1-1-2004

Evaluation of Multiple Models to Distinguish Closely Related Forms of Disease Using DNA Microarray Data: an Application to Multiple Myeloma

Johanna S. Hardin
Pomona College

Michael Waddell
University of Wisconsin - Madison

C. David Page
University of Wisconsin

Fenghuang Zhan
University of Arkansas - Medical Sciences

Bart Barlogie
University of Arkansas

See next page for additional authors

Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 10

Evaluation of Multiple Models to Distinguish Closely Related Forms of Disease Using DNA Microarray Data: an Application to Multiple Myeloma

Johanna Hardin, *Pomona College*
Michael Waddell, *University of Wisconsin, Madison*
C. David Page, *University of Wisconsin*
Fenghuang Zhan, *University of Arkansas for Medical Sciences, Little Rock*
Bart Barlogie, *University of Arkansas*
John Shaughnessy, *University of Arkansas for Medical Sciences*
John J. Crowley, *Cancer Research And Biostatistics*

Recommended Citation:
Hardin, Johanna; Waddell, Michael; Page, C. David; Zhan, Fenghuang ; Barlogie, Bart; Shaughnessy, John; and Crowley, John J. (2004) "Evaluation of Multiple Models to Distinguish Closely Related Forms of Disease Using DNA Microarray Data: an Application to Multiple Myeloma," *Statistical Applications in Genetics and Molecular Biology*: Vol. 3: Iss. 1, Article 10.
DOI: 10.2202/1544-6115.1018

Evaluation of Multiple Models to Distinguish Closely Related Forms of Disease Using DNA Microarray Data: an Application to Multiple Myeloma

Johanna Hardin, Michael Waddell, C. David Page, Fenghuang Zhan, Bart Barlogie, John Shaughnessy, and John J. Crowley

Abstract

Motivation: Standard laboratory classification of the plasma cell dyscrasia monoclonal gammopathy of undetermined significance (MGUS) and the overt plasma cell neoplasm multiple myeloma (MM) is quite accurate, yet, for the most part, biologically uninformative. Most, if not all, cancers are caused by inherited or acquired genetic mutations that manifest themselves in altered gene expression patterns in the clonally related cancer cells. Microarray technology allows for qualitative and quantitative measurements of the expression levels of thousands of genes simultaneously, and it has now been used both to classify cancers that are morphologically indistinguishable and to predict response to therapy. It is anticipated that this information can also be used to develop molecular diagnostic models and to provide insight into mechanisms of disease progression, e.g., transition from healthy to benign hyperplasia or conversion of a benign hyperplasia to overt malignancy. However, standard data analysis techniques are not trivial to employ on these large data sets. Methodology designed to handle large data sets (or modified to do so) is needed to access the vital information contained in the genetic samples, which in turn can be used to develop more robust and accurate methods of clinical diagnostics and prognostics.

Results: Here we report on the application of a panel of statistical and data mining methodologies to classify groups of samples based on expression of 12,000 genes derived from a high density oligonucleotide microarray analysis of highly purified plasma cells from newly diagnosed MM, MGUS, and normal healthy donors. The three groups of samples are each tested against each other. The methods are found to be similar in their ability to predict group membership; all do quite well at predicting MM vs. normal and MGUS vs. normal. However, no method appears to be able to distinguish explicitly the genetic mechanisms between MM and MGUS. We believe this might be due to the lack of genetic differences between these two conditions, and may not be due to the failure of the models. We report the prediction errors for each of the models and each of the methods. Additionally, we report ROC curves for the results on group prediction.

Availability: Logistic regression: standard software, available, for example in SAS. Decision trees and boosted trees: C5.0 from www.rulequest.com. SVM: SVM-light is publicly available

from svmlight.joachims.org. Naïve Bayes and ensemble of voters are publicly available from www.biostat.wisc.edu/~mwaddell/eov.html. Nearest Shrunken Centroids is publicly available from http://www-stat.stanford.edu/~tibs/PAM.

KEYWORDS: Microarray, Logistic Regression, Boosted Decision Trees, Ensemble of Voters, Support Vector Machines, Nearest Shrunken Centroid, Multiple Myeloma, MGUS

Author Notes: This work supported in part by grants (CA38926-17, CA90998-02) from the NCI/ NIH, grant 9987841 from the NSF, and grant 1T15LM007359-01 from the NLM.

1. Introduction

The molecular mechanisms of the related plasma cell dyscrasias monoclonal gammopathy of undetermined significance (MGUS) and multiple myeloma (MM) are poorly understood. The poor understanding has important clinical implications because MGUS is a benign plasma cell hyperplasia whereas MM is a uniformly fatal malignancy. Monoclonal gammopathies are characterized by the detection of a monoclonal immunoglobulin in the serum or urine and underlying proliferation of a plasma cell/B lymphoid clone. (Kyle and Rajkumar, 1999.) Patients with MGUS have less advanced disease and are characterized by a small detectable plasma cell population in the marrow (< 10%) and secretion of a monoclonal protein detectable in the serum (<30g/L), but they lack clinical features of overt malignancy (such as lytic bone lesions, anemia, or hypercalcemia.) Patients with overt MM have increased marrow plasmacytosis (>10%), serum M protein (>30g/L), and generally present with anemia, lytic bone disease, hypercalcemia, or renal insufficiency.

Approximately 2% of all MGUS cases will convert to overt MM per year (International Myeloma Working Group, 2003), but it is virtually impossible to predict which of these cases will convert. A difficulty in the clinical management of MM is the extreme heterogeneity in survival, which can range from as little as two months to greater than eight years with only 20% of this variability being accounted for with current clinical laboratory tests. Thus, there is a great need for more robust methods of classification and stratification of these diseases. There is now strong evidence in a variety of cancers that global gene expression profiling can reveal a molecular heterogeneity of similar or related hematopoietic malignancy (Golub et. al., 1999; Alizadeh et. al., 2000.) In MM, the most differentially expressed genes in a comparison of normal and malignant cells can be used to identify changes that may point to the basic mechanisms of cellular transformation (Zhan et. al., 2002). These unique signatures could also be used for the development of molecular diagnostics that may be capable of identifying malignant cells even in the absence of any clinical manifestations, thus providing a means of early detection and possible prevention. We anticipate that expression profiling will provide the means of differentiating MGUS and MM at the molecular level. Here, we show that various methodologies applied to global gene expression data identified a class of genes whose altered expression is capable of discriminating normal and malignant plasma cells as well as classifying some MGUS as “like” MM and others as “unlike” MM. The predictive power of this small subset of genes suggests that their deregulated expression may not only prove useful in creation of molecular diagnostics, but may also provide important insight into the mechanisms of MM development and/or conversion from the benign condition MGUS to the overly malignant and uniformly fatal MM.

In previous work, comparison of gene expression profiles of bone marrow plasma cells from 32 normal, healthy donors and 74 untreated patients with MM revealed highly significant differences in both qualitative and quantitative gene expression. A statistical analysis showed that expression of 120 genes distinguished MM from normal cells. A total of 50 genes showed significant down-regulation in MM, and 70 genes were up-regulated in MM (Zhan et. al., 2002). With an unsupervised two-dimensional hierarchical clustering of 5,483 genes, MM and normal samples could also be differentiated (Zhan et. al., 2002). Importantly, however, these studies were unable to distinguish MGUS from MM. Additionally, these studies simply identified genes that were different across groups, while we are interested in using models to predict group membership.

The findings of Zhan et. al. lead us to consider predictive genetic models for discriminating between malignant and healthy samples. Using six different approaches: logistic regression, decision trees, support vector machines (SVM), Ensemble of Voters with 20 best information gain genes (EOV), naïve Bayes, and Nearest Shrunken Centroids (NSC), we identified subsets of genes that discriminated between the three types of samples (MM, MGUS, and normal healthy samples.) Additionally, using the genes from the original predictive models (MM versus normal), an analysis of the MGUS samples found that the MGUS samples were genetically much more similar to the MM samples than to the normal subjects. The identification of these classes of genes gives insight into the disease through their genetic mechanisms. The most interesting result of this work is the lack of ability of the methods to successfully discriminate between MM and the related disease MGUS, which has a similar laboratory presentation but is void of any clinical symptoms. Thus, these data suggest that in spite of its benign clinical course and only 2% conversion rate, MGUS may have genetic features that make it indistinguishable from the fatal MM. We hypothesize that the reason for our prediction failures is not due to inaccurate models, but rather, they are due to genetic mechanisms that divide MGUS into two distinct groups: those that will eventually convert to MM (and are grouped with the MM samples) and those that will remain inactive (and are grouped separately.)

2. BACKGROUND

We ran all six models on microarray data derived from Affymetrix (version 5) high density oligonucleotide microarray analysis. We compared 218 untreated MM samples, 45 healthy samples, and 21 samples designated as MGUS. We chose to use the normalization algorithm available from the Affymetrix software. Information on normalization and standardization of the microarray data is available on Affymetrix's website:
www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf

DOI: 10.2202/1544-6115.1018

3. Methods

Various methods were employed with two goals in mind. The first goal is to identify genes whose over or under expression are apparently essential in the comparison of healthy samples, MGUS samples, and malignant MM samples. The second goal is to identify optimal methods for use in analyzing microarray data and specifically methods applicable to analyzing microarray data on samples from MGUS and MM patients. Previous work has been done in identifying lists of genes that discriminate between the two types of samples (Zhan et. al., 2002; Chauhan et. al., 2002), but, to our knowledge, this is the first work that has been done on simultaneously identifying discriminatory genes and evaluating models to predict and describe the differences between myeloma, MGUS, and healthy samples.

Recently, the literature has contained numerous methods for discrimination between two or more classes of microarray samples. For example, Golub et. al. give an ad hoc measure of discrimination (Golub et. al., 1999); Tibshirani et. al. present a method called nearest shrunken centroid which improves on nearest centroid classification (Tibshirani et. al., 2002); and Dudoit et. al. present a comparison of multiple classification methods including Fisher's linear discriminant analysis, nearest neighbor classification, classification trees, boosted trees, bagged trees, and aggregate classifiers (Dudoit et. al., 2002.) We use six classification methods to provide results from a range of different techniques. In the same way as Dudoit et. al., we use classification trees as one of our models; additionally, we tried boosted trees, but these did not have any advantage over the more simple classification trees. Among the methods in the Dudoit et. al. paper, trees were intermediate performers; among our methods, trees were our worst performers. We feel that our other methods will produce classifications as least as good as or better than methods currently being used to discriminate between two microarray samples.

For each of the following methods (and each of the comparisons), we employed 10-fold cross validation to estimate the prediction error. Using 10-fold cross validation, $1/10^{th}$ of the data was removed (the 'test' data), and the entire model was created using only the remaining 90% of the data (the 'training' data.) The test data were then run through the training model and any misclassifications were noted. Error rates were computed by compiling the misclassifications from each of the 10 independent runs. Empirical results suggest that 10-fold cross validation may provide better accuracy estimates than the more common leave one out cross validation (Kohavi, 1995.)

Each of our methods is listed below and can be considered as a particular algorithm. For most of the methods we must first subset the list of ~12,000 genes due to computational limitations. Therefore, our results will be a comparison of

each of the complete algorithms that produced the result. We refer to the complete algorithm as the method or model.

3.1 Logistic Regression

The logistic procedure creates a model that predicts a binary value of the outcome (e.g., MM=1, normal = 0) based on numeric values of the gene:

$$\Pr(MM \mid X) = \frac{\exp(x^t \beta)}{1 + \exp(x^t \beta)}$$

where $x^t \beta = \beta_0 + x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p$ for the p most significant genes $X = \{ x_1, x_2, \ldots x_p \}$. The resulting value represents a predictive probability, for example, of being in the MM sample (predictive value close to one) or of being in the normal sample (predictive value close to zero.) The structure allows for knowledge of the uncertainty in predicting the group membership of future samples. For example, a new sample might be classified as MM with a predictive probability of 0.53, albeit with much less confidence than another sample whose predictive probability is 0.99. A gene with a positive value for the β coefficient indicates that an increase in gene expression correlates to a higher probability of classification to the more highly diseased group.

Logistic regression was applied using SAS software on the signal data, using two different gene selection methods. First, for each independent run of the data, the best subset function was used to find a small group of genes which best predicted the two groups of interest (MM vs. normal, MM vs. MGUS, or MGUS vs. normal.) The best subset function in SAS finds the best model of a given size out of the full set of parameters (either all the genes or a subset of genes.) To choose between the models of different sizes, we penalized the score function by the number of parameters, and took the model with the highest penalized score function. Because of computing limitations, we were required to first subset the genes (from 12625 genes down to about 50 genes) from the full set. This first subset was arrived at by ranking the contribution (p-value) of each gene individually in a univariate logistic model. We call this procedure the "best" subset selection for logistic regression.

Second, for each independent run of the data, forward selection was used to find a subset of genes which best predicted the two groups of interest. Again, because of computing limitations, we were required to first subset the genes (from 12625 down to 500 genes) from the full list. We maintain that the list of 500 genes give sufficient variety in the contribution to the model that it is equivalent to forward selection from the full set of 12625 genes. We call this procedure the "forward" subset selection for logistic regression.

DOI: 10.2202/1544-6115.1018

3.2 Decision Trees

Decision tree induction algorithms begin by finding the single feature (gene) that is most correlated with class, where the measure of correlation may be the correlation coefficient, mutual information (as in the system ID3), Gini index (as in the system CART), or one of several others. For concreteness of the present discussion, we will use mutual information and take the classes to be MM vs. normal (the same discussion applies to MM vs. MGUS and MGUS vs. normal). For each gene the algorithm computes the *information gain* of the detection and of the optimal split point for the real-valued measure (signal.) Information gain is defined as follows. The entropy of a data set is $-p \log_2 p - (1-p) \log_2(1-p)$ where p is the fraction of samples that are of class MM. A split takes one data set and divides it into two data sets: the set of data points for which the gene has a value below the split point (a particular value) and the set of data points for which the gene has a value above the split point. The information gain of the split is the entropy of the original data set minus the weighted sum of entropies of the two data sets resulting from the split, where these entropies are weighted by the fraction of data points in each set. This split yields a "decision stump," or decision tree with one internal node (split point value), as illustrated in Figure 1. If the new "leaf" or data set contains data points with different class values, then the algorithm recursively splits these "impure" nodes in the same fashion, until all decision nodes are pure. In practice, to avoid over-fitting, typical decision tree systems then "prune" the tree to get a smaller tree that is nearly consistent with the data though not necessarily completely consistent. Each leaf then makes the majority class prediction among data points that end at that leaf. In the present work we apply the decision tree system C5.0 (www.rulequest.com), one of the most widely used data mining algorithms on the market. C5.0 uses information gain to score splits; information about pruning is available on the aforementioned website.

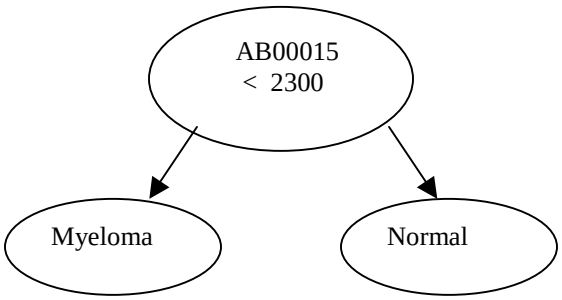

Figure 1. Example of a decision tree with one internal node, also called a "decision stump."

3.3 Ensemble of Voters (EOV)

Even with pruning, decision trees can sometimes over-fit the data. One approach to avoid over-fitting is to learn an *ensemble*, or a set of classifiers (in the present case, trees) that will vote on each new case to be predicted. The simplest ensemble scheme uses the *n* best-scoring classifiers (e.g., by information gain or accuracy), where predictions are made by unweighted voting. Progressively more sophisticated ensemble schemes include weighted voting of the top *n* classifiers, bagging (Breiman 1996) and boosting (Freund and Schapire, 1996). Our "Ensemble of Voters" (EOV) approach is the simplest ensemble scheme we could imagine. It predicts by unweighted voting of the top *n* decision stumps; a decision stump is the simplest decision tree -- a decision tree with a single internal node, or decision node, as shown in Figure 1. We rather arbitrarily committed to n=20 decision stumps before experimentation, largely for purposes of model comprehensibility; this number of genes is large enough to be interesting yet small enough not to be overwhelming. Afterward we verified that the results were not sensitive to this choice of n unless it went below 10 or above 100. In addition to this simplest ensemble scheme, we also employed boosting for trees (Freund and Schapire, 1996.) With boosting, after a model (e.g., tree) is learned, the few training data points that are inconsistent with the model are given increased weight (e.g., replicated), and training is repeated. This process repeats a number of times. Prediction is again by a vote, but where each model's vote is weighted by its training set accuracy.

In our experiments, boosting did not improve the performance of C5.0, so we do not report the results of boosted trees in this paper. We do report the results of EOV. It is worth noting that our EOV approach is similar to that employed by Golub and colleagues (Golub et al., 1999) with the exception that EOV uses unweighted rather than weighted voting. This observation raises the question of whether weighting the votes could perhaps improve performance of EOV. Of course, there are many different ways to choose the weights. One very natural choice leads to a widely-used algorithm in data mining and machine learning, known as "simple Bayes" or "naïve Bayes," which we describe next.

3.3 Naïve Bayes

Naïve Bayes is so named because it makes the (often) naïve assumption that all features (e.g. gene expression levels) are conditionally independent given the class value (e.g. MM or normal). In spite of this naïve assumption, in practice it often works very well. Like logistic regression, naïve Bayes returns a probability distribution over the class values. The model simply takes the form of Bayes' rule with the naïve conditional independence assumption:

$$\Pr(MM \mid X) = \frac{\Pr(x_1 \mid MM)...\Pr(x_n \mid MM)\Pr(MM)}{\Pr(X)}$$

DOI: 10.2202/1544-6115.1018

In practice we can ignore the denominator, and compute both *Pr(MM|X)* and *Pr(Normal|X)* normalized to sum to one. When X is a real-valued signal (gene expression value), we discretize using the split point that maximizes information gain on the training data (i.e., we use the best-scoring decision stump for this gene). Having discretized, we can again use the frequency in the data to estimate *Pr(x|MM).* One important methodological point is that the discretization step must be repeated on every fold of cross-validation using only the training data for that fold, or else the accuracy estimates for the approach will be overly-optimistic. When the features are discretized in the manner described, they are in fact the decision stumps considered by EOV. Hence naïve Bayes can be seen as a variant of EOV where the votes are weighted according to probabilities. We mentioned decision stumps in this section not to say that we are using decision stumps to calculate P(x_j|MM) (which is calculated using standard methods.) Instead, we hope to show a parallel between the underlying models built using naïve Bayes and built using ensembles of voters. In a sense, one can think of a naïve Bayes net as simply an ensemble of voting decision stumps where each stump's vote is weighted by its P(x_j|MM).

The question remains of how badly our performance is hurt by the naïve assumption in naïve Bayes. To determine this, we compare naïve Bayes with Bayesian network learning. A naïve Bayes model corresponds to a very simple form of Bayesian network, shown in Figure 2 below.

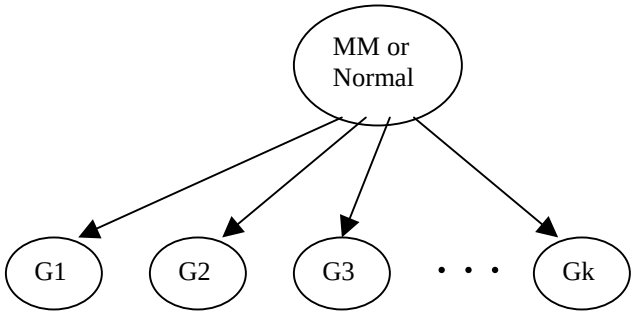

Figure 2. Structure of a naïve Bayes model where the class is MM vs. Normal and the features are genes G1, G2, …, Gk.

We used the top 10% of genes, as selected by information gain, in the naïve Bayes model. Ten percent was chosen rather arbitrarily so as to not overfit the data by recursive feature selection. We did not want to bias our accuracy results by trying all possible number of features and using the best one. The selection of features was repeated on every fold of cross-validation using only the training data for that fold.

Additionally, we tried applying Bayes' Net (Power Predictor) to the data, but the resulting model was consistently the Naïve Bayes model or a model that

did a worse job of prediction. For that reason, we leave those results out. We are currently working on an algorithm to find a better Bayes' Net structure.

3.4 Nearest Shrunken Centroid (NSC)

Nearest Shrunken Centroid classification is an extension of naïve Bayes and linear discriminant analysis (LDA). (Specifically, NSC with no shrinkage is equivalent to naïve Bayes using a Gaussian kernel and diagonal LDA.) The classification process is done by computing centroids for each class based on gene expression, and then the centroids are shrunken toward an overall centroid given by the entire data set. The shrinking process happens by a user-defined "threshold." The threshold is a specified value which is subtracted from each dimension of the centroid estimate. However, a centroid cannot be shifted across zero. So, if the centroid is a vector (1.2, 7, -.5) and the threshold is 3, the shrunken centroid will be (0, 4, -3.5). The threshold is chosen so as to minimize the prediction error. A sample is allocated to the closest centroid, using Euclidean distances. The process of shrinking allows for some genes to be removed from the prediction process, thus reducing the effect of noisy genes. This method is due to Tibshirani, Hastie, Narasimhan, and Chu, and further details are given in Tibshirani et. al., 2002.

3.5 Support Vector Machines

Support vector machines (SVMs) (Vapnik, 1998; Cristianini, 2000) are another novel data mining approach that has proven within the last three years to be well suited to gene expression microarray data (Brown et. al., 1999; Furey et. al., 2000.) At its simplest level, a support vector machine is an algorithm that attempts to find a linear separator between the data points of two classes. SVMs seek to maximize the *margin*, or separation between the two classes. Maximizing the margin can be viewed as an optimization task that can be solved with quadratic programming techniques. Of course, in practice there may be no good linear separator of the data. Support vector machines based on "kernel methods" can efficiently identify separators that belong to other functional classes. A commonly used kernel is the Gaussian kernel. Nevertheless, for gene expression microarray data, it has been repeatedly demonstrated empirically that simple linear SVMs give better performance (Brown et. al., 1999; Furey et. al., 2000) than SVMs with other kernels. Because it has been repeatedly observed that support vector machine performance can be improved by prior feature selection, the top 10% of the features as selected by information gain were provided to the SVM. Once again, 10% was chosen rather arbitrarily, and the selection of features was repeated on every fold of cross-validation using only the training data for that fold.

DOI: 10.2202/1544-6115.1018

4. Results

As mentioned, each model was tested using 10-fold cross validation to obtain error (misclassification) rates. For each of 10 runs of the data, 10% of the sample was removed and the prediction model was created. Then, using the created model, the test sample was predicted into groups, and the accuracy was recorded. After completing all 10 runs, the accuracy values were accumulated into the following table (Table 1.) (A balanced error rate can be computed from these values by taking a weighted average of the correct classification rates with respect to sample size.)

Table 1.

% correctly classified	MM	Normal	# genes	MM	MGUS	# genes	MGUS	Normal	# genes
Logistic – best	96.8%	84.4%	10	91.3%	38.1%	10	91.1%	71.4%	10
Logistic-forward	97.3%	84.4%	4	93.1%	38.1%	8	81.0%	91.1%	3
Trees	96.3%	91.1%	4	89.0%	23.8%	7	57.1%	91.1%	3
EOV	97.7%	100%	20	67.0%	90.5%	20	71.4%	100%	20
Naïve Bayes	97.3%	100%	250	93.1%	57.1%	250	76.2%	100%	250
SVM	97.7%	95.6%	250	96.8%	23.8%	250	76.2%	100%	250
NSC	95.9%	100%	11	77.1%	85.7%	102	98.2%	100%	408

There does not appear to be one methodology that stands out from the rest in terms of predicting group membership. In the difficult classification of MM vs. MGUS, Ensemble of Voters classifies the most MGUS correctly (90.48%), but the fewest MM correctly (66.97%.) Using naïve Bayes or nearest shrunken centroids may produce the best classification, though they do not seem to be appreciably better than the other methods. All the methods appear to be able to classify MM vs. Normal quite well and MGUS vs. Normal almost as well. However, as mentioned in the introduction, Trees appear to be the worst classifiers.

Notice that in Table 1 we have also provided the number of genes used to classify the dichotomous groups. These numbers are based on the full model which uses complete data to discriminate between the two groups (and not the cross validation models.) A small number of genes is useful in being able to interpret the model biologically. However, with a small number of genes, it is likely that the model won't be as robust as a model with a larger number of genes. For example, in the forward logistic regression, the three genes that were selected in the MGUS vs. Normal classification were the exact same genes selected in four of the cross validation models, but five of the cross validation models did not

include any of the three complete model genes. In contrast, 186 out of the 250 (74.4%) of the genes used in the Naïve Bayes comparison of MM vs. Normal were present in all cross validation models. It seems as though stability in modeling is a trade-off to biological interpretation with a small number of genes.

For each method, we have identified models for predicting group membership; we do not report the models here because we do not yet have knowledge of the compatibility of the quantitative values given by Affymetrix chips across different laboratories. However, because our results are based on cross validation, we feel that the models we have created would predict samples from our laboratory with a reasonable degree of accuracy (for Normal vs. MM and Normal vs. MGUS.) We have not provided models here, but for each comparison we have provided a list of genes that show up in the final model of at least two of the methods (tables 5, 6, and 7.) We would like to point out that the models classifying MM vs. MGUS had more overlap (22 overlapping genes) than the models classifying MM vs. Normal (8 overlapping genes) or MGUS vs. Normal (6 overlapping genes.) A possible explanation for this is that there are probably numerous genes that distinguish MM and normal samples because the two groups are quite distinct. However, the genetic similarities between MM and MGUS lead us to smaller number of genes that are different across the two groups. This dearth of distinguishing genes conditions any good model to contain some of the same limited number of genes.

4.1 Meta-Voting

As an additional step to improve the prediction capabilities of our methods, we calculated a "meta" prediction value. For a subset of the procedures we calculated the marginal predicted group and then gave a final prediction as the top voted group. (A sample is classified in a group if at least three of the five methods predict that group.) Results are given in Table 2.

Table 2.

total % correctly classified	MM vs. Normal	MM vs. MGUS	MGUS vs. Normal
Logistic – best	93.93%	88.70%	84.85%
Trees	95.45%	87.87%	78.79%
EOV	95.45%	69.04%	90.91%
Naïve Bayes	95.45%	89.96%	92.42%
SVM	96.97%	90.38%	92.42%
Meta	96.97%	91.21%	90.91%

From the table it is apparent that the meta voting procedure does not improve the overall results appreciably.

DOI: 10.2202/1544-6115.1018

4.2 Receiver Operator Characteristic (ROC) Curves

A Receiver Operating Characteristic (ROC) curve demonstrates the relationship between sensitivity (correct prediction to the more diseased group) and specificity (correct prediction to the less diseased group.) Figures 3, 4, and 5 give the ROC curves for the comparison of the different classifications; MM vs. normal, MM vs. MGUS, and MGUS vs. normal, respectively. The comparison of MM vs. MGUS is challenging for all the methods. For example, naïve Bayes has a high sensitivity but at the cost of low specificity. For even mediocre values of specificity, the sensitivity drops off quite rapidly. In order to have a high sensitivity for any of the methods (that is, in order to have very few false positives of MM) we compromise our ability to predict MGUS accurately (specificity.) However, for the two other comparisons, we see that high sensitivity does not come at the expense of high specificity.

4.3 Prediction of MGUS

The models that classify the MM and normal samples into distinct groups may also be able to be used as a predictive model for samples that are not clearly in either group based on clinical data. As a whole, the MGUS samples are healthy (except for high levels of immunoglobulins) but clinically appear malignant. Applying the MM vs. normal model to the MGUS samples will give us an idea as to which group each MGUS sample belongs. Table 3 provides the prediction distribution for the MGUS samples into the MM and normal groups based on five of the models which compared MM to normal samples. On average, about 75% of the MGUS samples are classified as MM, and about 25% are classified as normal. One possible reason for this split is that the 25% who are classified as normal may not have disease progression. Due to the fact that MGUS is a slow moving disease, we do not yet have progression data on these samples, so we cannot test this hypothesis. Regardless, the similarity of MGUS to MM (even in the model that was derived without any MGUS) gives additional evidence that the MGUS is actually genetically much more similar to the MM than to the normal samples.

Table 3.

MM vs. Normal (predicting MGUS)		
% MGUS classified as:	MM	Normal
Logistic – best	76.19%	23.81%
Trees	85.71%	14.23%
EOV	61.90%	38.10%
Naïve Bayes	76.19%	23.81%
SVM	80.95%	19.05%

In order to better understand the mechanisms behind the poor classification of the MGUS samples (when compared to MM), we tabulated the number of MGUS classified as MM for five of the methods, logistic regression (best subsets), EOV, Trees, Naïve Bayes, and SVM. Of the 21 MGUS samples, the misclassification rates are given in Table 4.

Table 4.

# MGUS Misclassified	Logistic (best)	EOV	Trees	Naïve Bayes	SVM
Logistic (best)	13	2	11	7	10
EOV		2	2	2	2
Trees			16	8	12
Naïve Bayes				9	8
SVM					16

There were 13 MGUS samples misclassified using the logistic procedure; 10 of the 13 were also misclassified using SVM, and 11 of the 13 were misclassified using Decision Trees. All of the EOV misclassifications were also misclassified using the other methods. Of the 9 misclassification using Naïve Bayes, almost all of them were misclassified using logistic, SVM, and Trees. This cross tabulation indicates that the misclassified MGUS samples are continuously getting misclassified which lends evidence to a possible subset of MGUS samples that are genetically similar to the MM samples.

5. Conclusion

In this manuscript we have compared six different statistical and data mining algorithms for their ability to discriminate normal, hyperplastic (MGUS), and malignant (MM) cells based on the expression patterns of ~12,000 genes. The models were highly accurate in distinguishing normal plasma cells from abnormal cells, however they displayed a modest failure in the discrimination between the hyperplasic cells and malignant cells. A goal of this study was to develop or modify statistical and data mining tools in order to capture a small subset of genes, from massive gene expression data sets, capable of accurately distinguishing groups of cells, e.g. normal, precancerous, and cancerous cells, with the ultimate goal to create sensitive and reproducible molecular-based diagnostic tests. Genes that our models established as important in prediction are listed in tables 5, 6, and 7. In addition, future studies will be aimed at using a similar strategy to identify a minimum subset of genes capable of discriminating subgroups of disease for risk stratification and prognostics. It is particularly important to understand the genetic mechanisms for multiple myeloma as the overall survival in MM is highly variable, with some patients surviving as long as

DOI: 10.2202/1544-6115.1018

10 years and others dying within several months of diagnosis. Current microarrray studies require the isolation of large numbers of cells that necessitate advanced facilities and expertise. Our studies represent the first step toward streamlining this process, as a smaller subset of genes (10-20) with a high predictive power allows for a massive reduction in scale, which in turn will make development of a commercial test more amenable to mass production and hence widespread clinical use.

MGUS is the most common plasma cell dyscrasia occurring in up to 2% of the population over age 50 (Kyle et. al., 2002.) The differentiation of MGUS and MM is based on a combination of clinical criteria such as the amount of bone marrow plasmacytosis, the concentration of monoclonal immunoglobulin, the presence of bone lesions, and kidney malfunction. Especially in early phases of MM, the differential diagnosis may be associated with a degree of uncertainty. Thus, it is imperative to determine if post-genome era technologies can be used to overcome these limitations. However, results from our studies suggest that development of a molecular test for discrimination based on global gene expression patterns will be more of a challenge than originally anticipated. The results pose somewhat of a paradox in that although MGUS has all the features of malignancy, the disease lacks clinical symptoms, and in addition, very few MGUS cases will convert to overt MM over the lifetime of the patient.

One possible reason for the inability of the models to discriminate MGUS from MM is that MGUS represents at least two different diseases (as hypothesized by the overlap in misclassification of MGUS samples stated in section 4.3.) In simplistic terms, MGUS can be viewed as a disease that will remain indolent or one that will convert to overt malignancy. Since ours is a prospective study, we do not have outcome data on the MGUS population which would enable a separation of the MGUS samples into two distinct diseases. With outcome data, it may be possible to show that the models are, in fact, more accurate than realized. Accruing sufficient numbers of stable and progressive MGUS cases along with sufficient follow-up time will help resolve this hypothesis.

The failure of the models to differentiate the two disease types could be related to the limitations of the current methodologies. The microarray profiling utilized here only interrogated 1/3 of the estimated 35,000 human genes (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001), thus it is possible that a whole genome survey would reveal discriminating features. We are currently investigating the possibility by performing microarray analysis with the new Affymetrix U133 GeneChip system, which is thought to interrogate all human genes. It is also possible that the full genome analysis will reveal no significant differences. Such a revelation could mean any of a variety of possibilities: (1) there is no genetic difference between the two diseases, (2)

only the MGUS that are classified as MM are genetically similar to the MM, and the clinical tests are unable to identify that distinction, (3) the current microarray technology is not specific enough to measure the genetic differences between the two diseases, (4) the methods described above are not appropriate for this type of analysis. If (1) or (2) is true, results would point to other determinants of an indolent or malignant course such as genetic predisposition or somatic DNA mutations not manifest in gene expression, a unique environmental exposure interacting with these predisposing genetic traits, or a non-tumor cell microenvironment or "soil" that promotes plasma cell growth.

In conclusion, it is anticipated that strategies like those employed here will allow the creation of new molecular diagnostic and prognostic tests and should provide useful insight into the genetic mechanisms of neoplastic transformation.

DOI: 10.2202/1544-6115.1018

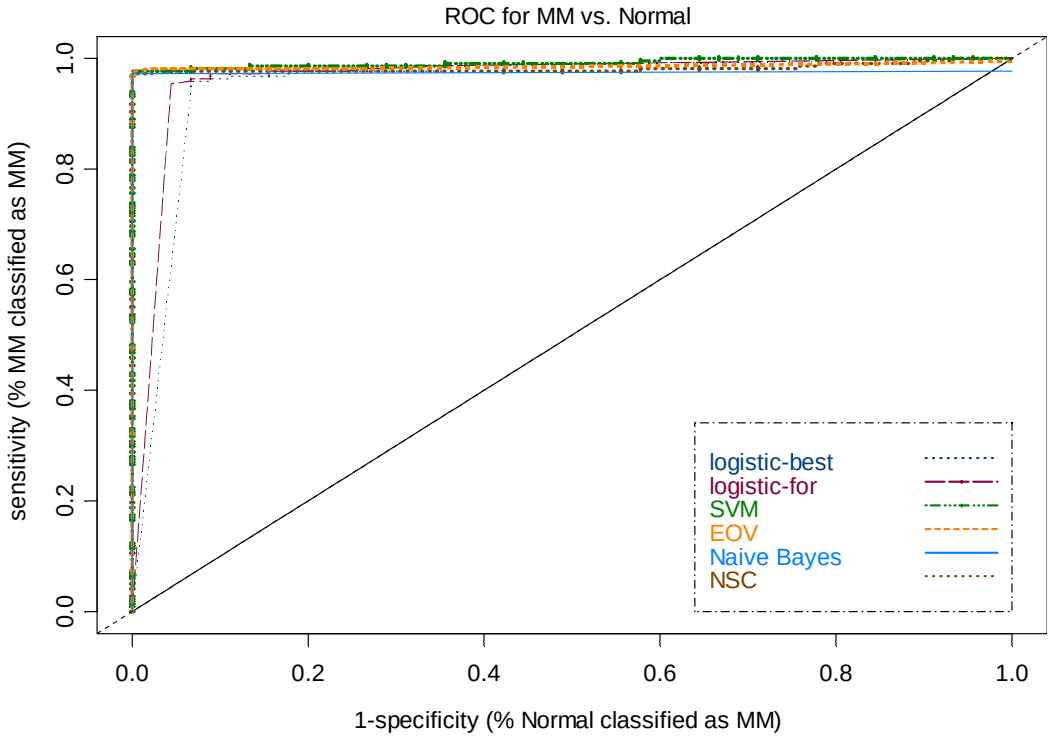

Figure 3. ROC curves. Six models are compared on each curve: logistic regression (best and forward), SVM, EOV, naïve Bayes, and NSC. The x-axis models one minus the specificity while the y-axis models the sensitivity. The diagonal line across the graph represents the line y=x. The relationship between MM and normal is fairly straightforward to model with any of the methods. (Lines are not drawn for decision trees because of the computing complications associated with creating ROC curves for that method.)

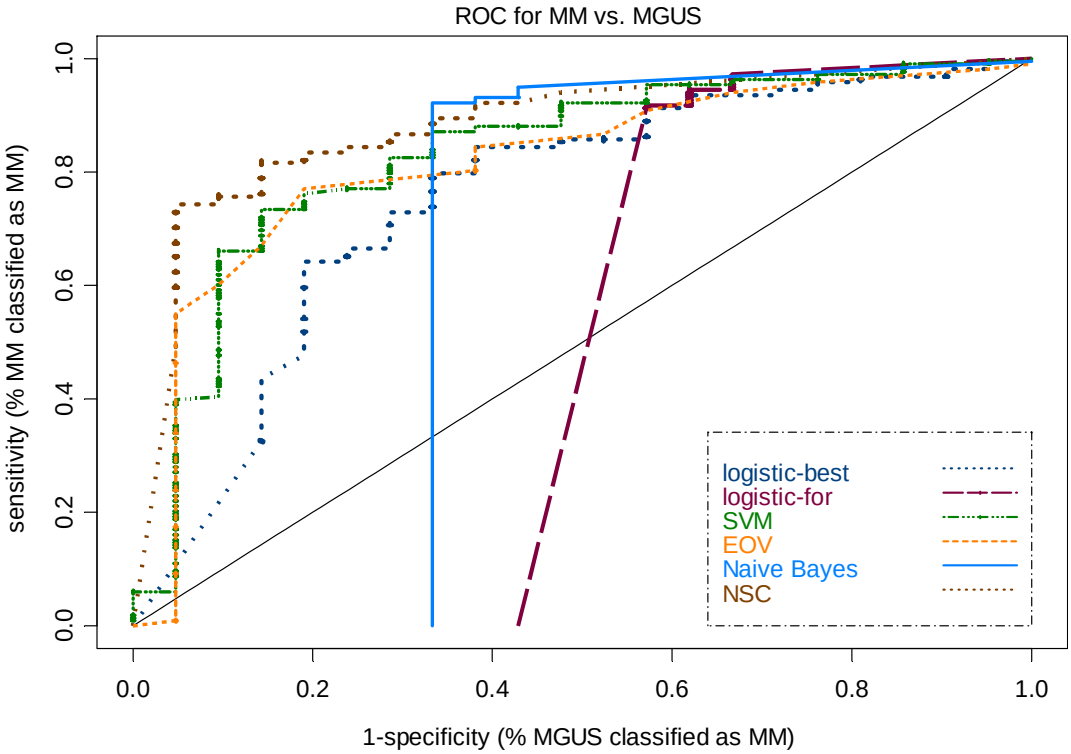

Figure 4. ROC curves. Six models are compared on each curve: logistic regression (best and forward), SVM, EOV, naïve Bayes, and NSC. The x-axis models one minus the specificity while the y-axis models the sensitivity. The diagonal line across the graph represents the line y=x. The relationship between MM and MGUS is difficult to characterize which is reflected in the lack of ability to keep specificity high for high values of sensitivity for any of the models. (Lines are not drawn for decision trees because of the computing complications associated with creating ROC curves for that method.)

DOI: 10.2202/1544-6115.1018

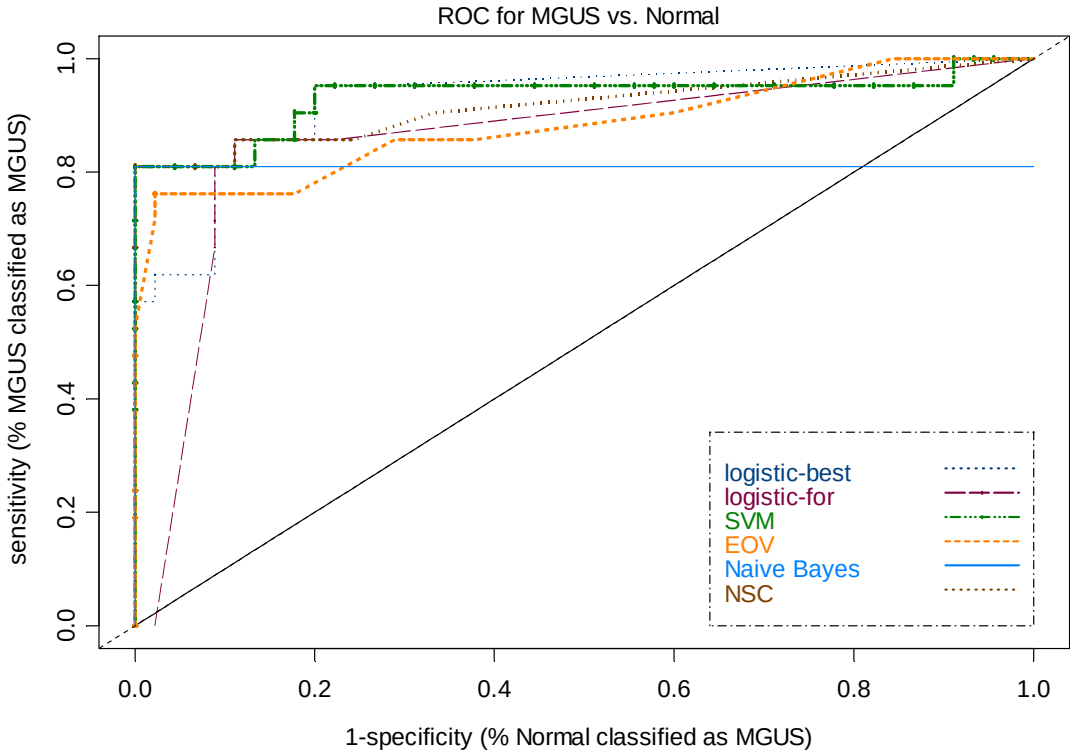

Figure 5. ROC curves. Six models are compared on each curve: logistic regression (best and forward), SVM, EOV, naïve Bayes, and NSC. The x-axis models one minus the specificity while the y-axis models the sensitivity. The diagonal line across the graph represents the line y=x. The relationship between MGUS and normal is fairly straightforward to model with any of the methods. (Lines are not drawn for decision trees because of the computing complications associated with creating ROC curves for that method.)

Table 5. Genes which are essential in at least two of the methods when comparing MM vs. Normal plasma cells.

Array ID	Title
L02867-689_at	paraneoplastic antigen
M25915-36780_at	"clusterin (complement lysis inhibitor, SP-40,40, sulfated glycoprotein 2, testosterone-repressed prostate message 2, apolipoprotein J)"
U14187-34573_at	ephrin-A3
W26381-39490_f_at	ADP-ribosylation factor GTPase activating protein 1
X16832-37021_at	cathepsin H
X76079-1988_at	"platelet-derived growth factor receptor, alpha polypeptide"
AF088219-37085_g_at	lysozyme homolog
AB009598-35179_at	beta-1,3-glucuronyltransferase 3 (glucuronosyltransferase I)

Table 6. Genes which are essential in at least two of the methods when comparing MGUS vs. Normal plasma cells.

Array ID	Title
AA522530-39827_at	HIF-1 responsive RTP801
AB005297-35897_r_at	brain-specific angiogenesis inhibitor 1
D80010-38098_at	lipin 1
X62025-32204_at	"phosphodiesterase 6G, cGMP-specific, rod, gamma"
X66945-424_s_at	"fibroblast growth factor receptor 1 (fms-related tyrosine kinase 2, Pfeiffer syndrome)"
X80026-40094_r_at	Lutheran blood group (Auberger b antigen included)

DOI: 10.2202/1544-6115.1018

Table 7. Genes which are essential in at least two of the methods when comparing MM vs. MGUS plasma cells.

Array ID	Title
AA976838-41764_at	apolipoprotein C-I
AB014590-36520_at	KIAA0690 protein
AB020687-37684_at	"solute carrier family 21 (organic anion transporter), member 9"
AF007155-40472_at	"Homo sapiens clone 23763 unknown mRNA, partial cds"
AI762213-32821_at	lipocalin 2 (oncogene 24p3)
AJ130718-33731_at	"solute carrier family 7 (cationic amino acid transporter, y+ system), member 7"
D82348-38811_at	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase
J02854-39145_at	"myosin, light polypeptide 9, regulatory"
L36033-33834_at	stromal cell-derived factor 1
M12529-608_at	apolipoprotein E
M30257-583_s_at	vascular cell adhesion molecule 1
M34379-37096_at	"elastase 2, neutrophil"
M73255-41433_at	vascular cell adhesion molecule 1
M83667-1052_s_at	"CCAAT/enhancer binding protein (C/EBP), delta"
U03057-39070_at	"singed-like (fascin homolog, sea urchin) (Drosophila)"
X03084-38796_at	"complement component 1, q subcomponent, beta polypeptide"
X84740-1188_g_at	"ligase III, DNA, ATP-dependent"
X95735-36958_at	zyxin
Z22971-31438_s_at	CD163 antigen
Z82244-33802_at	heme oxygenase (decycling) 1
J03358-35133_at	fer (fps/fes related) tyrosine kinase (phosphoprotein NCP94)
L36033-33834_at	stromal cell-derived factor 1

REFERENCES

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P., and Staudt, L. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123-140.

Brown, M., Grundy, W., Lin, D., Christianini, N., Sugnet, C., Ares, M., and Haussler, D. (1999) Support vector machine classification of microarray gene expression data. UCSC-CRL 99-09, Department of Computer Science, University California Santa Cruz, Santa Cruz, CA.

Chauhan, D., Auclair D., Robinson, E.K., Hideshima, T., Li, G., Podar, K., Gupta, D., Richardson, P., Schlossman, R.L., Krett, N., Chen, L.B., Munshi, N.C., and Anderson, K.C. (2002) Identification of genes regulated by Dexamethasone in multiple myeloma cells using oligonucleotide arrays. *Oncogene*, **21**, 1346-1358.

Cristianini, N. and Shawe-Taylor, J. (2000) An Introduction to Support Vector Machines and other kernel-based learning methods, **Cambridge University Press**.

Dudoit, S., Fridlyand, J., Speed, T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77-87.

Freund, Y., and Schapire, R.E. (1996) Experiments with a New Boosting Algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning (ICML'1996)*. Morgan Kaufmann, 148-156.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., and Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906-914.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and

Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**, 860-921.

International Myeloma Working Group. (2003) Criteria for the classification of monoclonal gammopathies, multiple myeloma, and related disorders; a report of the international myeloma working group. *British Journal of Haematology* **121**, 749-757.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).*

Kyle, R. A., and Rajkumar, S. V. (1999) Monoclonal gammopathies of undetermined significance. *Hematol Oncol Clin North Am,* **13**, 1181-202.

Kyle, R. A., Therneau T. M., Rajkumar, S. V., Offord, J. R., Larson, D. R., Plevak, M. F., Melton, L. J. III. (2002) A long term study of prognosis in monoclonal gammopathy of undertermined significance. *N Engl J Med*, **346,** 546-549.

Tibshirani, R., Hastie, T., Balasubramanian, N., Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, **99**, 6567-6572.

Vapnik, V. (1998) Statistical Learning Theory, **John Wiley & Sons**.

Venter, J. C., Adams, M. D., Myers, E. W., et. al. (2001) The sequence of the human genome. *Science*, **291,** 1304-51.

Zhan, F., Hardin, J., Kordsmeier, B., Bumm, K., Zheng, M., Tian, E., Sanderson, R., Yang, Y., Wilson, C., Zangari, M., Anaissie, E., Morris, C., Muwalla, F.,Van Rhee, F., Fassas, A., Crowley, J., Tricot, G., Barlogie, B., and Shaughnessy, Jr., J. (2002) Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. *Blood,* **99**, 1745-1757.