Class Prediction and Discovery Using Gene Expression Data

Donna K. Slonim, Pablo Tamayo, Jill P. Mesirov, Todd R. Golub, Eric S. Lander *

September 30, 1999

Abstract

Classification of patient samples is a crucial aspect of cancer diagnosis and treatment. We present a method for classifying samples by computational analysis of gene expression data. We consider the classification problem in two parts: *class discovery* and *class prediction*. Class discovery refers to the process of dividing samples into reproducible classes that have similar behavior or properties, while class prediction places new samples into already known classes. We describe a method for performing class prediction and illustrate its strength by correctly classifying bone marrow and blood samples from acute leukemia patients. We also describe how to use our predictor to validate newly discovered classes, and we demonstrate how this technique could have discovered the key distinctions among leukemias if they were not already known. This proof-of-concept experiment paves the way for a wealth of future work on the molecular classification and understanding of disease.

*Whitehead/MIT Center for Genome Research, One Kendall Square bldg. 300, Cambridge, MA 02139. Contact author's email address: slonim@genome.wi.mit.edu.

1 Introduction

Classification of patient samples is a crucial aspect of cancer diagnosis and treatment. Current classification methods rely primarily on the cancer's tissue of origin (for example, whether a tumor first developed in the lung or the brain) and on the microscopic appearance and location of cancerous cells. However, there are many clinically-relevant distinctions that can only be made in hindsight. For example, tumors of identical appearance may progress at very different speeds, some growing aggressively and demanding equally aggressive treatment, others remaining so inactive that the best course might be no treatment at all. Unfortunately, these often can be distinguished only by observing the patient over time and discovering whether or not the initial treatment was sufficiently aggressive. Thus, researchers continue to search for new methods of classification that might predict the course of the disease at the time of diagnosis.

Recent technological advances in monitoring gene expression may help. Although the blueprints encoding all human genes are present in each cell, only a fraction of the proteins they can produce are active in any particular cell. The process of transcribing a gene's DNA sequence into the RNA that serves as a template for protein production is known as *gene expression*. A gene's expression level indicates the approximate number of copies of that gene's RNA produced in a cell; this is thought to correlate with the amount of the corresponding protein made. While the traditional technique for measuring gene expression, the Northern blot assay, is labor-intensive and produces only an approximate quantitative measure of expression, new technologies have greatly improved the resolution and the scalability of gene expression monitoring. "Expression chips," manufactured using technologies derived from computer-chip production, can now measure the expression of thousands of genes simultaneously.

It has been suggested that gene expression may provide the additional information needed to improve cancer classification and diagnosis. In this paper, we present a proof-of-concept study supporting the idea. Our expression experiments are performed using Affymetrix oligonucleotide arrays [9, 14] that measure the expression of 6817 known human genes in each patient sample. However, our methods are not restricted to any particular microarray technology.

We present a method for performing classification of patient samples by gene expression analysis. We separate the general classification problem into two challenges: *class discovery* and *class prediction*. Class discovery refers to the process of dividing samples into groups with similar behavior or properties. For example, the determination of a system for grading tumors by degree of progression is a class discovery process. In contrast, the class prediction problem corresponds roughly to diagnosis: given a set of known classes, determine the correct class for a new patient. To see why this is called "prediction" rather than "diagnosis," consider the case where classes are based on how a patient will respond after two years of treatment with a certain drug. A class predictor for this problem would suggest whether the patient would benefit from treatment *before* performing the two-year experiment to test the prediction. The development of such a prediction method would clearly have profound implications on the diagnosis and treatment of disease.

This paper focuses on developing a method for class prediction. Class discovery by gene expression data has been attempted through a variety of clustering techniques [4, 5, 13]; in Section 5 we describe how our prediction technique can be used to improve the class discovery process as well. To develop these methods, we consider the simplified problem of predicting membership in one of just two classes. In Section 6, we discuss ways of extending the methods to construct multi-class predictors.

In our experiments we measure the expression levels of approximately 6800 genes. Most of these genes, however, are probably not relevant to the class distinction we want to predict. Thus, a class predictor needs a method, whether explicit or implicit, for focusing on the relevant genes.

For our predictor, we select genes explicitly using the methods described in Section 2. In Section 3 we describe how to use the chosen genes to predict and how to evaluate the method's success. Our prediction method is designed to be extremely robust and to permit experimentation with different schemes for gene selection, prediction, and confidence evaluation. Despite this strong empirical orientation the method can be viewed in a theoretical Bayesian framework, as discussed in Section 3.3.

Sections 4 and 5 describe experiments in which we applied the method to the classification of acute leukemia patients; Section 4 focuses on class prediction while Section 5 addresses the class discovery problem. Section 6 proposes several directions for future work in this area.

We need a few basic definitions before proceeding. A data set consists of a set of gene expression measurements for m genes in each of n samples (generally, one from each patient). Each gene in the data set can be represented by a *gene expression vector* $g \in \mathcal{R}^n$ showing the gene's expression in each of the n samples. Note that these measurements are actually estimates of the gene's expression level; even with the latest technology the process of measuring gene expression is somewhat noisy [9]. However, measurements are thought to be reproducible within roughly a factor of two. In practice, we restrict our expression values to be above some minimum positive threshold, so we never have negative gene expression, which would be difficult to interpret. A class vector $c \in \{-1, 1\}^n$ represents the two-class distinction we wish to predict; $c_i = 1$ if sample i is in class 1, and -1 otherwise.

2 A method for choosing significantly correlated genes

In this section we address the problem of choosing predictive genes. We consider desirable characteristics of predictive genes and we define a quantitative metric for evaluating these characteristics. Not all interesting class distinctions are determined by gene expression levels alone. Cellular differences may be regulated by alternate splice variants or by methylation, neither of which would necessarily be evident from expression chip data. We must therefore ask whether there are any genes at all whose expression data is likely to be predictive of the specific class distinction. If there are such genes, we still face several issues in choosing the right set of genes to use as predictors. This section explores each of these aspects of gene selection.

2.1 A metric for gene selection

If the exact class distribution functions were known, the problem of metric selection and class prediction would be straightforward from a Bayesian perspective [6, 10, 1]. Unfortunately, this is not the case. We therefore take a more empirical approach.

In choosing predictive genes, we look for two characteristics. First, a predictive gene's typical expression in one class should be quite different from its typical expression in the other. Second, aside from the differences in expression that are explained by the class distinction, there should be as little variation as possible. So we want a gene selection metric that favors genes where the range of the expression vector is large, but where most of that variation is due to the class distinction.

We designed a metric P with this property. Let c be a class vector and let g be the expression vector of a gene over n samples. Define the within-class mean μ_1 to be the mean expression level of samples in class 1, and the within-class standard deviation σ_1 as the standard deviation of expression in these samples. Define μ_2 and σ_2 similarly for class 2. Then we can can define a correlation metric $P(g, c) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$, which measures relative class separation.

Many other gene selection metrics could be used as well; we considered several. Most measure either the degree of similarity between gene expression and the class vector (*correlation metrics*),

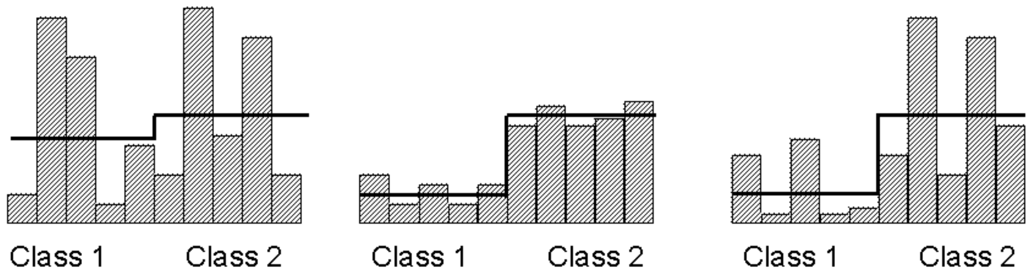

Figure 1: Expression profiles for 3 genes, each in 10 samples (samples 1–5 in class 1, 6–10 in class 2). Dark horizontal lines indicate within-class mean expression levels. The gene profiled on the left is unlikely to predict well because the class means are quite close; the expression of this gene gives us little power to distinguish between classes. The class means for the center gene profile are identical to those for the rightmost profile; both are well-separated. Of the two, the central one shows less variation around those means and so is likely to be a better predictive gene. The relative class separation metric, P, is designed to capture these properties.

or the difference between the two (*distance metrics*). These included the Pearson correlation coefficient ($1/n \sum_i^n g_i c_i$) between the class vector and the normalized gene expression vectors. We also considered a simple Euclidean distance ($1/n \sum_i^n (g_i - c_i)^2$), and some other distance-based metrics such as the Manhattan and Battacharyya distances which have traditionally been employed as measures of class separation [6, 10]. The best performance (i.e., the most accurate prediction) was obtained with the relative class separation metric defined above. This is probably a consequence of the fact that it accounts for both the class separation and the spread around class means. The Euclidean distance achieved almost the same performance but the other metrics were somewhat less accurate.

2.2 Neighborhood analysis

Having chosen a correlation metric, we consider whether there are *any* genes likely to be good predictors of the given class distinction. To answer this question, we use a permutation test we refer to as *neighborhood analysis*. Consider the class vector c and all the gene expression vectors as points in n-dimensional space. The idea of neighborhood analysis is simply to look at a neighborhood of a fixed size around c and count the number of gene expression vectors within it. We compare this to the number of expression vectors within the neighborhood of the same size around a random permutation of c. By trying many random permutations of c, we can determine if the neighborhood around c holds more gene expression vectors than we'd expect to see by chance. If so, we conclude that the class distinction represented by c is likely to be predictable from the expression data.

For example, Figure 2a shows the neighborhoods around a hypothetical class distinction c and a random class distinction c'. Within the neighborhood of size k around c there are many more genes than appear in the same-sized neighborhood around c'. This distinction holds whether k is a measure of distance or of correlation; the only thing that changes is the direction of the inequality.

Figure 2b shows a neighborhood analysis plot distinguishing six normal kidney samples from six renal cell carcinomas, cancerous tumors derived from the same tissue. The y-axis shows the number of genes within the neighborhood around c (the vector representing the class distinction between normal and cancerous samples) and the x-axis indicates the size of the neighborhood (i.e.,

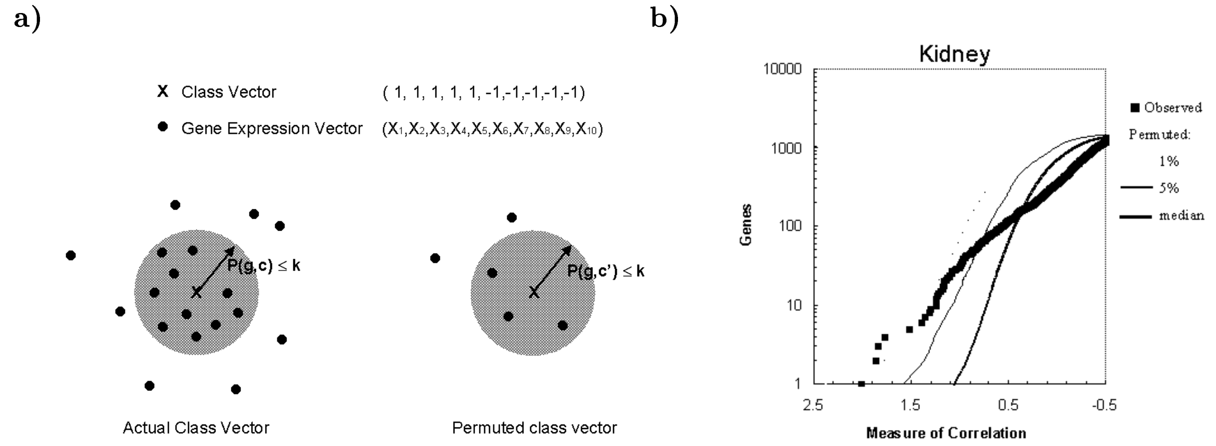

Figure 2: Neighborhood analysis. a) A schematic diagram of neighborhoods around real and randomly permuted class vectors. b) Plot of observed $P(g, c)$ distinguishing 6 kidney samples from 6 renal cell carcinomas, compared to randomly permuted class distinctions.

$P(g, c)$). So an observed data point at (1.3, 10) indicates that the gene g with the 10-th highest correlation with c has $P(g, c) = 1.3$, or that there are 10 genes within the neighborhood defined by $P(g, c) \geq 1.3$. The observed data intersects the 5% significance line at 72 genes, with $P(g, c) = 0.85$, indicating that for 5% of the random vectors c', neighborhoods of size 0.85 contain as many genes as we saw in the neighborhood of size 0.85 around the real class distinction c. We interpret the existence of genes above the 5% significance level as an indication that the class distinction is likely to be predictable by gene expression data. The 5% significance level is sufficient since we are examining only 12 samples; if the distinction is truly predictable, as this plot implies, we should see the significance level increase as the data set grows.

2.3 Choosing a prediction set S

Once the neighborhood analysis graph shows that there are genes significantly correlated with class distinction c, there are some decisions to make in choosing a set of genes for prediction. Our goal is to choose a set S of k genes such that most genes in S are likely to be predictive of the class distinction in future samples. We could simply choose the top k genes by the absolute value of $P(g, c)$, but this allows for the possibility that, for example, all genes might be expressed at a high level in class 1 and a low level (or not at all) in class 2. We've found that predictors often do better when they include some genes expressed at high levels in each class. So we perform separate neighborhood analyses for positive and negative $P(g, c)$ scores, and we choose the top k_1 genes (highly expressed in class 1) and the bottom k_2 genes (highly expressed in class 2).

Finally, we need to determine how sensitive our prediction method is to the exact number of genes used and to choose k_1 and k_2 accordingly. There are several competing constraints. We want to limit the number of genes to those shown to be significantly correlated by the permutation test, perhaps at the 1% level. Furthermore, suppose that there were a single gene whose expression level was perfectly correlated with the class distinction in the available data. We still might like a predictor that includes more than one gene, in order to provide robustness against noise and to allow us to estimate prediction accuracy, as described in Section 3.1. Additional genes may be active in different biological pathways or may provide independent estimates of activity along the same pathway; in either case they can add information that can be combined to improve prediction.

On the other hand, this argument cannot be extended indefinitely. If there are a thousand genes that are significantly correlated with a class distinction, it's unlikely that they all represent different biological mechanisms. Their expression patterns are probably dependent, so that the thousandth gene would be unlikely to add information not already provided by the previous 999. However, that thousandth gene *does* add noise to the system.

In general, then, there is a tradeoff between the amount of additional information and robustness gained by adding more genes, and the amount of noise added. The optimal size of the prediction set is likely to vary somewhat due to the genetics of the class being predicted (whether there are many independent classes of genes correlated with the class distinction, or just a single co-regulated pathway; whether many genes or few).

We therefore evaluated two different methods for choosing k_1 and k_2. The first chooses all genes above the 1% significance level in neighborhood analysis, but sets a maximum of 50 genes in each direction to avoid being dominated by noise. The second method tries many different values for $|S|$, with constraint that k_1 and k_2 are roughly equal. These models are each evaluated in cross-validation on roughly half the data. The performance of the best model is then tested on the remainder of the data.

In Section 4 we use the second approach, but preliminary results comparing the two methods indicate that they provide roughly equivalent predictive ability. Similar results for predictors using different numbers of genes (as discussed in Section 4) indicate that this prediction method is not highly sensitive to the exact number of genes used. This allows us to simply choose a reasonable-sized prediction set according to the guidelines mentioned here.

3 Prediction by weighted voting

Once we've chosen S, we're ready to try predicting new samples. We assume that we have a set of samples called the *training set* whose correct classifications are already known, and a *test set* of additional samples whose classes are currently unknown, at least to the algorithm.

To determine the classification of a new sample in the test set, we use a simple weighted voting scheme. Each gene in S gets to cast its vote for exactly one class. The gene's vote on a new sample x is weighted by how closely its expression in the training set correlates with c. The vote is the product of this weight and a measure of how informative the gene appears to be for predicting the new sample.

Intuitively, we'd expect the gene's expression in x to look like that of either a typical class-1 sample or a typical class-2 sample in the training set. So we compare expression in the new sample to the class means in the training set. We define a "decision boundary" b halfway between the two class means. The vote corresponds to the distance between the decision boundary and the gene's expression in the new sample. So each gene casts a weighted vote $V = \text{weight}(g) \cdot \text{distance}(x, b)$. (We use $P(g, c)$ as the weight of gene g; formal definitions of voting and normalization as implemented in [7] are described in Appendix A.)

The weights are defined so that positive votes count as votes for membership in class 1, negative ones for membership in class 2. The votes for all genes in S are combined; V_1 is the sum of all positive votes and V_2 the sum of all negative votes. The winner, and the direction of the prediction, is simply the class receiving the larger total vote.

3.1 The tradeoff of reliability vs. utility

Intuitively, if one class receives most of the votes and the other class has only a token representation, it seems reasonable to predict with the majority. However, if the margin of victory is slight, a prediction for the majority class seems somewhat arbitrary and can only be done with low confidence. We therefore define the "prediction strength" (PS) to measure the margin of victory:

$$PS = \frac{V_{\text{winner}} - V_{\text{loser}}}{V_{\text{winner}} + V_{\text{loser}}}.$$

Since V_{winner} is always greater than V_{loser}, PS varies between 0 and 1.

As one might expect, typical prediction strengths for incorrect predictions tend to be much lower than those for correct predictions. Thus, the PS measure provides a quantitative way of defining a tradeoff between a "reliable" predictor (one that is almost always correct but sometimes refuses to predict), and a "useful" predictor (one that makes a prediction in every case, but may be incorrect sometimes) [11]. If it is essential to make a prediction every time, one can always predict in favor of the winning class, however small the margin of victory. If the cost of an incorrect prediction is high, one can choose a PS threshold below which predictions are not made. For example, when the predictions are used to diagnose or direct treatment of cancer patients, as in Section 4, an incorrect prediction could have potentially devastating results. Therefore, we choose a conservatively high PS threshold (0.3) to minimize the chance of making an incorrect diagnosis (and potentially treating a patient with the wrong chemotherapy regimen).

3.2 Evaluating the method

The preceding discussion suggests two criteria for evaluating a predictor: the error rate, or percentage of incorrect predictions from the total number of predictions made; and the "no-call" rate, or the percentage of samples for which no prediction was made (due to a PS below the threshold).

When the number of samples is limited, we evaluate the model by n-way cross-validation: remove a single sample from the data set, use the remaining $n-1$ samples as the training set, and test the algorithm's ability to predict the withheld sample. This process is repeated for each of the n samples in turn, and the error and no-call rates are calculated over the entire data set.

When additional samples become available, we use them to test the best model found in the cross-validation step. If the initial data set is large enough to divide in two, we may still benefit from a cross-validation step in a number of ways. We can try models with different numbers of genes in cross-validation, and pick the best one as the final model to apply to future data. We may also evaluate the tradeoff between error and no-call rate at the cross-validation stage. By plotting the cumulative cross-validation error rate (with a PS threshold of zero) against the prediction strength, one can determine a reasonable choice for the PS cutoff to use in the test set.

3.3 The prediction scheme from a Bayesian perspective

Our approach of separating gene selection, prediction, and confidence evaluation can be cast in a Bayesian framework [6, 10, 1]. In this formalism, the prediction of new samples is based on the log likelihood ratio. For example, if we assume that the class distributions p_1 and p_2 are normal with equal variances –and that the mean and variance of the classes can be effectively estimated using the training set– then the vote for gene g is the log likelihood ratio: $V_g = (x - \frac{\mu_1+\mu_2}{2})(\frac{\mu_1-\mu_2}{\sigma^2})$. (The derivation of this expression is outlined in Appendix B.) In this expression one can identify the two factors (distance and weight) used in our voting scheme. A Bayesian rule that minimizes the error assigns the predicted class according to the sign of V_g. Prediction is based on the sign of the sum of V_g (assuming independence) over all genes (i.e., the prediction strength is $\sum_g V_g / \sum_g |V_g|$).

A number of potential improvements of the method can be identified from this perspective: fitting non-normal empirically-determined distributions to the data (e.g., the LaPlace distribution);

removing the symmetry assumption so that the class boundary is shifted as a function of the class distributions; removing the independence assumption in the voting scheme by using linearly-independent gene components rather than the genes themselves; etc. These enhancements have the potential to increase the accuracy of the prediction scheme at the cost of sacrificing robustness and simplicity. More work will be needed to evaluate their effects and practical potential.

4 Application: Classifying patient samples

We applied this approach to the problem of classifying acute leukemias [7]. Acute leukemias can broadly be divided into two classes, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), that respectively originate from cells of either myeloid or lymphoid origin [12]. The two diseases appear identical under the microscope, and indeed were thought to be a single disease for many years. However, correct diagnosis is critical, since they respond best to different treatment regimens. Diagnosis currently requires a number of distinct clinical tests, each performed by specialized labs and analyzed by experts. While most diagnoses are correct, mistakes still occasionally occur. Because the acute leukemias are well understood and can generally be predicted correctly, they are a good test case for class prediction methods. Furthermore, a few genes whose expression serves as a marker of acute leukemia type are already known, indicating that the class distinction is likely to be predictable from gene expression data.

We obtained a set of 38 leukemia samples (11 AML, 27 ALL). Samples were derived from bone marrow (BM) taken at time of diagnosis (i.e., before treatment). We used these samples as our training set. For further testing, we later obtained an additional 34 samples (14 AML, 20 ALL) as a test set. While 25 of the test samples were derived from bone marrow, 9 came from peripheral blood (PB) samples, which are thought to be considerably more heterogeneous than the bone marrow samples. It thus remained to be seen whether they could be predicted accurately by a model trained on cleaner data.

We first performed neighborhood analysis in the training set and found many genes significantly correlated with the ALL/AML distinction. Figure 3 shows that there are about 700 genes above the 1% level in each direction.

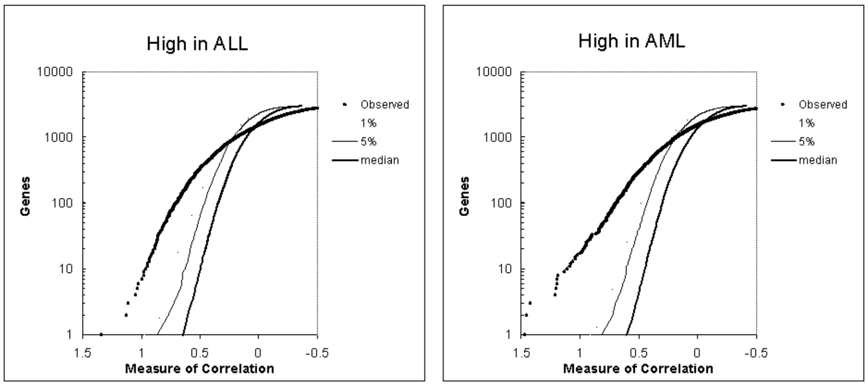

Figure 3: Neighborhood analysis for AML and ALL samples.

We then performed cross-validation, with a PS cutoff of 0.3. All trials with at least 3 genes had an error rate of zero with from 1 to 4 no-calls. Since there were several hundred genes correlated with the ALL/AML distinction at the 1% level, we somewhat arbitrarily chose to use a 50-gene predictor. This model made 36 predictions, all correct, out of 38 samples; one of the two no-calls would have been an error if predicted.

We went on to evaluate the performance of the method on a test set of 34 additional leukemia samples. The 50-gene model (trained on all 38 of the samples in the training set) predicted 29 of the 34 samples, all of them correctly. Of the remaining 5 (4 BM, 1 PB), only two (including the PB sample) would have been predicted incorrectly had the PS threshold been set at zero. A complete breakdown of the samples, their origins (PB or BM), their predictions and prediction strengths can be found on our web site (www.genome.wi.mit.edu/MPR).

Next, we built a predictor to distinguish between the two key subclasses of ALL, those arising from T-cells and those arising from B-cells. While the distinctions between AML and ALL are fairly dramatic, those between T-ALL and B-ALL are more subtle, leading us to expect that prediction might be more difficult. However, neighborhood analysis (Figure 4) showed about 200 genes significantly correlated with the distinction. In cross-validation, a single predictor built with 50 genes made 32 calls (PS $\geq .3$) out of 33 samples; all 32 were correct. A predictor built with all 200 significant genes gave essentially the same results: 32 samples were correct with PS $\geq .3$, and the same sample fell below the PS threshold as in the 50-gene model. This illustrates one of the strengths of the weighted voting scheme – its relatively low dependence on the exact number of genes chosen. In general, the average prediction strength drops as the number of genes increases above a certain point, but the drop is gradual enough to allow some flexibility in gene selection.

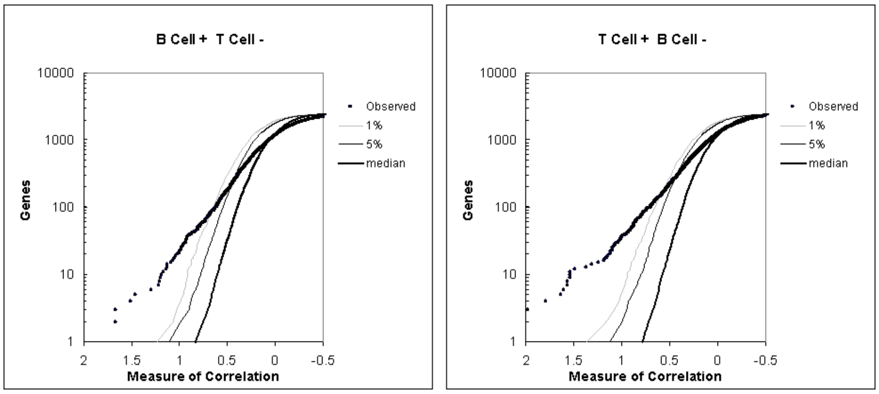

Figure 4: Neighborhood analysis for the B-cell / T-cell distinction in ALL samples.

Finally, in a set of 15 AML samples for which we had long-term follow-up information, we attempted to predict which patients would go into remission following chemotherapy. The 15 samples were divided into "treatment failure" and "treatment success" groups (8 and 7 samples, respectively). Neighborhood analysis, however, showed no genes correlated with the distinction at the 1% or even the 5% significance level. Only the top gene out of 6800 looked more highly correlated than we'd expect to see by chance, but at roughly the 10% significance level. Despite this, we attempted prediction and found that the only good predictor (making two errors in 15 samples) contained only this single gene, HOXA9. Interestingly, HOXA9 was already known to be involved in AML pathogenesis [2]; overexpression of HOXA9 causes leukemia in transgenic mice [8]. Thus, we suspect that this gene may truly be predictive of outcome despite the lack of statistical significance in our permutation test; more samples are needed to test this hypothesis. Error rates for predictors with additional genes were generally above 30%. We also conclude, therefore, that prediction with genes that show *no* correlation in excess of that expected by chance is unlikely to succeed. This supports our proposal for using neighborhood analysis to determine whether a class distinction is predictable.

5 Application: Verifying proposed classes

Class discovery (as opposed to class prediction) for the acute leukemias required many years of medical research, and cancer classification is still an active area of research today. It has been suggested that computational analysis of gene expression may be a useful approach to expediting the discovery of new, clinically-significant classes. To this end, a variety of approaches to sorting and clustering gene expression data have been proposed [4, 5, 13]. However, regardless of the method used for class discovery (whether self-organizing maps (SOMs) or cluster trees or conclusions drawn from years of clinical observation), the challenge we face is in validating the clusters. Any clustering algorithm will find clusters of samples in expression data. However, given relatively few samples and thousands of gene expression vectors, one needs to show that the class distinctions discovered are real and biologically interesting, rather than coincidental artifacts of the data. We propose improving and validating the clusters by testing predictability.

We expect that if clusters reflect true biological structure, the distinction should be predictable in additional samples. We therefore clustered the 38 leukemia samples in our training set using the self-organizing map method implemented in our GENECLUSTER software [13]. When we asked for two clusters of samples, the dominant distinction was nearly that of ALL/AML: cluster A contained 1 AML and 24 ALL samples, while cluster B contained 10 AML and 3 ALL samples. Thus, all but four of the samples were consistent with the known ALL/AML distinction.

Were the four remaining samples anomalies or were they actual improvements over the ALL/AML distinction? To investigate this question, we tested the predictability of the SOM-derived classes (A and B rather than ALL and AML). Testing prediction of derived classes is different from predicting known classes in that we do not know the correct answers for new data. However, we can test consistency by examining prediction strengths in cross-validation. When we performed cross-validation on the SOM-derived classes, two of the four anomolous samples were not called (PS $< .3$), and a third, the lone AML sample in class A, was incorrectly predicted to come from class B (the class containing the majority of the AML samples). These errors account for three of the four samples not predicted correctly in cross-validation, showing that the majority of the class distinction, that part consistent with ALL/AML, is still predictable. In this way, cross-validation can be used to refine classes discovered by other means.

Furthermore, the distribution of prediction strengths for distinguishing classes A and B by cross-validation was significantly higher than we'd expect for a random class distinction. We predicted 100 random permutations of the class distinction using the same data and generated a histogram showing the median PS (over the 38 samples) for predicting each of the random distinctions (Figure 5). In contrast, the median PS for the SOM-derived class distinction was 0.86, noticeably higher than the highest (0.66) of the 100 random distinctions tried.

Next, we asked GENECLUSTER to divide the set of leukemia samples into four clusters. This time the clustering method not only discovered the ALL/AML distinction, but also divided the ALL samples largely by T-cell or B-cell lineage. There were two clusters of B-cell ALLs, one of T-cell ALLs, and one of AMLs. The key question was whether the distinction between the two B-cell classes was a new but meaningful biological difference, or simply an artifact resulting from our asking the algorithm to produce more clusters than the data would support. We tested this by building predictors for all six possible pairwise class comparisons between the four clusters. The median prediction strength for distinguishing the two B-cell clusters (0.40) was much lower than for any of the other distinctions (0.65-0.94), suggesting that the distinction is more likely to be an artifact of the clustering method than a true new discovery.

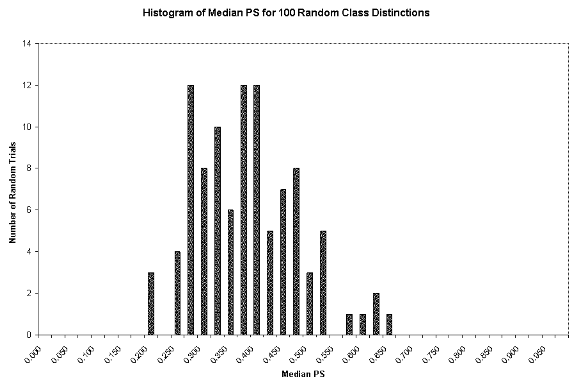

Figure 5: Histogram of median prediction strengths (over 38 samples) in each of 100 random class distinctions. By comparison, the median prediction strength for the class distinction derived by our clustering algorithm was 0.86.

6 Discussion and Conclusion

We have described a method for class prediction from gene expression data and illustrated its potential by pilot experiments in of cancer classification. However, it is worth stressing that these results are not limited to diagnosis, nor even to the field of cancer research. We have already mentioned the possibility of using these methods to predict patient outcomes or responses to treatment. Notably, the genes selected for our method often appear to be directly relevant to the process being studied, and thus may provide clues to gene function or leads for drug development. In general, one can imagine using similar methods to predict *any* trait or characteristic that is evident at the transcriptional level. Furthermore, neighborhood analysis allows us to determine which distinctions are evident transcriptionally. The number of potential applications is unlimited.

However, a great deal remains to be done. We have been fortunate in that in the majority of our test cases, classes are distinguished by a large number of genes whose expression indicates class membership for all the samples we tested. However, one could imagine predicting more subtle distinctions where no one biological pathway is responsible for all the cases in either class.

Furthermore, to be universally applicable a prediction method must be able to distinguish between multiple classes. Certainly, there are *ad-hoc* methods for combining the binary class predictors described here into multi-class predictors. As we did in Section 5, one can do pairwise comparisons between a set of classes; however, this becomes time-consuming as the number of classes grows. Another approach is to predict each sample's membership in either a class or its complement and repeat the process for each of the known classes. Ideally, a sample will show strong evidence (indicated by a high PS) for membership in one class only. In practice, this is moderately effective when the sample sizes for each class are sufficiently large, but the method still scales linearly with the number of classes. The desirable solution, an elegant extension of the weighted voting scheme for distinguishing multiple classes, is non-trivial since individual genes that distinguish between all classes simultaneously are unlikely to exist.

Thus, a useful extension of the method would allow different sets of genes to be responsible for predicting various subsets of the target classes. Califano, *et al.* have suggested the application of a Bayesian approach to find all maximal patterns of correlated genes [3]. We have also considered using canonical discriminant analysis to provide an approach intermediate in complexity and, potentially, in predictive power. We plan to compare the strength of these methods in a future paper.

Acknowledgments

We thank Christine Huard and Michelle Gaasenbeek for technical assistance in generating the leukemia data and analyzing chip quality; Mignon Loh, James Downing, Clara Bloomfield, and Mark Caligiuri for providing patient samples; Andrea Califano, Malcolm Williamson and Frank Shen for insights into comparative statistical methods; and Nathan Seimers, Michael Angelo, Jane Staunton, Hilary Coller, Jean-Paul Comet, and the rest of the Whitehead's Molecular Pattern Recognition Group for helpful discussions.

References

[1] James O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer series in Statistics, 2nd edition, 1985.

[2] J. Borrow, *et al.* The t(7;11)(p15;p15) translocation in acute myeloid leukaemia fuses the genes for nucleoporin NUP98 and class I homeoprotein HOXA9. *Nature Genetics*, 12:159–167, 1996.

[3] Andrea Califano, Gustav Stolovitzky, and Yuhai Tu. Analysis of gene expression microarrays: A combinatorial approach. Unpublished Manuscript, 1999.

[4] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

[5] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, USA*, 95:14863–14868, 1998.

[6] K. Fukunaga. *Statistical Pattern Recognition.* Academic Press, 2nd edition, 1990.

[7] Todd Golub, Donna Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill Mesirov, Hilary Coller, Mignon Loh, James Downing, Mark Caligiuri, Clara Bloomfield, and Eric S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, page to appear, October 1999.

[8] E. Kroon, J. Krosl, U. Thorsteinsdottir, S. Baban, A.M. Buchberg, and g. Sauvageau. Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. *EMBO Journal*, 17:3714–3725, 1998.

[9] David Lockhart, *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[10] B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1st edition, 1996.

[11] R. L. Rivest and R. Sloan. Learning complicated concepts reliably and usefully. In *Proc. 1st Annual Workshop on Computational Learning Theory*, pages 69–79, San Mateo, CA, 1988. Morgan Kaufmann.

[12] S. Robbins, R. Cotran, V. Kumar, and F. Collins. *Pathologic Basis of Disease.* Saunders, 6th edition, 1999.

[13] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric Lander, and Todd Golub. Interpreting gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences, USA*, 96:2907–2912, 1999.

[14] L. Wodicka, H. Dong, M. Mittmann, M.H. Ho, and D. Lockhart. Genome-wide expression monitoring in saccharomyces cerevisiae. *Nature Biotechnology*, 15:1359–1367, 1997.

Appendix A: Weighted voting

This appendix provides details of the weighted voting scheme used for prediction. Recall from Section 3 that each gene casts a vote $V =\text{weight}(g)\ \text{distance}(x, b)$.

Formally, consider a single gene represented by gene expression vector g, and let x be the raw expression level of that gene in a new sample whose class we want to predict. Let $\tilde{x} = \log_{10} x$, and let $\tilde{g} = (\log_{10}(g_1), \ldots, \log_{10}(g_n))$. Let μ and σ represent the mean and standard deviation of $\tilde{g}$. Then we define the normalized vector $\tilde{g}_{\text{norm}}$ by

$$\tilde{g}_{\text{norm}} = \left(\frac{\tilde{g}_1 - \mu}{\sigma}, \ldots, \frac{\tilde{g}_n - \mu}{\sigma}\right)$$

$$\text{and } \tilde{x}_{\text{norm}} = \frac{\tilde{x} - \mu}{\sigma}.$$

(Note that $\tilde{x}_{\text{norm}}$ is normalized by the mean and standard deviation over the training set only.) We define the class means for $\tilde{g}_{\text{norm}}$:

$$\mu_1 = \frac{\sum_{(i \in \text{Class 1})} \tilde{g}_{\text{norm}_i}}{|\text{Class 1}|}$$

and μ_2 similarly. Finally, we set the "decision boundary" $b = (\mu_1 + \mu_2)/2$ halfway between the average class 1 expression level and the average class 2 expression level. Then the gene's vote V is simply the gene's weight in the training set (in this case, its correlation with c), multiplied by the distance of the new sample from the decision boundary:

$$V = P(g, c)(\tilde{x}_{\text{norm}} - b).$$

Positive votes are counted as votes for the new sample's membership in class 1, negative votes for class 2. To see why this works, recall that $P(g, c) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$. So $P(g, c)$ is positive if $\mu_1 > \mu_2$ and negative if $\mu_1 < \mu_2$. Suppose that $\mu_1 > \mu_2$. If x looks like a typical sample in class 1, $\tilde{x}_{\text{norm}} > b$, so $\tilde{x}_{\text{norm}} - b > 0$ and the weighted vote will be positive. However, if x looks like a typical class 2 sample, $\tilde{x}_{\text{norm}} - b < 0$ and the weighted vote will be negative. A similar argument holds if $\mu_1 < \mu_2$: the signs of $(\tilde{x}_{\text{norm}} - b)$ are reversed but cancel with the negative sign of $P(g, c)$, so the weighted votes are still positive for class 1 and negative for class 2.

Appendix B: Derivation of Bayesian log likelihood

Recall that in Section 3.3 we've assumed that the two class distributions of gene expression, p_1 and p_2, are normal distributions with equal standard deviations represented by σ. Then

$$V_g = -\ln \frac{p_1}{p_2} = \frac{1}{2}\frac{(x - \mu_1)^2}{\sigma^2} - \frac{1}{2}\frac{(x - \mu_2)^2}{\sigma^2}$$

$$V_g = \frac{1}{\sigma^2}\left(2x(\mu_2 - \mu_1) + \mu_1^2 - \mu_2^2\right)$$

$$V_g = \left(x - \frac{\mu_1 + \mu_2}{2}\right)\left(\frac{\mu_2 - \mu_1}{\sigma^2}\right)$$

In our framework, the first term can be interpreted as the distance between the new sample and the decision boundary, while the second term can be viewed as the gene's weight.