BIOINFORMATICS

Vol. 18 no. 7 2002
Pages 961–970

Deriving quantitative conclusions from microarray expression data

Adam B. Olshen [1, 2] and Ajay N. Jain [1,]*

[1]Comprehensive Cancer Center, Cancer Research Institute, and Department of Laboratory Medicine, University of California, San Francisco, CA 94143-0128, USA and [2]Department of Epidemiology and Biostatistics, Memorial Sloan–Kettering Cancer Center, 1275 York Avenue, NY 10021, USA

Received on July 21, 2001; revised on December 10, 2001; January 19, 2002; accepted on January 24, 2002

ABSTRACT

Motivation: The last few years have seen the development of DNA microarray technology that allows simultaneous measurement of the expression levels of thousands of genes. While many methods have been developed to analyze such data, most have been visualization-based. Methods that yield quantitative conclusions have been diverse and complex.

Results: We present two straightforward methods for identifying specific genes whose expression is linked with a phenotype or outcome variable as well as for systematically predicting sample class membership: (1) a conservative, permutation-based approach to identifying differentially expressed genes; (2) an augmentation of K-nearest-neighbor pattern classification. Our analyses replicate the quantitative conclusions of Golub *et al.* (*Science*, **286**, 531–537,1999) on leukemia data, with better classification results, using far simpler methods. With the breast tumor data of Perou *et al.* (*Nature*, **406**, 747–752,2000), the methods lend rigorous quantitative support to the conclusions of the original paper. In the case of the lymphoma data in Alizadeh *et al.* (*Nature*, **403**, 503–511,2000), our analyses only partially support the conclusions of the original authors.

Availability: The software and supplementary information are available freely to researchers at academic and non-profit institutions at http://cc.ucsf.edu/jain/public.

Contact: ajain@cc.ucsf.edu

INTRODUCTION

Biological data such as those derived from DNA microarrays present unique challenges. Many thousands of measurements can be made on single biological samples (Alizadeh *et al.*, 1998; Iyer *et al.*, 1999), complicating the process of identifying *specific* patterns or correlations. Generally, the goal of such experiments is to identify some subset of items being measured that share a particular property (e.g. co-expression) or whose behavior is correlated with some outcome (e.g. patient survival).

Even with the largest practical sample sizes, array-based experiments involve vastly more measurements than samples. The most popular methods of analysis involve clustering, typically hierarchical clustering (Eisen *et al.*, 1998). Cluster-based methods by themselves generally do not directly yield statistically meaningful results. While cluster-based approaches can be useful for generating hypotheses, one generally cannot conclude that if two items belong to the same cluster, then their behavior is quantitatively related in a statistically significant sense. More quantitative approaches to microarray data analysis have been published (Golub *et al.*, 1999; Hastie *et al.*, 2001; Ben-Dor *et al.*, 2000), but the set of methods are quite diverse, and often very complex. In this paper, we present straightforward methods for addressing common questions about the information contained in microarray-based expression data.

There are typically two types of questions of interest: (1) questions about the variables; and (2) questions about the biological samples. With the former, we are concerned with which genes or genomic loci, either individually or in groups, are related to some phenotype, outcome, biological pathway, or other concept in a quantitative way. With the latter, we are concerned with making predictions about the behavior of samples, e.g. probability of treatment failure within a specified time-frame given expression data from a tumor. Type 2 questions quickly give rise to type 1 questions: given a predictive model, one wants to understand which particular variables are responsible for the accurate predictions. Direct statistical approaches that account for the high ratio of variables to samples can address questions of the first type. Pattern classification and machine learning approaches address questions of the second type.

We present the application of two methods to three

*To whom correspondence should be addressed.

published microarray expression data sets. The first, a simple and conservative approach to permutation testing, is used to address type one questions. The second, a simple augmentation of K-nearest-neighbor pattern classification, is used to address type two questions. We also employ a clustering method to illustrate the data. For data sets with strong signals, these methods provide direct and quantitative substantiation of many hypotheses suggested by visualization-based approaches.

METHODS

Description of the analytical methods follow a discussion of normalization and filtering of microarray data.

Data normalization and filtering

There are two popular technologies for generating large-scale expression data: Affymetrix's oligonucleotide-based hybridization, and spotted cDNA-based comparative hybridization (Schena *et al.*, 1995; Lockhart *et al.*, 1996). We analyze three data sets: one Affymetrix-based and two cDNA-based sets. Each of the cDNA-based sets used a common reference sample from which to compute ratios (Perou *et al.*, 2000; Alizadeh *et al.*, 2000).

Data normalization of some sort is required with the cDNA-based method, since the method relies on the ratio of hybridization of a test sample to that of a reference sample and there is no experimental way to precisely control the stoichiometry of the two samples. Common practice is to center the data by the median or mean ratio, and possibly to scale the data by the standard deviation, and in what follows, we indicate which normalization method is employed. We consider these data in log-space for two reasons: (1) this helps reduce the effects of outliers on normalization; and (2) the distribution of log expression values in self/self hybridizations is symmetric about 0, but the distribution of ratios exhibits skew to the right of 1. For the Affymetrix data, normalization may be beneficial to reduce the effects of varying dynamic range from sample to sample. However, such normalizations are not technically required, and we have not employed any for analysis of these data. When displaying Affymetrix data, however, we transform the data into the space of log-ratios. This is done by taking the $\log_2$ of each gene divided by the median of that gene across a set of experiments.

Quantitation methods for expression often yield quality measures that are associated with each expression measurement. Depending on a large variety of factors, many different criteria are employed to eliminate data that are likely to be noisy. Since this process requires knowledge of control experiments that are not often reported, we have either followed the authors' conventions or simply considered the data that they have provided after their preferred filtering procedures.

Hypothesis testing: questions about genes

In considering questions about the relationship of individual genes to phenotype, survival, or any other variable, one possibility is to perform the appropriate statistical test on every gene individually. Those genes whose p-values are low enough (say below 0.05) would be termed significant. Depending on context, for two-class data as we have in our three examples, the two-sample t-test (with equal or unequal variances), a rank-based test such as the Wilcoxon rank sum test, or the chi-square test could be applied (Rice, 1988).

This approach has a serious problem in a microarray setting. For example, suppose that one has 200 genes, and therefore performs 200 tests at the standard $p = 0.05$ level. If there were truly no underlying difference in expression levels between the two classes for any gene, one would still expect to observe approximately 10 significant genes (200×0.05). In fact, under the assumption that the tests are statistically independent, the probability of getting at least one significant result is $1–0.95^{200} = 0.999\,96$. This problem, termed the *multiple comparisons problem*, is of special concern in gene expression studies.

One proposed solution to the multiple comparisons problem is to divide the level, that is to say, the probability of type I error, of the test by the number of tests attempted. This method is called the *Bonferroni method* (Rice, 1988). For 200 genes, the cutoff for significance would be $0.05/200 = 0.00025$. This adjustment offers a conservative method for significance assessment, but it requires the use of a statistical test with a known distribution.

We use instead a permutation-based method. The approach computes p-values that are adjusted for the number of tests undertaken but in a way that can be less conservative than the Bonferroni method. Specifically, p-values for specific genes are based on the largest test statistic expected for the given data set if there were no relationship between any gene and a given outcome. The algorithm is as follows (for testing at the 0.05 level) :

(1) fix a statistic by which differences between classes can be assessed;

(2) compute the statistic for every gene;

(3) randomly permute the class labels (thus breaking the relationship between the expression data and the class designation);

(4) compute the same test statistic for every gene as in (2) using the permuted labels. Save the maximum statistic;

(5) repeat (3) and (4) many times (say 10 000 times, or for all permutations of the labels if feasible);

(6) find the 95th percentile of the distribution of maxima. Term this the *critical value*;

(7) any gene with an associated statistic that is bigger than the critical value determined in (6) is termed significant.

In practice, it often is the case that the critical values determined by this method are nearly as conservative as those based on the Bonferroni adjustment, but there are two benefits to this approach. The first is that any test statistic can be used, not just those with known distributions. The second is that it is easy to implement and runs in reasonable time on standard desktop hardware.

There are three important subtleties. First, the critical value determined above can be conservatively applied to any gene whose test statistic exceeds the critical value, not just the maximal test statistic. Suppose that the mth best gene has a test statistic greater than our critical value. If we recomputed the critical value using the mth best test statistic on each permutation iteration, the 95th percentile of that distribution would be *less* than the critical value determined previously. Since our cutoff for significance is higher, our method is more conservative than that method would be. Second, in gene expression data, many genes have strong correlations with one another from sample to sample. The procedure above does not penalize this redundancy since the maximum test statistic on each permutation will not change with the addition of more variables correlated with the initial ones. Note that the Bonferroni correction yields a more pessimistic result in the case of a large number of very strongly correlated variables versus the corresponding smaller set of uncorrelated variables (see **Results and Discussion**). Third, as the number of samples becomes extremely small, the implicit assumption that the distributions of a statistic for each gene will be roughly comparable may begin to fail. In these cases, the method will tend to yield overly conservative critical values.

Westfall and Young (1993) describe several resampling methods to address the multiple comparisons problem. Others have applied permutation-based methods in assessing significance of gene expression data (Dudoit *et al.*, 2000; Tusher *et al.*, 2001; Golub *et al.*, 1999). We have applied them successfully to data from comparative genomic hybridization (Jain *et al.*, 2001) and microarray-based comparative genomic hybridization (Wilhelm *et al.*, 2002). Some of the referenced methods are less conservative than what we present here, but they are also less straightforward.

Classification: questions about samples

Another important problem is to accurately predict sample classes based on patterns of expression across multiple genes. There is an enormous body of literature in pattern classification and machine learning germane to such problems (Duda and Hart's classic text; Duda and Hart, 1973 or more recent works by Mitchell, 1997, or Ripley, 1996, cover many methods in detail). Despite the wealth of competing methods, it is often the case that a very simple method, K-nearest-neighbors (Duda and Hart, 1973) is difficult to beat. The class of an observation of unknown class is based on the classes of its K nearest neighbors. If the neighbors are not all of the same class, the class is assigned based on a majority vote. Advantages of the KNN approach include ease of implementation, minimal parameterization, and straightforward interpretation of the resulting classification.

In order to construct a classifier, we must choose K, a distance metric, and (optionally) a method to select a subset of genes S_G of size G on which to compute distances. For the latter we use the absolute value of the two-sample t-statistic with equal variances. For the ith gene the statistic is

$$t_i = \frac{|\overline{X_i} - \overline{Y_i}|}{S_i\sqrt{\frac{1}{n_i} + \frac{1}{m_i}}},$$

where $\overline{X_i}$ and $\overline{Y_i}$ are the average (log) expression values corresponding to class 1 and 2, respectively, n_i and m_i are the number of non-missing samples, and

$$S_i = \sqrt{\frac{(n_i - 1)U_i + (m_i - 1)V_i}{n_i + m_i - 2}},$$

where U_i and V_i are the variances of the (log) expression values. Because many genes will have no predictive power and therefore only contribute noise to the classification, it is often beneficial to classify using subsets of the genes. In this work, we have employed the scaled Euclidean distance metric. Under this distance metric, the vector of expression levels for a gene is normalized to length one before computing distance so that single genes do not have disproportionate influence on the distance calculation. So the distance $D(j, k)$ between the jth and kth sample is

$$D(j, k) = \sqrt{\sum_{i \in S_G} (Z_{ij} - Z_{ik})^2},$$

where Z_{ij} is the scaled (log) expression value for the ith gene on the jth sample and Z_{ik} is the same for the ith gene and kth sample.

Assessing predictive performance of classifiers can be done by blind testing of performance on unseen data and cross-validation. The former is preferable, but it is only practicable in situations where there is a sufficient amount of data to form a training and test set. One of the data sets analyzed here has that characteristic. For two of the sets,

there are few samples and no obvious basis for splitting into training and test, so we use leave-one-out cross-validation. The cross-validation procedure designates one sample as a test set and the remaining N-1 as a training set. The K-NN rule developed on the training set is applied to the test sample. This procedure is repeated with every observation serving once as the test sample, so a class prediction is made for every sample, without benefit of any information about that sample.

Even though the K-NN approach has few parameters to choose, cases arise where classification performance is highly sensitive to the choice of these parameters. So, to avoid all contaminating effects of knowledge of the training data in predicting new data, we use a systematic method for determining K (the number of neighbors) and G (the size of the gene subset) from the training set in each cross-validation iteration above. So, just on the training set:

(1) choose a value of K and G;

(2) drop one sample and predict its class based on the K-NN rule for the remaining samples;

(3) repeat (2) for every sample and calculate the number of errors over the whole training set;

(4) repeat 1–3 for various combinations of (K, G);

(5) pick the (K,G) that gives the smallest error in (3). Employ this (K, G) for this iteration of cross-validation;

Note that even given a successful prediction result, one cannot specifically ascribe a causal relationship between the genes forming the classifier and the sample classes. This is because *many* different sets of genes may generate equivalent classification performance (this has been termed the credit-assignment problem; (Mitchell, 1997)).

Dudoit *et al.* (2001) found that K-NN was generally the best classifier among a group of statistical choices for three gene expression data sets. Ben-Dor and colleagues (2000) used K-NN as a starting point for further exploration of classification methods for these types of data. A number of other papers include classification methods in the context of gene expression (Brown *et al.*, 2000; Moler *et al.*, 2000; Golub *et al.*, 1999).

Clustering

As is common practice in analyzing microarray expression data, we also employ hierarchical clustering (Eisen *et al.*, 1998; Duda and Hart, 1973). We use this primarily to visualize the data and to explore the relationships among distance metric, variable selection, and classification. All the cluster diagrams displayed here were constructed using single linkage with the t-statistic used for variable selection and scaled Euclidean as the distance metric.

RESULTS AND DISCUSSION

LEUKEMIA

Golub *et al.* (1999) reported on the expression of 7070 genes for 72 subjects using Affymetrix high density oligonucleotide arrays (data available at http://www.genome.wi.mit.edu/MPR/). Each sample is identified by leukemia type: acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). The data were collected in two separate patient series, allowing for a natural definition of training and testing sets: a training set of 38 (27 ALL vs 11 AML) and a test set of 34 (20 ALL vs 14 AML). The authors applied clustering, neighborhood analysis, and a voting scheme to address visualization, correlation, and predictive classification, respectively. The numerical prediction results were 29 of 34 subjects correctly classified into the ALL/AML subtypes, with five not classified, being considered too close to call.

We applied the methods described above to these data without filtering or normalization. A t-test (equal variance), using an unadjusted critical value of 2.02 ($p = 0.05$), yields 1636 genes (23.1%) from the training set that are nominally significant (meaning without any adjustment). The permutation-based adjustment, which has a critical value of 5.16, yields 40 of these 1636 genes as significant (gene list available via http://cc.ucsf.edu/jain/public). To assess the appropriateness of the permutation-based critical value, we considered the t-test for the significant genes on the test set. Note that the test set was on a slightly different patient population, with different proportions of ALL and AML samples.

Of the genes that were significant on the training set with the unadjusted critical value, only 585 (36%) were nominally significant on the test set. However, of the 40 that were significant after adjustment, 32 (80%) were nominally significant. Given that the uncorrected approach identified a majority of genes that were not nominally significant on the test set, the permutation-based cutoff appears to be more appropriate. The Bonferroni adjusted critical value is 5.25 for these data, just slightly higher than the value we derived from the permutation adjustment. The degree of internal correlation among the genes is not high. By taking 707 randomly selected genes from the original data and adding 9 noisy copies of each (10% uniform multiplicative noise), we constructed a new data set of 7070 genes in which the internal correlation is high. In this case, while the Bonferroni critical value is the same, the permutation adjusted critical value is 4.58, thereby accounting for the internal structure in the data.

Using the 100 genes with highest absolute t-statistics from the training set, we clustered the samples from both the test and training set data (note that many different gene set sizes work well). The results can be seen in Figure 1. The test samples cluster strongly with the training samples

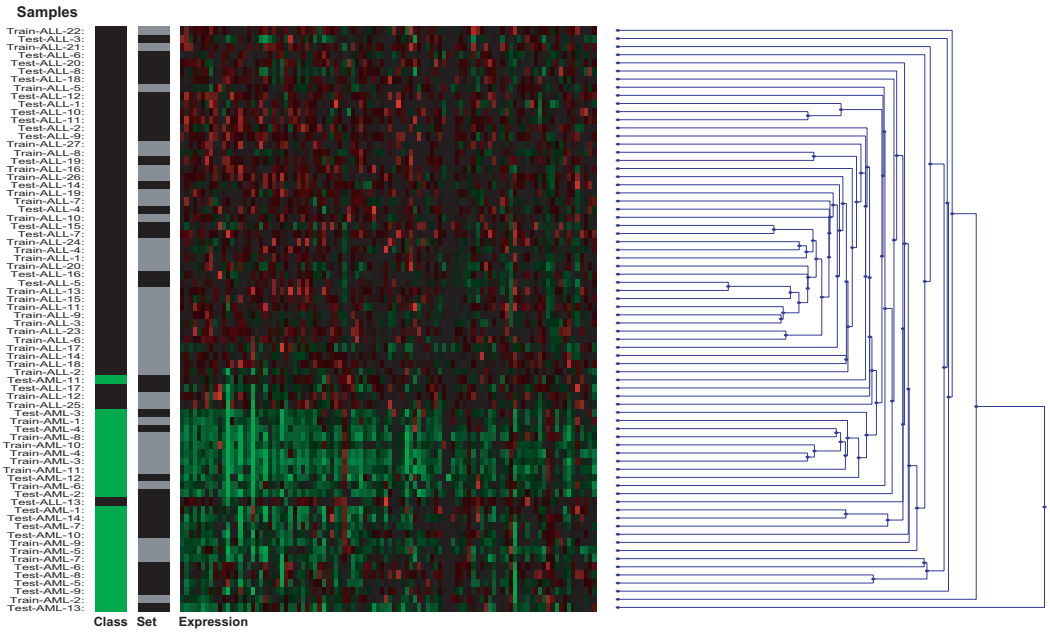

Fig. 1. Clustering of the leukemia data using the best 100 genes chosen on the training set. For the Class variable, black is ALL and green is AML. For the Set variable, gray is the training set and black is the test set. The data was standardized for display for each gene by taking the log2 of the original expression values divided by the median value across samples. Green represents expression values above the median and red represents values below the median. The clustering of samples by class is very strong. Note that almost all the best genes have higher expression in AML than in ALL.

of their own classes. This clear separation of classes in terms of expression was borne out by our systematic classification experiments. We used cross-validation on the training set to choose the number of genes G, selected by the rank order of the t-statistic, (between 10 and 7070) and the number of nearest neighbors K (1, 3, or 5) in our K-NN classifier. The best performing classifier on the training set (37 out of 38 correct) had 5000 genes and 5 neighbors. This classifier applied to the test set correctly classified 32 out of 34 samples (94%), with all 20 ALL samples correct and 12 out of 14 AML samples correct.

We also explored the robustness of classification performance under different choices of the number of genes selected and number of neighbors used (see Figure 2). The ALLs are perfectly classified for almost any set of parameters, except for large G and $K = 1$. The AMLs are more difficult to classify, with between 10 and 14 correctly classified. The mistakes made by different classifiers are focused on two of the AML samples. Note that for small gene sets (e.g. the 40 most significantly differentially expressed), randomly selected gene subsets of the same size have poorer classification performance (data not shown). A scheme using voting among classifiers over all the parameter sets examined yielded only a single (AML) test sample classified incorrectly and one (AML) ambiguity. We also experimented with rank-based methods both for choosing genes and calculating distance, and got similar results.

There appear to be strong effects on gene expression related to the phenotypic difference of lymphoblastic versus myeloid leukemia. In this case, even with an unfavorable ratio of genes to samples (about 200:1 on each set), it is possible to identify specific genes that are correlated with class in a univariate sense. The distinctive patterns are sufficiently regular that excellent classification performance is achievable using a very straightforward approach.

Breast Cancer

Perou *et al.* (2000) used cDNA microarrays to assess the expression of 8102 genes on breast tumor samples from 42 individuals (data available at http://genome-www.stanford.edu/molecularportraits/). Of these, 38 have data for the clinically determined estrogen receptor (ER) tumor status and 32 have clinical ERB-B2 tumor status. These immunohistochemistry-based measurements of protein level are scored as either absent or present (*negative* or *positive* below). In addition to exploring clusterings of the data, which reveal a diversity of expression signatures, the authors indicated there was a

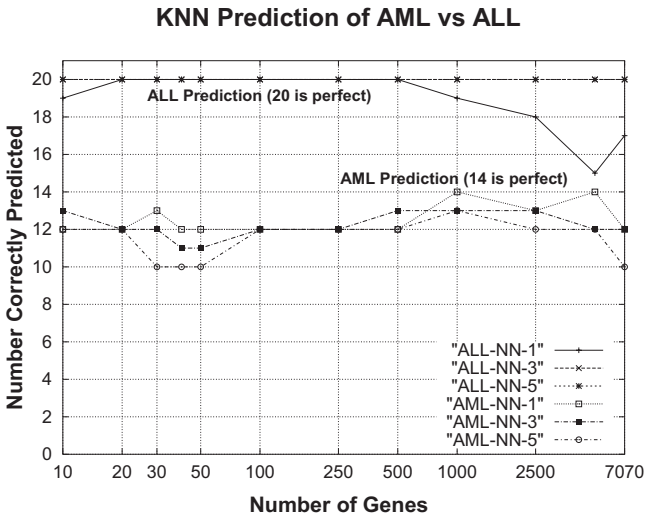

Fig. 2. K-NN prediction of the 34 members in leukemia data test set as K and the number of predictors varies. Note that the ALL samples are perfectly predicted over a wide range of parameters. The AML samples average two incorrect predictions. There is relatively little influence of either K or G, with some reduction in performance being seen at high G with K = 1.

strong correlation between the gene expression of ESR1 (estrogen receptor 1) and ERB-B2 (avian erythroblastic leukemia viral oncogene homolog 2) and the clinically measured proteins.

We carried out the same type of analysis described for the leukemia data, but here we focused on the genes that differentiated the samples labeled by the clinically determined ER and ERB-B2 status. We analyzed the data set of 1753 genes provided by the authors that can be found in Figure 1 of their paper. The genes in this set were characterized by having little missing data (less than 20%) and a least three samples that varied four-fold from the median abundance. Once the genes were selected, samples and genes were centered to have medians of zero. For the subjects that had more than one profile, the pre-surgery one was used. There are 159 genes (10.1%) that are significantly different between the ERB-B2 positive and negative samples using the two-sample t-test with equal variances. After a permutation-based determination of the critical value, 4 genes are significant. Two are the actual ERB-B2 cDNAs (a third copy of the ERB-B2 cDNA is borderline significant). The others are MLN64 (steroidogenic acute regulatory protein related) and GRB7 (growth factor receptor-bound protein 7). Both genes are known to be co-amplified with ERB-B2 (Kauraniemi *et al.*, 2001), and GRB7 is a known target for ERB-B2 (Jane *et al.*, 1997). For ER, 326 genes (18.6%) are significant without accounting for multiple comparisons. After the permutation-based adjustment, only four genes are significant: two copies of Estrogen Receptor 1 and two (of four) copies of GATA-binding Protein 3. While it should be unsurprising that expression and protein levels directly correlate, the correlation between expression and clinical status of ER and ERB-B2 is strong enough to be distinguished from background, even among thousands of genes, requiring no prior knowledge of the relationship between the genes and gene products. Note that the permutation adjustment is technically unnecessary for the ERB-B2 and ESR1 expression measurements since there is an independent reason to assess those correlations.

A cluster diagram of expression and ERB-B2 status based on the best 100 genes selected using the t-statistic can be found in Figure 3. The clustering largely segregates the classes. There is a group of eight genes, including the three ERB-B2 genes, that have high expression for half of the positive samples. Cross-validated K-NN correctly predicts 28/32 (88%), with all 20 samples that are ERB-B2 negative correctly predicted and 8/12 that are positive correctly predicted. Again this should not be surprising, since the gene set from which the selection takes place contains ERB-B2. However, after discarding the five genes with labels related to ERB-B2, 19 out of 20 negative samples are classified correctly; and 7 out of 12 positive samples are classified correctly. So, accurate prediction is possible without using the ERB-B2 genes. It is not clear, however, that the genes used in these predictive classifiers are causally related to the functional ERB-B2 pathway.

A comparable analysis with the ER data yielded similar results. The clustering in Figure 4 shows that the samples cluster together by class. The K-NN classifier correctly predicts 89% of the samples correctly, with 28 of 29 ER negative samples correct and 6 of 9 ER positive samples correct. Eliminating the three genes that are labeled Estrogen Receptor 1 leads to exactly the same classification. Thus ESR1 genes are not necessary for accurate prediction of ER protein status.

These results are encouraging in that it is possible to detect effects that are clearly causal from large-scale gene expression experiments: the relationship between ERB-B2 and ER at the protein level and the expression level. However, since strong classification performance predicting a phenotype is possible absent the obvious causal genes, caution should be taken in interpreting even clearly positive analytical results.

Lymphoma

Alizadeh *et al.* (2000) used cDNA microarrays to assess the expression of approximately 18,000 genes on 96 samples from lymphoma patients (data available at http:/llmpp.nih.gov/lymphoma). The authors performed two-way hierarchical clustering (first on genes and then

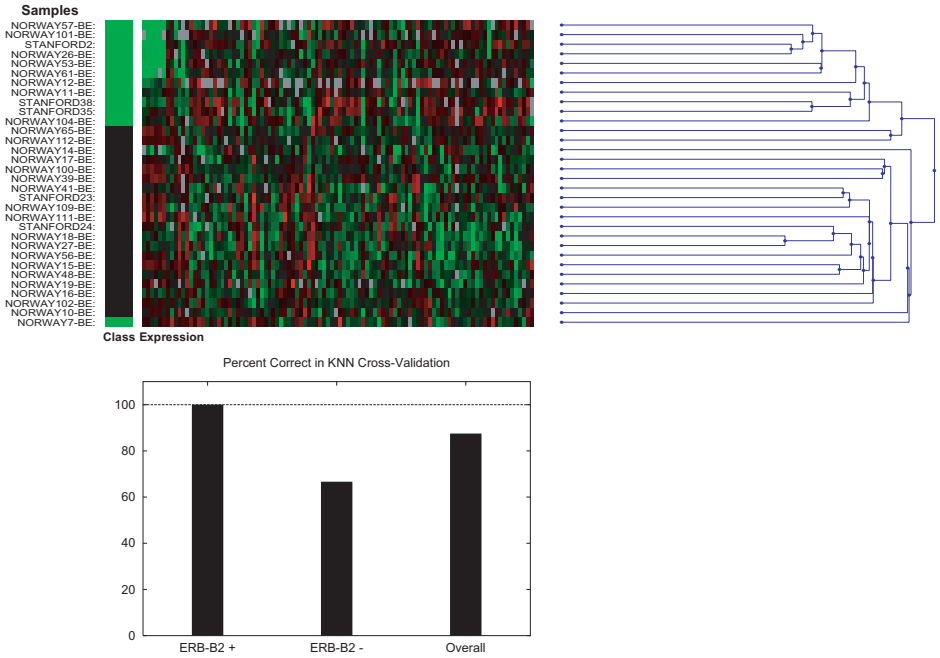

Fig. 3. Top: Cluster diagram for the breast samples based on the 100 genes that best separate the ERB-B2 positive samples from the ERB-B2 negative samples. The class variable is ERB-B2 protein (black for negative, green for positive). Only one positive sample does not cluster with members of its own class. Bottom: Plot of classification performance under cross-validation for prediction of ERB-B2 clinical status based on gene expression data. Overall performance is 88%.

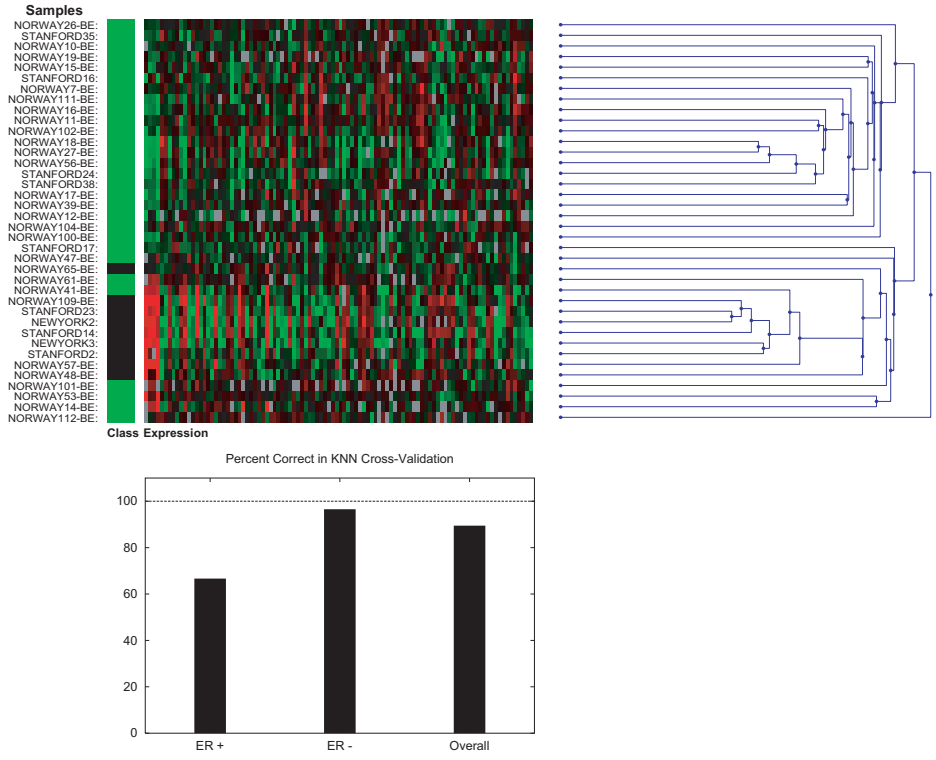

Fig. 4. Top: Cluster diagram for the breast samples based on the 100 genes that best separate ER positive samples from ER negative samples. The class variable is ER protein (black for negative, green for positive). The cluster diagram shows a strong separation between the positives and negative samples. Bottom: Plot of classification performance under cross-validation for prediction of ER clinical status based on gene expression data. Overall performance is 89%.

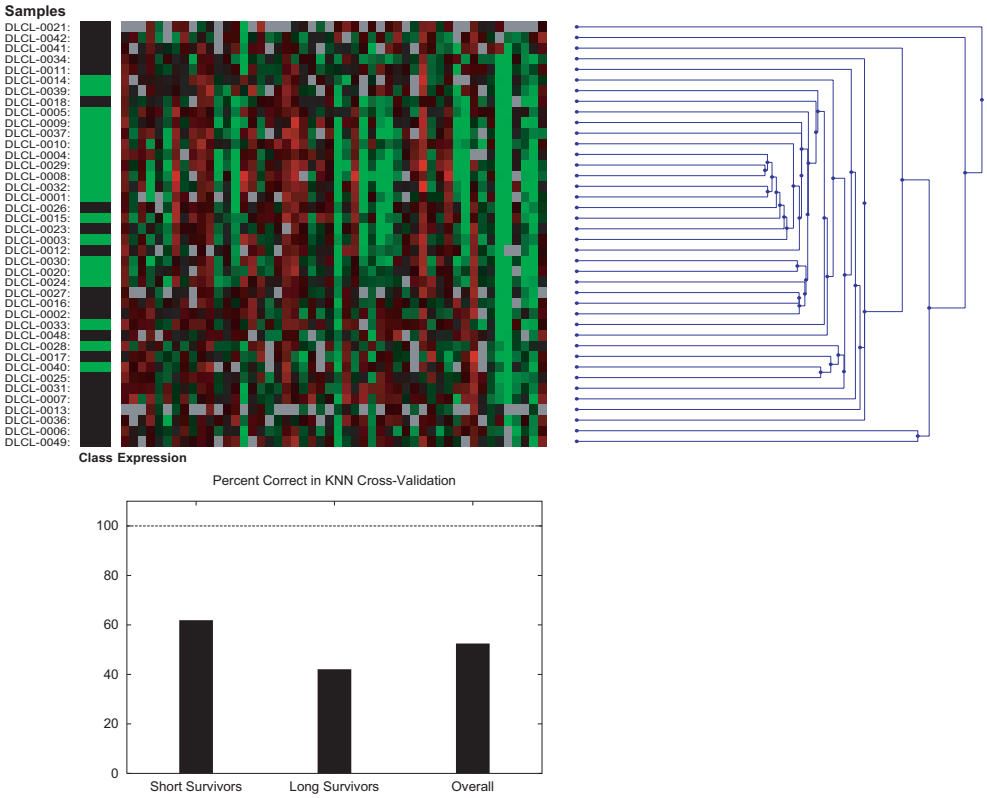

Fig. 5. Top: Cluster diagram for the lymphoma samples using the best 50 genes. The class variable is black for short survivors and green for long survivors. Note the lack of clear expression differences between the classes. Bottom: Plot of classification performance under cross-validation for prediction of survival class based on gene expression data.

samples). From the original clustering, they identified 148 genes that define the germinal center B-cell signature. They then clustered only the diffuse large B-cell (DLBCL) samples using these genes. From the second round of clustering, they identified two sub-types of DLBCL, which they termed *GC B-like* and *Activated B-like*. They found that subjects with the GC B-like form lived significantly longer than those with the Activated B-like form. Among the subset of 20 patients with low clinical risk, they found this same relationship.

We focus here on the relationship between gene expression and survival, seeking a direct link. There is a natural partitioning of the patient samples based on survival times. Of the 40 DLBCL samples, there were 21 subjects who had survival times less than 40 months, and 19 subjects who had survival times of at least 50 months. We call the former group *short survivors* and the latter group *long survivors*. We considered all genes with non-missing expression values for at least 10 short survivors and at least 10 long survivors, resulting in a data set of 6712 genes. The $\log_2$ expression values for each sample were normalized to mean zero and variance one.

Our methods failed to find significant univariate expression differences between the long and short survivors. While there are 471 genes that were significantly different between the two groups using the unadjusted t-test, there were zero significant genes after a permutation adjustment. Figure 5 shows the clustering of these data based on the 50 genes with the highest absolute t-statistics. The separation between the survival classes is not very strong considering that the genes were chosen to distinguish them. A formal cross-validated K-NN procedure predicts 21 out of 40 correctly, no different from chance when there is no signal. This result is at odds with the clear observation in the original paper relating expression to survival.

Given this negative result, we looked more closely at Alizadeh's methods using their data set of 4026 (4005 of which were in our data set) genes that went into their initial clustering. Of the 148 genes that define the germinal center B-cell signature, only 39 have significantly different expression between the survival groups by a t-test, uncorrected for multiple comparisons. So our direct method of choosing gene subsets that separate the survival-based

classes would not find this group of genes. Among the subjects with GC B-Like DCL, 12 were long survivors and 7 short survivors. Among those with Activated B-Like DCL, 5 were long survivors and 16 were short survivors. The putative subtypes segregate our survival classes with just 70% accuracy.

We wanted to assess the degree to which we could support the notion that the two clusters of DLBCL were true subclasses, since the hypothesis put forward is that these are truly biologically distinct types of lymphoma. We labeled the samples according to membership in the GC B-Like and Activated B-Like DCL subclasses, as defined by Alizadeh *et al.* using specific gene sets derived from their cluster analysis. Our results suggest that the distinction between these subtypes is real. After a permutation correction for multiple comparisons, 36 genes were found to be significant (19 of which are found in the author's cluster). The cross-validated K-NN classified correctly assigned 37 out of 40 samples to the correct DLBCL subclasses. Classification performance was robust across a wide variety of parameters (including no selection of a gene subset). Our analyses support the notion that there are two distinct subclasses of DLBCL, as proposed by Alizadeh *et al.*, but it is not clear to what extent these subclasses are in fact defined by the 148 genes of the germinal center B-cell signature.

CONCLUSIONS

We have presented the application of two straightforward methods for quantitative analysis to three expression microarray data sets: a simple and conservative approach to permutation testing and a simple augmentation of K-nearest-neighbor pattern classification. Our analyses replicate the quantitative conclusions of Golub *et al.* (1999), with slightly better classification results, using far simpler methods. In the case of Perou *et al.* (2000), our methods lend quantitative support to the conclusions of the original paper, but they also yield an ambivalent message. While it is possible to detect statistically significant causal relationships against the background of a large number of gene expression measurements, it is also possible to construct predictive models using subsets of genes lacking the primary causal biological agents. In the case of the data in Alizadeh *et al.* (2000), our analyses support the existence of the subtypes hypothesized by the authors, but the components of gene expression involved in the subtypes are unclear.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the National Cancer Institute (CA89520 and CA64602).

REFERENCES

Alizadeh,A., Eisen,M.B., Brown,P.O. and Staudt,L.M. (1998) Probing lymphocyte biology by genomic-scale gene expression analysis. *J. Clin. Immunol.*, **18**, 373–379.

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

Ben-Dor,A., Bruhn,L.K., Friedman,N., Nachman,I., Schummer,M. and Yakhini,Z. (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–583.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristiani,N., Sugnet,C., Furey,T.S., Ares,Jr,M. and Haussler,D. (2000) Microarray gene expression data using support vector machines. *PNAS*, **97**, 262–267.

Duda,R.O. and Hart,P.E. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T. (2000) Statistical methods for identifying differentially expressed genes. Unpublished (*Berkeley Stat. Dept. Technical Report #578)*.

Dudoit,S., Fridlyand,J. and Speed,T. (2001) Comparison of discrimination methods for the classification of tumors using gene expression data. To appear in *JASA*.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**, 14863–14868.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hastie,T., Tibshirani,R., Botstein,D. and Brown,P. (2001) Supervised harvesting of expression trees. *Genome Biol.*, **2**, 1–12.

Iyer,V., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C.F., Trent,J.M., Staudt,L.M., Hudson,Jr,J., Boguski,M.S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.

Jain,A.N., Chin,K., Borresen-Dale,A., Erikstein,B.K., Lonning,P.E., Kaaresen,R. and Gray,J.W. (2001) Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with P53 status and patient survival. *PNAS*, **98**, 7952–7957.

Jane,P.W., Lackmann,M., Church,W.B., Sanderson,G.M., Sutherland,R.L. and Daly,R.J. (1997) Structural determinants of the interaction between the erbB2 receptor and the Src homology 2 domain of Grb7. *J. Biol. Chem.*, **272**, 8490–8497.

Kauraniemi,P., Barlund,M., Monni,O. and Kallioniemi,A. (2001) New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. *Cancer Res.*, **61**, 8235–8240.

Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittman,M., Wang,C., Kobayashi,M., Horton,H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Mitchell,T.M. (1997) *Machine Learning*. McGraw Hill, New York.

Moler,E.J., Chow,M.L. and Mian,I.S. (2000) Analysis of molecular profile data using generative and discriminative methods. *Physiol. Genomics*, **4**, 109–126.

Perou,C.M., Sorlie,T., Eisen,M.B., Van de Rijn,M., Jeffrey,S., Rees,C.A., Pollock,J.R., Ross,D.T., Johnsen,H. *et al.* (2000) Molecular portraits of human breast tumors. *Nature*, **406**, 747–752.

Rice,J. (1988) *Mathematical Statistics and Data Analysis*. Wadsworth, Pacific Grove, CA.

Ripley,B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

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

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

Westfall,P.H. and Young,S.S. (1993) *Resampling-based Multiple Testing*. Wiley, New York.

Wilhelm,M., Veltman,J.A., Olshen,A.B., Jain,A.N., Moore,D.H., Presti,J.C., Kovacs,G. and Waldman,F.M. Array based CGH for the differential diagnosis of renal cell cancer. *Cancer Res.*, (in press).