies (Thall, 1995) for determining the minimum number of subjects required in clinical trial studies, when a new drug is discovered and its efficacy is being evaluated. In this case, the minimal sample size determined from power analysis is used to ensure statistical reliability of an efficacy measure.

This reliability issue can also be central in other applications involving any statistical analysis, with this study giving only one example. For instance, correlations between genes are often considered in microarray studies in the search for co-regulated genes. A small number of samples will result in unreliable correlation coefficients, so when additional samples are included, the estimated correlation coefficients will show a high degree of variability. Thus, the appropriate sample size should be determined by power analysis to ensure that the distribution of correlation coefficients is reliable. Another application is the construction of a regression model using gene expression data to estimate the level of an important cellular variable. For instance, gene expression regression models of urea level in liver tissues should also be supplemented by power analysis to determine the sample size for the model to have statistically reliable regression parameters. Although the range of uses is broad, the appropriate statistical measures and their distributions should be carefully chosen in these sorts of applications.

Power analysis determines the sample size based on the assumption of homogeneous sampling from the entire population of each class (i.e. a disease subtype as in this study). Therefore, during sample collection, if the number of samples suggested by power analysis does not cover the broad population of each subtype to capture the inherent variance of the population, the distributions of parameters will be biased toward the type of samples collected. As a result, a poor sampling can make power analysis appear to underestimate the necessary sample size. Furthermore, statistical inference based on the calculated parameters can be misleading. The FDA has recently noticed the importance of broad sampling and requested pharmaceutical industries to include clinical trial studies on pediatric patients in order that the efficacy measure should not be biased to adults. As a result, a well-designed sampling strategy is required together with a reasonable estimate of sample size calculated from power analysis to ensure statistical reliability.

This study uses linear combinations of individual genes as variables in the classifier instead of the individual genes themselves. Although the discriminatory genes used for the classifier are chosen based on Wilks' lambda score and the error rate calculated through LOOCV, the number of selected genes is usually still large (50 or more depending on the situation). If all individual genes are considered independently in constructing a classifier, and new samples are classified using the sum of all gene contributions to the classifier, the classifier will not capture the interaction of the genes and may be biased to redundant characteristics. In addition, the parameters in the classifier will be subject to statistical variations of the individual genes. If all the genes are considered together as seen in multiple discriminant analysis (MDS), it may be difficult to estimate the model parameters due to the large number of discriminatory genes and singularity in the data. On the other hand, the linear combinations of individual genes obtained from FDA capture the important discriminating characteristics at the outset because the algorithm seeks the most relevant directions (weights) for separation of classes. Thus, the number of variables used for the classifier is significantly reduced to several FDA projection variables (the number of classes – 1), while capturing in a large degree the discriminating characteristics in data. This reduction in variables is achieved without significant accuracy cost in discrimination.

The use of FDA also reduces the amount of noise obscuring the information content of the data. Signals that nearly appear to be random noise will be filtered out during the process of obtaining the weights for the linear combinations. Just as the first few PCs in PCA usually capture the important patterns and the last few PCs only random noise, the first few discriminant functions in FDA captures the important discriminating characteristics in the data. Only systemic noise that happens to have similar patterns to the real signals may be retained in the data projected through the linear combination.

Finally, the interactions and relative contributions of the individual genes to the classification can be interpreted from the discriminant weights in the linear combinations, improving the understanding the discriminant features in the data. As a result, the FDA classifier using linear combinations as variables can provide the preferable aspects in classification, including robustness in performance, noncomplexity in modeling and improvement in interpretation.

ACKNOWLEDGEMENTS

This work was supported by the Engineering Research Program of the Office of Basic Energy Science at the Dept. of Energy, Grant No. DE-FG02-94ER-14487 and DE-FG02-99ER-15015 and NIH Grant number 1-R01-DK58533-01. We also acknowledge support by the Merck Genome Research Institute (MGRI).

REFERENCES

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Cohen,J. (1988) *Statistical power analysis for the behavioral sciences*, 2nd edn, Erlbaum, Hillsdale, NJ.

Dillon,W.R. and Goldstein,M. (1984) *Multivariate Analysis*. Wiley, New York.

Dudoit,S. *et al.* (2001) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* (to appear), *UC Berkeley Statistics Technical Report.*

Dudoit,S., Fridlyand,J. and Speed,T.P. (2000) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.* (to appear), *Technical Report* **576**.

G power reference material. http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/reference/reference_manual_02.html#noncentral.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Holter,N.S. *et al.* (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Natl Acad. Sci. USA*, **97**, 8409–8414.

Johnson,R.A. and Wichern,D.W. (1992) *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey.

Kraemer,H.C. and Thiemann,S. (1987) *How Many Subjects? Statistical Power Analysis in Research*. Sage, Newbury Park, CA.

Lachenbruch,P.A. and Mickey,M.R. (1968) Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11.

Lockhart,D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Mace,A.E. (1974) *Sample-size determination*. Krieger, Huntington, NY.

Olson,C.L. (1974) Comparative robustness of six tests in multivariate analysis of variance. *J. Amer. Stat. Assoc.*, **69**, 894–908.

Other F-tests. http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/reference/reference_manual_09.html#t3.

Pan,W., Lin,J. and Le,C. (2001) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Biostatistics*, University of MN Technnical Report.

SAS/STAT User's Guide (1989) 4th edn, SAS Institute, Cray, NC.

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

Stephanopoulos,G. *et al.* (2002) Mapping Physiological States from Microarray Expression Measurements. *Bioinformatics*, in Press.

Storey,J.D. and Tibshirani,R. (2001) *Estimating false discovery rates under dependence with applications to DNA microarrays*, Stanford Technical Report.

Thall,P.F. (1995) *Recent Advances in Clinical Trial Design and Analysis*. Kluwer, Boston.

Thomas,J. *et al.* (2001) An efficient and robust statistical modeling approach to discover diffentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 5116–5121.

West,M. *et al.* (2001) DNA microarray data analysis and regression modeling for genetic expression profiling. *CAMDA*.

Zhao,G. and Maclean,A.L. (2000) A comparison of canonical discriminant analysis and principal component analysis for spectral transformation. *Photogramm Eng. Rem. S*, **66**, 841–847.

Zien, *et al.* (2001) Microarrays: How many do you need? *Proceedings of the Sixth Annual International Conference on Computational Biology. RECOMB*.