# Machine Learning Methods Applied to DNA Microarray Data Can Improve the Diagnosis of Cancer

Eric Bair
Dept. of Statistics
Stanford University
Stanford, CA 94305-4065
ebair@stanford.edu

Robert Tibshirani
Depts. of Health, Research, & Policy, and Statistics
Stanford University
Stanford, CA 94305-4065
tibs@stat.stanford.edu

## ABSTRACT

The morbidity rate of cancer victims varies greatly for similar patients who receive similar treatments. It is hypothesized that this variation can be explained by the genetic heterogeneity of the disease. DNA Microarrays, which can simultaneously measure the expression level of thousands of different genes, have been successfully used to identify such genetic differences. However, microarray data typically has a large number of features and relatively few observations, meaning that conventional machine learning tools can fail when applied to such data. We describe a novel procedure called "nearest shrunken centroids" that has successfully detected clinically relevant genetic differences in cancer patients. This procedure has the potential to become a powerful tool for diagnosing and treating cancer.

## Keywords

Microarrays, shrunken centroids, classification

## 1. OVERVIEW

When a patient is diagnosed with cancer, various clinical parameters are used to assess the risk of metastasis and death in that patient. However, despite numerous advances in the field, our ability to determine the risk of morbidity is extremely limited. Tumors that appear indistinguishable under the microscope may have drastically different effects on the patient.

It has long been known that cancer is a genetic disease. Thus, it is commonly believed that these differences in the clinical outcome of cancer can be explained by differences in the genetic profile of the tumor. Unfortunately, until recently, our ability to directly observe the genetic makeup of a tumor was extremely limited.

This is changing, however, with the advent of DNA microarrays. Microarrays can simultaneously measure the expression levels of thousands of genes in an organism. Thus, they have the ability to detect differences between tumors at the molecular level. This is illustrated in Figure 1. Under the microscope, the two types of lymphoma appear to be identical. However, gene expression profiling reveals that the two tumor types are actually distinct at the molecular level.

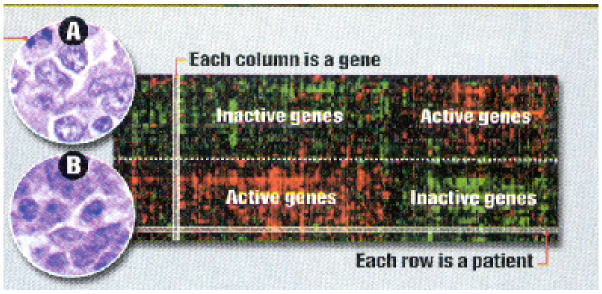

Figure 1: DNA Microarrays can identify differences between tumors that are not detectable under a microscope. Using conventional microscopic analysis, the lymphoma cells in groups A and B appear to be identical. Microarrays analysis shows that different genes are active and inactive in these two groups, indicating that they represent distinct disease subtypes.

The ability to identify such subgroups has important implications for the diagnosis and treatment of cancer. Suppose one subtype of cancer is likely to metastasize whereas another subtype is not. The patients who have a high risk of metastasis would need to be treated aggressively, whereas the other patients could be given a less invasive treatment (or no treatment at all). If there is no way to distinguish between these subtypes, all patients would need to be given the aggressive treatment. However, this is highly undesirable, because current treatments for cancer, such as surgery or chemotherapy, have extremely severe side effects. (In fact, some cancer patients have died as a result of chemotherapy.) If we could successfully identify the patients with a high risk of metastasis and death, we could give them the appropriate treatment while sparing other patients from the noxious side effects that such treatment would entail.

This is essentially a classification problem. Given a number of features (gene expression levels), we wish to predict which type of cancer is present in a patient. Many machine learning procedures have been developed for this type of problem. (See, for example, [4; 6].)

Unfortunately, these existing machine learning procedures cannot be directly applied to microarray data. The number of features is extremely large compared to the number of observations, causing most machine learning procedures to fail. Moreover, it is important to identify to identify which genes are the best predictors of tumor type. Using the expression level of thousands of different genes to make a diagnosis is not practical with existing technology (and may never be feasible). If we can identify a small subset of predictive genes, we could use these genes to raise antibodies suitable for immunostaining. Alternatively, it may be possible to develop RNA-based diagnostic tests using RT-PCR. Both of these techniques, however, can only be applied to a relatively small number of genes. Thus, it is important to identify which genes are necessary to perform such a classification.

## 2. PAM: A TOOL FOR CLASSIFYING TUMORS BASED ON MICROARRAY DATA

Recent attempts to classify tumors using microarray data have used statistical methods [3; 5; 7] and artificial neural networks [8]. We describe an alternative method that performs well on a wide variety of problems. It is also easy to understand and interpret.

PAM is based on a technique known as "nearest shrunken centroids." We illustrate the utility of this method by applying it to a microarray data set of small round blue cell tumors (SRBCT) of childhood [8]. The data set consists of measurements of the gene expression level of 2308 genes from 88 patients. Four different types of tumors were represented: Burkitt lymphoma (BL), Ewing sarcoma (EWS), neuroblastoma (NB), and rhabdomyosarcoma (RMS). The authors divided these patients into 63 training cases and 25 test cases. Using neural networks, they classified the test observations with 100% accuracy. Their model used 96 genes.

Tibshirani et al. [10] first analyzed this data set using a nearest centroid classifier. (For a description of this technique, see [6].) A nearest centroid classifier calculates the distance between a given test sample (patient) to the class centroid of each of the four classes. The test sample is classified to the class for which this distance is the smallest.

The class centroids of the SRBCT data are shown in Figure 2 (gray bars). When a nearest centroid classifier is applied to this data, it makes a total of five errors on the 20 test samples. This result shows that nearest centroid classifiers can be successfully applied to microarray data. It has several advantages compared to existing methods for classifying microarray data. In particular, it can be easily applied to problems with more than one class.

However, this technique still has several drawbacks. The technique of Khan et al. [8] makes zero test errors on the same data set. Moreover, nearest centroids requires that all 2308 genes be used for the classification. It would be desirable to develop a classifier with greater accuracy that uses fewer genes.

## 3. DESCRIPTION OF NEAREST SHRUNKEN CENTROIDS

To overcome the shortcomings of the nearest centroid classifier, Tibshirani et al. [10] proposed a modification of the nearest centroid algorithm known as "nearest shrunken centroids." The idea behind nearest shrunken centroids is the following: We calculate each class centroid as we would in a nearest centroid classifier. Then we divide each centroid by the within class-standard deviation for each gene. This gives greater weight to genes whose expression is stable among patients in the same class. Then we apply soft thresholding to the resulting normalized class centroids. If the normalized centroid is small, it is set to zero (and hence disregarded for the remainder of the calculation). By so doing, we obviously reduce the number of genes that are used in the final predictive model. One would hope that this would improve the accuracy of the model as well, since we would remove irrelevant genes.

Suppose there are $n$ patients and $p$ genes. We will let $x_{ij}$ denote the expression of the $i$th gene of the $j$th patient. Also, suppose there are $K$ classes. Let $C_k$ denote the indices of the $n_k$ samples in class $k$. Then the component of the centroid corresponding to the $i$th gene in the $k$th class is given by

$$\bar{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_k} \tag{1}$$

and the overall centroid corresponding to the $i$th gene in the $k$th class is

$$\bar{x}_k = \sum_{j=1}^{n} \frac{x_{ij}}{n} \tag{2}$$

Let

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i + s_0)} \tag{3}$$

where $s_i$ is the pooled within-class standard deviation for gene $i$:

$$s_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \tag{4}$$

and

$$m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}} \tag{5}$$

(The quantity $s_0$ in the denominator is a positive constant included to prevent the possibility that a gene with a low expression level could produce a large $d_{ik}$ by chance. It has the same value for all genes. One possibility is to let $s_0$ equal the median of the $s_i$'s.)

Equation (3) can be written as

$$\bar{x}_{ik} = \bar{x}_i + m_k(s_i + s_0)d_{ik} \tag{6}$$

Now, we apply soft-thresholding to these centroids. Let

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+ \tag{7}$$

where

$$t_+ = \begin{cases} t & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

We choose the optimal value of $\Delta$ by cross-validation. (By default, the PAM software tests 30 possible values of $\Delta$ ranging from 0 to the value of the largest centroid in the data set. The optimal $\Delta$ is chosen to be the value for which the cross-validation misclassification error rate is minimized.) Then we define the "shrunken centroids" to be

$$\bar{x}'_{ik} = \bar{x}_i + m_k(s_i + s_0)d'_{ik} \tag{9}$$

Note that if $d'_{ik} = 0$ for all $k$ for a given $i$, then all of the shrunken centroids are zero, and gene $i$ does not contribute to the final classification.

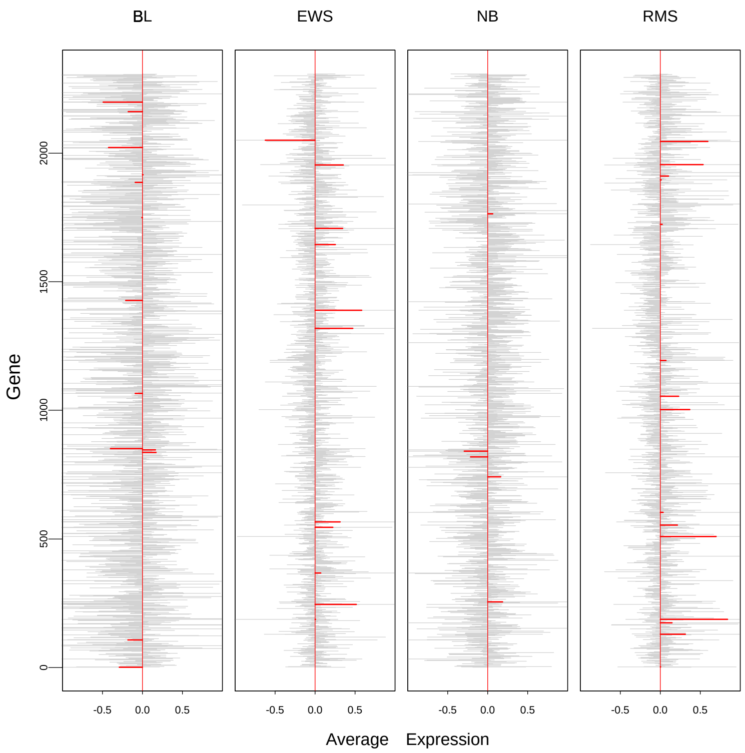

Figure 2: Centroids (gray) and shrunken centroids (red) for the SRBCT data set. The overall centroid has been subtracted from the centroid of each class. The horizontal units are log ratios of expression, and the order of the genes is arbitrary.

Now suppose that we have a "test patient" with expression levels $x^* = (x_1^*, x_2^*, \ldots, x_p^*)$. We wish to classify $x^*$ to the class whose "shrunken centroid" is nearest to $x^*$. Let

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{(s_i + s_0)^2} - 2\log \pi_k \tag{10}$$

(Here, $\pi_k$ represents the prior probability of class $k$, that is, the proportion of class $k$ in the population. If $\pi_k$ is unknown, it can be estimated from the data, or we can let $\pi_k = 1/K$ for all $k$.) Then the classification rule is $C(x^*) = \ell$ where

$$\delta_\ell(x^*) = \min_k \delta_k(x^*) \tag{11}$$

If we wish to estimate the probability that $x^*$ belongs to a given class, we may do so in the following manner:

$$\hat{p}_k(x^*) = \frac{\exp\left(-\delta_k(x^*)/2\right)}{\sum_{l=1}^{K} \exp\left(-\delta_k(x^*)/2\right)} \tag{12}$$

(This is analogous to the procedure used to estimate class probabilities in Gaussian linear discriminant analysis; see [6] for details.)

The discriminant scores in (3) are similar to those used in linear discriminant analysis (LDA). LDA uses the Mahalanobis metric to compute the distance between a given test observation and the class centroids (in vector notation):

$$\delta_k^{\mathrm{LDA}}(x^*) = (x^* - \bar{x}_k)^T W^{-1}(x^* - \bar{x}_k) - 2\log \pi_k \tag{13}$$

Here $W$ represents the pooled within-class variance/covariance matrix of the expression data. LDA has been successfully applied to a wide variety of prediction problems [6]. However, LDA cannot be directly applied to gene expression data, since the number of predictors (genes) is much greater than the number of samples (patients). As a consequence, the matrix $W$ is extremely large. Thus, any sample estimate of $W$ will be singular, and its inverse will be undefined.

Nearest shrunken centroids is thus similar to LDA, with several key differences. It assumes that the covariance matrix $W$ is diagonal. As noted earlier, it would be impossible to perform the necessary calculations without this assumption. Also, LDA uses the raw class centroids, whereas we use shrunken centroids. An important consequence of this fact is that there will be some genes for which $d'_{ik} = 0$ for all $k$. Such genes will not be used in the classification.

## 4. RESULTS ON THE SRBCT DATA

This procedure was applied to the SRBCT data of [8]; see [10] for the complete results. The value of $\Delta$ in Equation (3) was chosen by applying ten-fold cross-validation. Both the cross-validation error and test error were minimized when $\Delta = 4.34$. The error curves are shown in Figure 3. The resulting shrunken centroids are shown in Figure 2 (red bars). This model produced zero cross-validation errors and zero test errors. It required 43 genes. Thus, for this data set, nearest shrunken centroids produces accurate predictions using relatively few genes.

Figure 4 shows the 43 genes that were used to classify SRBCTs together with the value of their shrunken centroids for each of the four classes. Note that the genes with non-zero components in a given class are almost mutually exclusive.

Figure 5 shows the estimated probabilities of belonging to

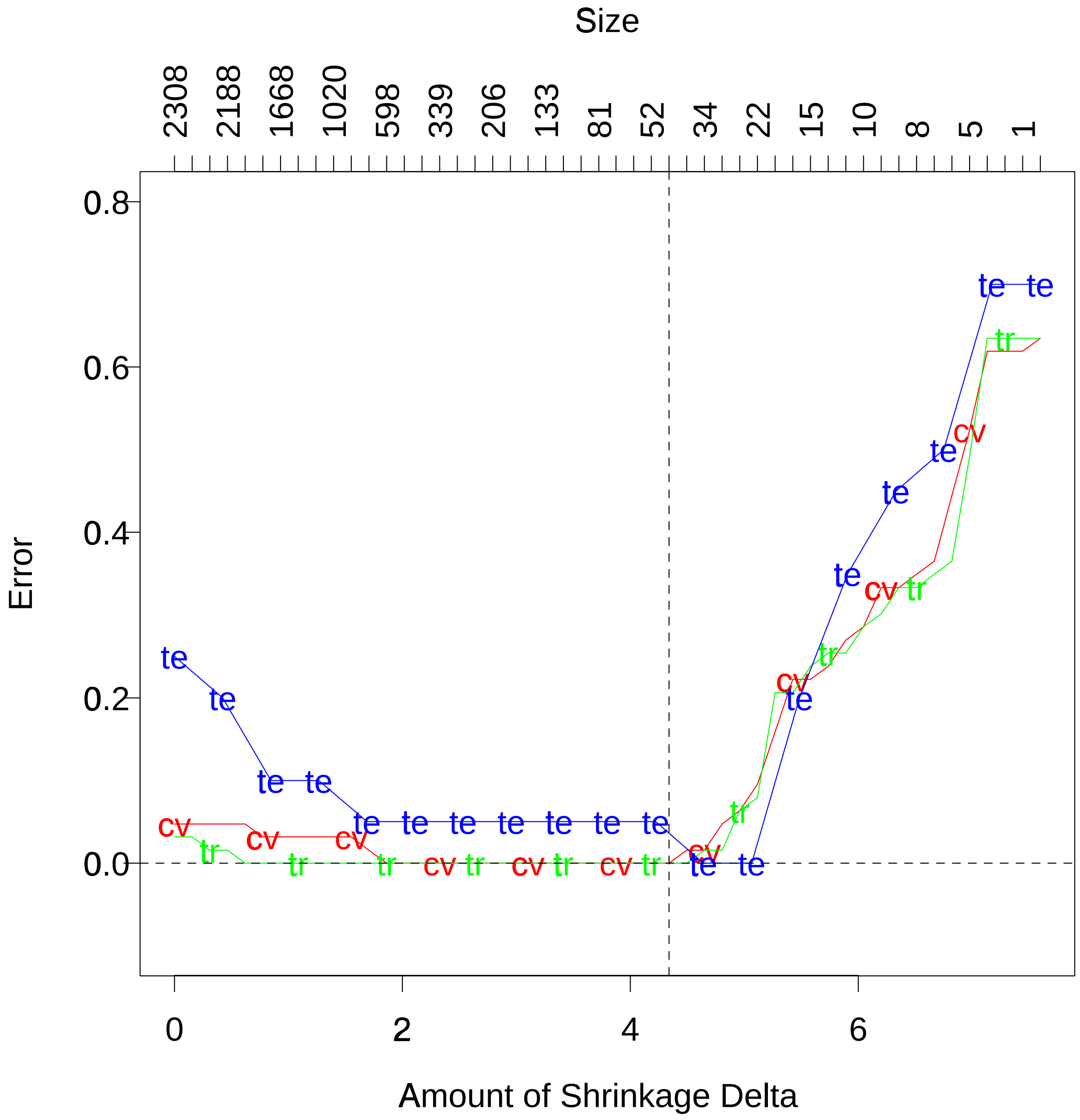

Figure 3: The error curves (training: tr/green, cross-validation: cv/red, and test: te/blue) resulting from applying nearest shrunken centroids to the SRBCT data. The value $\Delta = 4.34$ minimizes the cross-validation error rate. It produces a set of 43 genes.

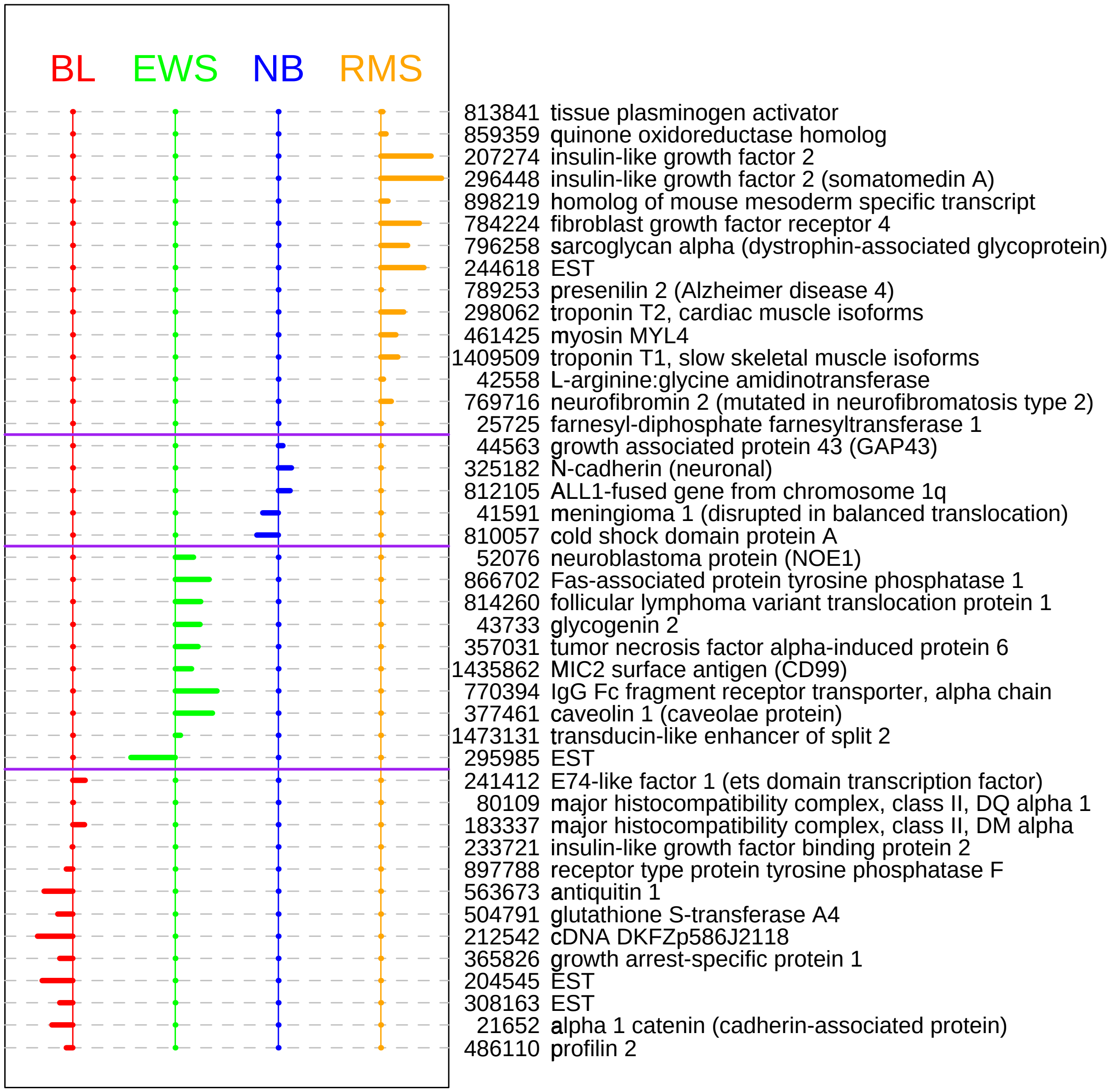

Figure 4: The values of $d'_{ik}$ for the 43 genes for which at least one $d'_{ik}$ is nonzero for the SRBCT data. Note that the genes with nonzero shrunken centroids in each class are almost mutually exclusive.

each class for each patient. For most patients, the estimated probability of belonging to the true class was significantly greater than the estimated probability of belonging to any other class. It is also interesting to examine the estimated probabilities of five test samples that were not SRBCTs. (These five samples are marked with a circle on the graph.) Note that the estimated probabilities for these five cases are significantly lower than the estimated probabilities for the true SRBCTs.

## 5. DIAGNOSIS OF CANCERS WHERE NO SUBTYPES ARE KNOWN TO EXIST

We have seen that nearest shrunken centroids has the potential to be a powerful tool for diagnosing cancer. When several cancer subtypes are known to exist, it can use gene expression information to distinguish between the subtypes using a small set of genes.

However, nearest shrunken centroids is a supervised learning procedure. It can only be applied in cases where subtypes of cancer are already known to exist. Unfortunately, no such subtypes have been identified for many types of cancer. In such cases, nearest shrunken centroids cannot be applied unless some putative subtypes can be identified.

Many types of cancer are suspected to be molecularly heterogeneous. For instance, diffuse large B-cell lymphoma (DLBCL) is the most common form of lymphoma among adults. Approximately 40% of DLBCL patients respond to chemotherapy and recover. The remainder will usually succumb to the disease. [9; 11] It is believed that this discrepancy is the result of variation among DLBCLs at the molecular level.

By itself, nearest shrunken centroids cannot be used to diagnose a patient with DLBCL. If two or more subtypes of DLBCL were known to exist and one subtype were more aggressive than the others, we could fit a nearest shrunken centroid classifier to determine which subtype is present in a given patient. However, no such subtypes have been clearly identified.

Nevertheless, if we use nearest shrunken centroids together with unsupervised learning methods, this problem becomes tractable. For example, [1] analyzed a microarray data set consisting of the expression levels of 3624 genes for 36 DLBCL patients. Using hierarchical clustering [2], they identified two putative subgroups of DLBCL, which they "GC B-like DLBCL" and "activated B-like DLBCL." They noted that patients with GC B-like DLBCL tend to live longer than patients with activated B-like DLBCL.

Although this result is intriguing, it has limited utility as a diagnostic tool. Hierarchical clustering (or any other type of clustering) can only be applied to a large group of patients. For this reason, clustering by itself cannot be used to construct a diagnostic tool. If a patient is diagnosed with DLBCL, the clinician must be able to perform a diagnosis on that individual patient. He cannot wait until a large group of other patients are diagnosed with DLBCL so that he can apply clustering to the entire group.

Nearest shrunken centroids can overcome this difficulty. After putative tumor subtypes have been identified, one can apply nearest shrunken centroids to attempt to diagnose which subtype is present in an individual patient. We would hope that the survival times of the patients differ between the predicted subtypes.

We tested this idea on the DLBCL data of [1]. There were 36 patients, of which 21 were classified as having activated B-like DLBCL and 15 were classified as having GC B-like DLBCL. We randomly divided these 36 patients into a training set of 18 patients and a test set of 18 patients. We fit a shrunken centroid classifier to the 18 training patients. The cross-validation error rate was minimized when $\Delta = 1.92$. This optimal model used 67 genes. It produced 3 cross-validation errors and 3 test errors. More importantly, however, patients in the predicted GC B-like DLBCL class lived significantly longer than patients in the predicted activated B-like DLBCL class (see Figure 6). Thus, nearest shrunken centroids can be used to help diagnose cancer even if no cancer subtypes are known to exist.

## 6. CONCLUSIONS

DNA microarrays have the potential to revolutionize the way we diagnose and treat cancer. In order to fully utilize their potential, however, we must develop tools to analyze them. Nearest shrunken centroids is a powerful tool for extracting useful information from microarray data. By identifying the genes that are necessary to differentiate between different types of cancer, it can help us to find candidates for raising antibodies for immunostaining. Moreover, our results raise the possibility of creating diagnostic tests based on RNA expression levels, perhaps using RT-PCR. Finally, we have shown that the methodology is still useful for estimating the survival of cancer patients when no subtypes have been identified. More accurate diagnoses can help clinicians to give each patient the appropriate therapy, which will increase the chances that the patient will survive, and help to spare the patient from the side effects of unnecessary treatments.

## 7. ACKNOWLEDGMENTS

Eric Bair was funded by an NSF Graduate Research Fellowship. Robert Tibshirani was partially supported by NIH grant 2 R01 CA72028, and NSF grant DMS-9971405. Several of the figures are reproduced from an earlier article by Robert Tibshirani that was published in PNAS [10].

## 8. REFERENCES

[1] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, Hajeer Sbet, Truc Tran, Xin Yu, John I. Powell, Lming Yang, Gerald E. Marti, Troy Moore, James Hudson, Jr., Lishen Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D. Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown, and Louis M. Staudt, *Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling*, Nature **403** (2000), 503–511.

[2] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein, *Cluster analysis and display of genome-wide expression patterns*, Proceedings of the National Academy of Sciences **95** (1998), 14863–14868.

[3] T.R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh,

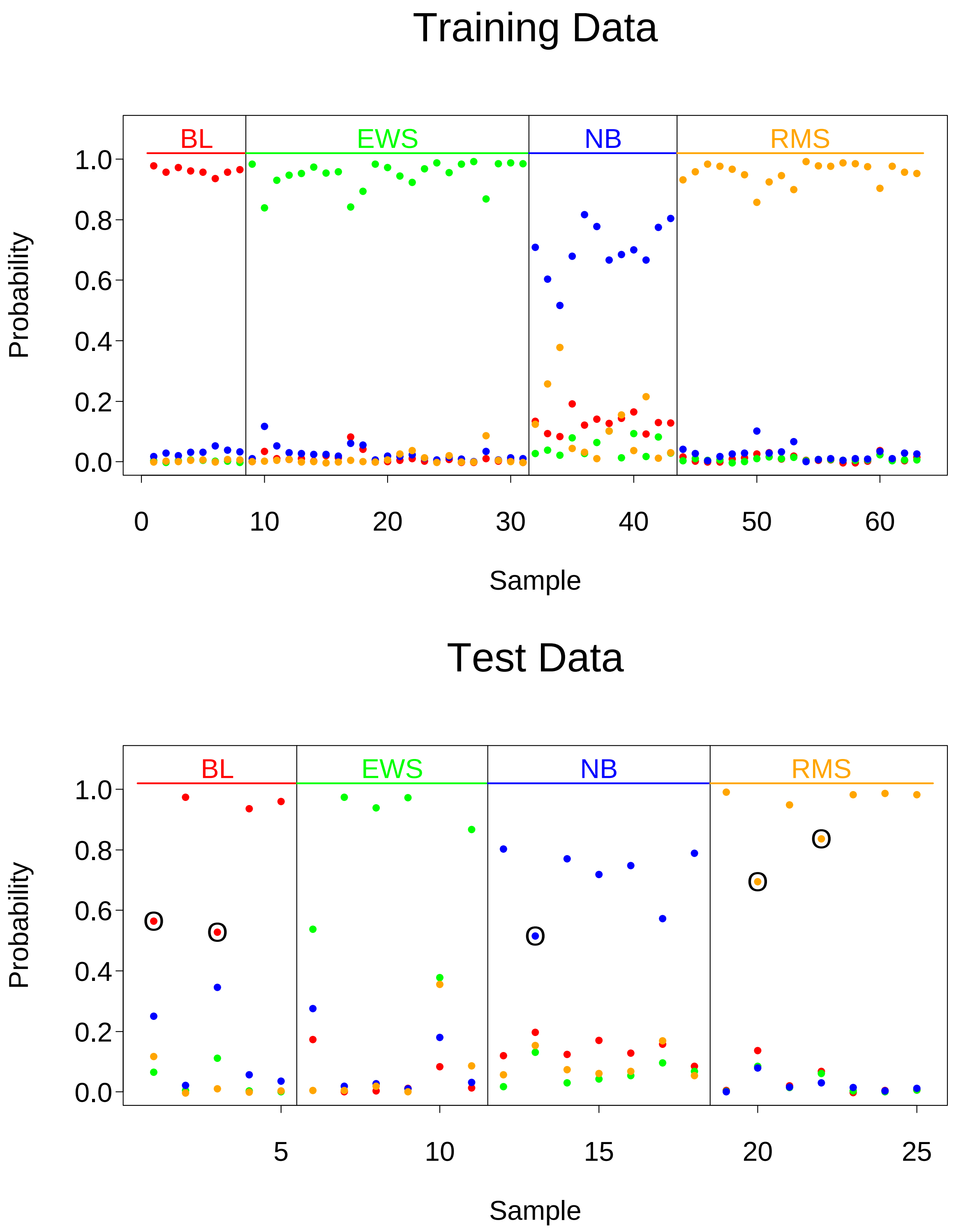

Figure 5: Estimated class probabilities for the SRBCT data. Samples are partitioned by true class (upper) and predicted class (lower). All of the 63 training samples are classified correctly, as are the 20 test samples that are known to be SRBCTs. Five of the test samples are actually not SRBCTs; they are marked with a circle. Note that their estimated class probabilities are significantly lower than the probabilities of the other test patients in each class.

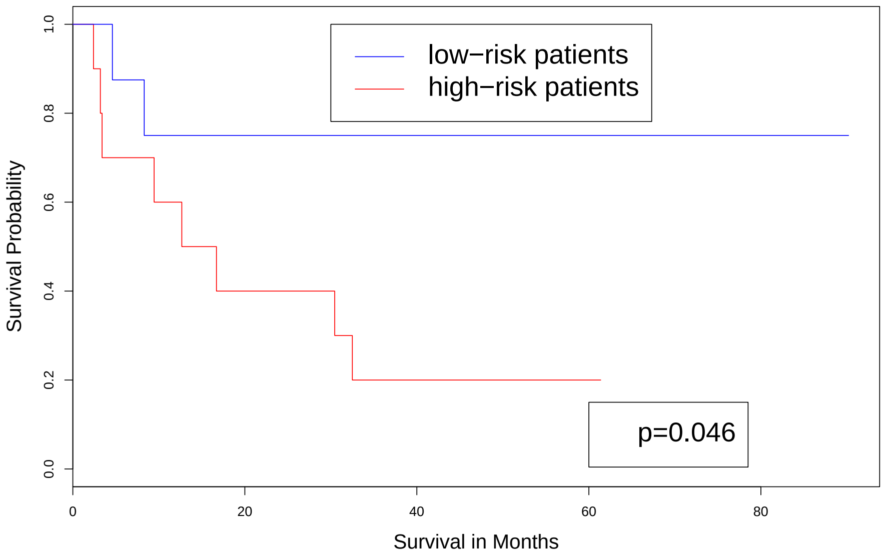

Figure 6: Survival curves of the predicted subclasses of the DLBCL data. First, hierarchical clustering was applied to an 18-patient "training set." Two distinct subclasses of DLBCL were identified. Patients in one subclass had significantly better survival than the patients in the other subclass. Once these subclasses were identified, a nearest shrunken centroid model was fit using these 18 training patients. Using this model, each patient in an 18-sample "test set" was classified as either "low-risk" or "high-risk." The "high-risk" patients had significantly poorer survival.

J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, Science **286** (1999), 531–536.

[4] A. D. Gordon, *Classification*, Chapman & Hall, Boca Raton, LA, 1999.

[5] Trevor Hastie, Robert Tibshirani, David Botstein, and Patrick Brown, *Supervised harvesting of expression trees*, Genome Biology **2(1)** (2001), 1–12.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer-Verlag, New York, NY, 2001.

[7] Ingrid Hedenfalk, David Duggan, Yidong Chen, Michael Radmacher, Michael Bittner, Richard Simon, Paul Meltzer, Barry Gusterson, Manel Esteller, Mark Raffeld, Zohar Yakhini, Amir Ben-Dor, Edward Dougherty, Juha Kononen, Lukas Bubendorf, Wilfrid Fehrle, Stefania Pittaluga, Sofia Gruvberger, Niklas Loman, Oskar Johannsson, Håkan Olsson, Benjamin Wilfond, Guido Sauter, Olli-P. Kallioniemi, Ake Borg, and Jeffrey Trent, *Gene-expression profiles in hereditary breast cancer*, The New England Journal of Medicine **344** (2001), 539–548.

[8] Javed Khan, Jun S. Wei, Markus Ringnér, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson, and Paul S. Meltzer, *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*, Nature Medicine **7** (2001), 673–679.

[9] Non-Hodgkin's Lymphoma Classification Project, *A clinical evaluation of the international lymphoma study group classification of non-hodgkin's lymphoma*, Blood **89** (1997), 3909–3918.

[10] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu, *Diagnosis of multiple cancer types by shrunken centroids of gene expression*, Proceedings of the National Academy of Sciences **99** (2002), 6567–6572.

[11] Julie M. Vose, *Current approaches to the management of non-hodgkin's lymphoma*, Seminars in Oncology **25** (1998), 483–491.