



# Computer Sciences Department

**Comparative Data Mining for Microarrays:  
A Case Study Based on Multiple Myeloma**

David Page, Fenghuang Zhan, James Cussens,  
Michael Waddell, Johanna Hardin, Bart Barlogie,  
John Shaughnessy, Jr.

Technical Report #1453

November 2002

UNIVERSITY OF  
WISCONSIN  
MADISON

# Comparative Data Mining for Microarrays: A Case Study Based on Multiple Myeloma

**David Page\***

Dept. of Biostatistics and  
Medical Informatics &  
Dept. of Computer Sciences  
University of Wisconsin  
Madison, WI 53706

**Fenghuang Zhan**

Lambert Laboratory of  
Myeloma Genetics  
University of Arkansas for  
Medical Sciences  
Little Rock, AR 72205

**James Cussens**

Computer Science Dept.  
University of York  
Heslington  
York YO10 5DD  
United Kingdom

**Michael Waddell**

Dept. of Computer Sciences &  
Dept. of Biostatistics and  
Medical Informatics  
University of Wisconsin  
Madison, WI 53706

**Johanna Hardin**

Southwest Oncology Group  
Fred Hutchinson Cancer  
Research Center  
Seattle, WA 98109

**Bart Barlogie**

Myeloma Institute for  
Research and Therapy  
University of Arkansas for  
Medical Sciences  
Little Rock, AR 72205

**John Shaughnessy, Jr.**

Lambert Laboratory of  
Myeloma Genetics  
University of Arkansas for  
Medical Sciences  
Little Rock, AR 72205

**Keywords:** gene expression microarrays, Bayesian networks, boosted decision trees, support vector machines, ensembles.

**Motivation:** Supervised machine learning and data mining tools have become popular for the analysis of gene expression microarray data. They have the potential to uncover new therapeutic targets for diseases, to predict how patients will respond to specific treatments, and to uncover regulatory relationships among genes in normal and disease situations. Comparative experiments are needed to identify the advantages of the leading supervised learning algorithms for microarray data, as well as to give direction in methodological decisions.

**Results:** This paper compares support vector machines, Bayesian networks, decision trees, boosted decision trees, and voting (ensembles of decision stumps) on a new microarray data set for cancer with over 100 samples. The paper provides evidence for several important lessons for mining microarray data, including: (1) Bayes nets and ensembles perform at least as well as other approaches but arguably provide more direct insight; (2) the common practice of throwing out low or negative average differences, or those accompanied by an "absent" call, is a mistake; (3) looking for *consistent* differences in expression may be more important than *large* differences.

**Availability:** All systems used are available online from the authors or others (Section 3). The new cancer data set is available online from the authors at <http://lambertlab.uams.edu/publicdata.htm>.

---

\* To whom correspondence should be addressed. Fax: 608-265-7916. Email: [dpage@cs.wisc.edu](mailto:dpage@cs.wisc.edu).

## 1. INTRODUCTION

Early studies of gene expression microarray data relied primarily on pairwise “fold-change” comparisons of values and on clustering. More recently, supervised data mining algorithms such as support vector machines (Furey *et al.*, 2000), ensemble methods (Golub *et al.*, 1999; Slonim *et al.*, 2000), and Bayes nets (Friedman *et al.*, 2000; Pe’er *et al.*, 2001) have become popular for analysis of microarray data.<sup>1</sup> At least the first two approaches have been tested on two-class cancer data sets of between 30 and 72 total samples. But questions remain about which supervised data mining algorithms are most appropriate for microarray data. Furthermore, even after one has committed to a particular algorithm, other important design decisions remain. For example, should some genes be omitted because of low or negative expression levels (e.g., in Affymetrix™ data)? Therefore, further application studies are needed to help give guidance in the selection and application of supervised data mining algorithms to microarray data. The purpose of this paper is to present one such application study. We compare SVMs, ensembles (voting), Bayes nets, decision trees, and boosted trees on a new, publicly-available cancer data set consisting of 105 highly purified plasma cell samples, from 74 newly diagnosed cancer samples and 31 normal healthy donors.<sup>2</sup> The cancer being studied is multiple myeloma, an incurable malignancy of immunoglobulin secreting plasma cells that grow and expand in the bone marrow. The microarray technology being employed is the Affymetrix oligo-based approach. The experiments provide evidence for the following lessons.

1. A directly comprehensible model, such as a Bayes net, decision tree, or ensemble of voters, has the advantage of exposing “trivially-accurate genes.” These are genes that provide no new insight but are highly-accurate owing to the nature of the disease or of sample collection. If the data mining goal is to obtain insight, and not just an accurate predictor, then trivially-accurate genes should be removed through consultation with a domain expert.
2. Unweighted voting and Bayes net learning provide accuracies at least as high as the other approaches (Figure 1) and arguably provide much more direct insight than they do.
3. Information gain provides a very different sort of insight than the traditional “fold-change” measure for comparing a gene’s expression measurements across different samples. Information gain looks for a high level of *consistency* in differential expression rather than *large* differences between some samples of different classes. Considering *consistency* alone works surprisingly well.
4. The ability of SVMs to consider the *magnitude* of the difference in expression in addition to the *consistency* of the difference appears to yield little or no benefit beyond methods that consider only the consistency of the difference.
5. Throwing out data based on low or negative average difference values or absent calls is a mistake, at least with data generated using Affymetrix technology before 2002.

To our knowledge, the present paper is the first to report a comparative experiment of such a wide variety of the leading supervised data mining algorithms on a gene expression microarray data set.

Figure 1 summarizes the accuracies of each of the five supervised data mining algorithms examined in this paper. Figures 2 and 3 summarize the highlights from voting – the genes that are most predictive of Myeloma vs. Normal according to information gain.

---

<sup>1</sup> Bayes nets are not necessarily a “supervised” approach, but when a class value is included as a variable they can be viewed as such. We employ a Bayes net learning algorithm tailored to classification and hence to supervised learning.

<sup>2</sup> This data set is likely to grow to the order of 500 samples in the next year.

Method	AC Only	AC+AD
Trees	90.5	98.1
Boosted Trees	96.2	99.0
SVMs	95.2	93.3
Vote	94.0	100.0
Bayes Nets	95.2	100.0

**Figure 1.** Summary of accuracies, by leave-one-out cross-validation, of data mining techniques applied to predicting Myeloma vs. Normal. The column labeled “AC” gives performance when using only the Affymetrix Absolute (Absent-Present) Call. “AC+AD” uses both Absolute Call and Average Difference. The significant (sign test, 0.05 level) differences are the following. Using AC only, all methods significantly outperform decision trees. Using AC+AD, all methods significantly outperform SVMs. Further discussion appears in Section 4.

Score	Gene	Accession Number	Split	MH	ML	NH	NL
0.80	<i>APOA2</i>	X04898	-777	74	0	1	30
0.74	<i>HERV K22 pol</i>	K03498	637	3	71	31	0
0.70	<i>TERT</i>	AF015950	-1610	70	4	0	31
0.70	<i>UMOD</i>	M15881	1119.1	0	74	28	3
0.70	<i>CDH4</i>	L34059	-278	74	0	3	28
0.66	<i>ACTR1A</i>	Z14978	3400.6	3	71	30	1
0.66	<i>MASP1</i>	D17525	-536.6	71	3	1	30
0.65	<i>PTPN21</i>	X79510	1256.1	6	68	31	0

**Figure 2.** The eight genes with the top information gain scores according to absolute call or average difference. All the top-scoring features were average difference features. “Score” is the information gain score, and “Split” is the value for Average Difference at which the split is made. MH is the number of samples that have class “Myeloma” (M) and an average difference higher (H) than the split value. ML is the number of samples with class “Myeloma” and average difference lower than the split value. NH and NL are analogous for samples with class value “Normal.”

Score	Gene	Accession Number	MH	ML	NH	NL
0.45	<i>HIF2</i>	X57129	57	17	0	31
0.44	<i>NCBP2</i>	D59253	57	17	0	31
0.43	<i>SM15</i>	U73167	56	18	0	31
0.43	<i>GCN5L2</i>	U57316	56	18	0	31
0.41	<i>MHC2 beta W52</i>	HT3779	12	62	29	2
0.41	<i>RNASE6</i>	U64998	15	59	30	1
0.41	<i>TNFRSF7</i>	M63928	15	59	30	1
0.41	<i>SDF1</i>	L36033	15	59	30	1

**Figure 3.** The eight top-scoring genes by information gain when using only Absolute (Absent-Present) Calls. We equate “H” with “Present” to retain the notation from Figure 2.

The remainder of the paper is organized as follows. Section 2 provides basic background about multiple myeloma and about the data set; this brief background is necessary for the sections that follow. Section 3 discusses the methodology of the data mining experiments. Section 4 discusses the results and the lessons that these results support. Section 5 summarizes the conclusions and identifies important questions and directions for further work.

## 2. BACKGROUND AND MATERIALS

Multiple myeloma is a cancer of antibody secreting plasma cells that grow and expand in the bone marrow. Although multiple myeloma is hypoproliferative (the cancer cells replicate at a relatively low rate), the disease is incurable and usually progresses rapidly after diagnosis, with bone demineralization, renal failure, anemia, and secondary infections resulting from immunosuppression as common causes of mortality.

A healthy body is capable of producing millions of distinct antibodies through a combinatorial process in which so called variable region (V), joining region (J), and diversity region (D) genes undergo site specific DNA recombination to create unique antigen binding domains. This process ensures that essentially all potential infectious agents will be recognized and eliminated during an immune response. Thus, a normal sample of plasma cells is *polyclonal*, containing a large number of plasma cells that each produce a different antibody. In contrast, a sample of plasma cells from a patient with multiple myeloma will be *monoclonal*, containing plasma cells that are all identical. This is due to cancer's nature as an uncontrolled growth and expansion of the progeny of a single aberrant cell, e.g. a plasma cell from a multiple myeloma patient. Hence by the time diagnosis of multiple myeloma is made, the plasma cells producing this one antibody have taken over the bone marrow, where all blood cells are normally produced. This eliminates the capacity of the bone marrow to produce the normal variety of antibody secreting plasma cells as well as the normal red and white blood cells. This leads to the anemia and immunosuppression mentioned above. This difference in expression between monoclonal and polyclonal samples is important for portions of the discussion in the next section.

The data were produced by purifying plasma cell samples from 74 newly-diagnosed multiple myeloma patients and 31 normal healthy donors, extracting total mRNA from these plasma cell samples, converting this RNA to biotinylated cRNA, and then hybridizing the cRNA to microarrays. Further details about this process are available from Zhan *et al.* (in press). The resulting Affymetrix output files are publicly available at [lambertlab.uams.edu/publicdata.htm](http://lambertlab.uams.edu/publicdata.htm).

Details about the Affymetrix process are available at [www.affymetrix.com](http://www.affymetrix.com) and are beyond the scope of this paper. But a very brief overview is needed for parts of the discussion that follows. One Affymetrix file is generated for each patient or sample. For each gene, this file contains two values. One is the Absolute Call (AC), taking values A (Absent), P (Present), or M (marginal, occurring only about 4% of the time in our data, which is consistent with many other Affymetrix data sets). The other value is the Average Difference (AD), which is a floating-point value that can be positive or negative. In a nutshell, AD compares hybridization with 25-mers that are known to appear in a gene against hybridization with the same 25-mers except that the middle (13<sup>th</sup>) base has been changed to its complement. A negative AD means the "mismatch" 25-mers tend to hybridize more than the "perfect match" ones. For this reason some researchers choose to ignore negative values or values for which the corresponding AC is Absent, while others choose to use these values. We chose not to eliminate or modify any values in the data, and we will return to this point several times in the discussion that follows.

Our set of 105 Affymetrix files, one per patient or sample, was converted into a single file with one row per sample and two columns per gene, one for the AC and one for the AD. In addition we added a final column holding the class value, "Myeloma" or "Normal." The file format is illustrated in Figure 4. We view a data point as a row, or sample, with the columns being features.

		Accession Number					Class		
		A28202		AB00014		AB00015		...	
Person	Person 1	P	1142.0	A	321.0	P	2567.2	...	myeloma
	Person 2	A	-586.3	P	586.1	P	759.0	...	normal
	Person 3	A	105.2	A	559.3	P	3210.7	...	myeloma
	Person 4	P	-42.8	A	692.1	P	812.0	...	normal
	...	...	...	...	...	...	...	...	...

**Figure 4.** Illustration of the input file form for data mining runs.

The goal of mining this data is to gain new insights into multiple myeloma development and also to identify new therapeutic targets. For example, suppose a gene is expressed at a consistently higher level in myeloma samples than in normal. Then perhaps the protein that is expressed from this gene has a key role in the development or progression of the disease. If so, then a small molecule that will bind to this protein, in a way that inhibits its activity, could be an effective drug to treat multiple myeloma. With this goal in mind, it is clear that comprehensibility of the data mining results is paramount. An accurate but incomprehensible predictor is of little value in the search for new therapeutic targets. Nevertheless, we measure accuracy of predictors because a comprehensible but inaccurate predictor also is of little value. *We seek accurate predictors that will provide insight into the disease.* The goal is the same in most other applications of data mining to microarray data for diseases.

### 3. METHODOLOGY

We ran all five data mining approaches listed in Figure 1 on the data in the form shown in Figure 4 (modulo syntactic changes required by the different software systems). For trees and boosted trees, we used C5.0 ([www.rulequest.com](http://www.rulequest.com)). For support vector machines, we used SVM<sup>light</sup> ([svmlight.joachims.org](http://svmlight.joachims.org)). We experimented with both linear support vector machines and Gaussian kernels but found linear SVMs performed better, which is consistent with the results of Furey *et al.* (2000) which used SVMs on similar but somewhat smaller cancer data sets. The voting algorithm we employed scored all features according to entropy-based information gain, kept the top scoring 1% of the features, and took a majority vote among these features. In other words, we used an unweighted ensemble of decision stumps with the number of stumps equal to 1% (70 in this case) of the number of features (roughly 7000).<sup>3</sup> The Bayes net application is slightly more involved and is motivated and described in the following paragraph.

Bayes net learning algorithms are being applied to attempt to uncover regulatory information from microarray data. They are very well suited in general to modeling probability distributions and revealing conditional dependencies among variables. But conventional wisdom holds that for a pure classification task Bayes nets are inferior to classification algorithms such as those named in the last paragraph. For this reason we would not have applied a Bayes net learner to the present task but for recent lessons from KDD Cup 2001 [Cheng *et al.*, 2002] (or see [www.cs.wisc.edu/~dpage/kddcup2001](http://www.cs.wisc.edu/~dpage/kddcup2001)). In that competition a Bayes net learner tailored to classification outperformed 113 other classification approaches on a task with similar properties to

<sup>3</sup> It is worth noting that one can obtain higher accuracies for voting in the “AC only” case by playing around with the percentage of features used in the ensemble. We chose 1% before running any experiments because it is a commonly used value. We did not modify it based on the cross-validation results because this is a kind of “cheating” that can give an overly-optimistic impression of performance.

the present one, albeit a drug design task. Our knowledge of this result led us to try that same algorithm, BayesNet PowerPredictor ([www.cs.ualberta.ca/~jcheng](http://www.cs.ualberta.ca/~jcheng)), with the same methodology employed there. Because the learning algorithm could not work with more than 30 features, we first used information gain to narrow the set of thousands of features to the top 30; analogous feature selection was used in the KDD Cup application. Because the underlying algorithm could not directly use continuous values, we discretized AD values based on whether they were greater than the split value that gave optimal information gain for that feature. We then applied BayesNet PowerPredictor. This approach can be viewed as a sophisticated form of weighted voting that takes into account not only the strength of correlation between a feature and class value, but also the correlations among different features.

#### 4. RESULTS

On the initial runs of voting, Bayes nets and trees, the features with highest information gain were genes associated with immune function, such as *IGL* and *IGHM*. These genes were absent or had very low expression in myeloma samples, but they were present or had higher expression in normal samples. As discussed in Section 2, this difference is because of the monoclonal versus polyclonal nature of the myeloma and normal plasma cell populations. These genes are unlikely to provide important disease specific information. Therefore, although the predictions were highly accurate, the domain experts in this work advocated removing all immunoglobulin (IG) and HLA genes, which we did through interaction with GeneCards ([bioinfo.weizmann.ac.il/cards](http://bioinfo.weizmann.ac.il/cards)). This removal of such “trivially-accurate” genes ensures that if a tool gives a high accuracy it will be based on novel insights into the disease. We are not certain whether this issue of trivially-accurate genes or predictors occurs in other applications of supervised data mining to microarray data. But we recommend (*lesson 1*) that those carrying out such applications begin by asking whether this might be the case for their application. All results reported in this paper are after the removal of all IG and HLA genes.

Having removed all IG and HLA genes, we then ran the five supervised data mining tools as described. We also ran them all using only the Absolute Calls, because (1) information gain as used by trees and voting sometimes overfits floating point features such as Average Difference, and (2) it is believed by some that Absolute Calls are less noisy. We wanted to test whether the results using AC only were as strong as AC+AD. All results were reported in Figure 1, copied below as Figure 5 for the reader’s convenience. The results were almost uniformly better with AC+AD. The lone exception is with SVMs and leads to our lesson 4, discussed near the end of this section. Because the AC+AD results were almost always better, we begin with a discussion of those.

Method	AC Only	AC+AD
Trees	90.5	98.1
Boosted Trees	96.2	99.0
SVMs	95.2	93.3
Vote	94.0	100.0
Bayes Nets	95.2	100.0

**Figure 5.** Leave-out-one cross-validation results for Myeloma vs. Normal (copied from Figure 1).

Voting and Bayes nets clearly produce the best results given AC+AD, although their differences with trees and boosted trees are not significant. *Lesson 2* is that, in addition to being at least as

accurate as any other approach, voting and Bayes nets provide greater direct insight than the other approaches. With the results of both of these tools, one can see at a glance which genes are important for distinguishing between the classes; this is not the case for boosted trees or SVMs. For example, contrast the list of top voters as given in Figure 2 (repeated as Figure 6 below), or the Bayes net in Figure 7, with a collection of ten weighted trees or a collection of SVM coefficients. Furthermore, while a single decision tree is at least as comprehensible as a Bayes net or collection of top voters, the single tree provides less information. A typical tree in the cross-validation runs uses only two or three genes, whereas voting allows one to immediately see all the genes that are most consistently differentially expressed.<sup>4</sup> Bayes nets and voting have a trade-off with regard to insight. The Bayes net exposes correlations among groups of genes that are not evident with the voting results. But the Bayes net may discard some genes of interest if a smaller subset has equal predictive power. The average number of features retained by the Bayes net learning algorithm over the cross-validation runs was roughly 20 out of the 30 provided to it.

Score	Gene	Accession Number	Split	MH	ML	NH	NL
0.80	<i>APOA2</i>	X04898	-777	74	0	1	30
0.74	<i>HERV K22 pol</i>	K03498	637	3	71	31	0
0.70	<i>TERT</i>	AF015950	-1610	70	4	0	31
0.70	<i>UMOD</i>	M15881	1119.1	0	74	28	3
0.70	<i>CDH4</i>	L34059	-278	74	0	3	28
0.66	<i>ACTR1A</i>	Z14978	3400.6	3	71	30	1
0.65	<i>MASP1</i>	D17525	-536.6	71	3	1	30
0.65	<i>PTPN21</i>	X79510	1256.1	6	68	31	0

Figure 6. Top voters according to information gain (copied from Figure 2).

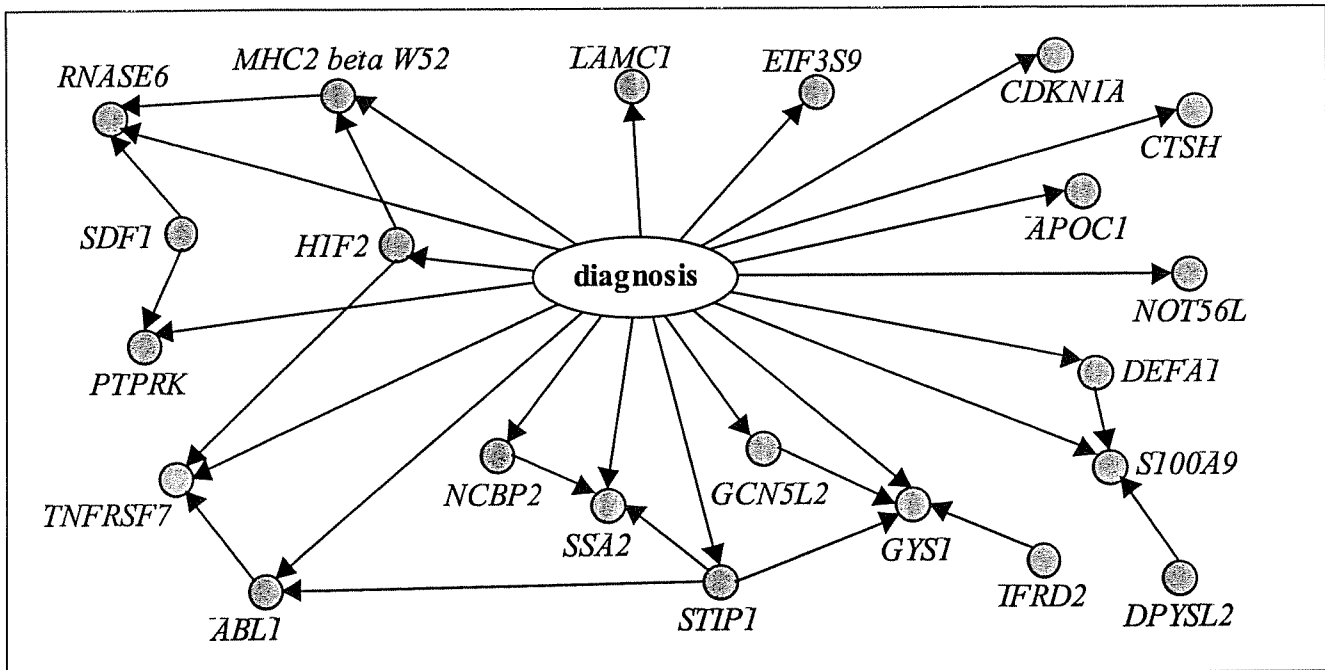


Figure 7. Structure of the Bayes net learned from AC only.

<sup>4</sup> Nevertheless, during boosting the later trees in the boosting run are more complex.



The Bayes net given “AC only” is shown in Figure 7. Note the large number of dependencies among features. In contrast, the Bayes net given AC+AD had *no* dependencies among features it retained: it had the structure of a naïve Bayes model. Nevertheless, the Bayes net learning system did recognize that 11 of the 30 features could be completely removed, because they provided no further information beyond that provided by the other 19. We suspect the reason for no dependencies was because each feature alone was very highly correlated with the class value – see the first 30 features in the appendix. The naïve Bayes net simply provided a weighted voting approach with a selected subset of genes.

Trees, boosted trees, voting, and Bayes nets as used here all ignore the *magnitude* of differences in expression level, or average difference (AD), and look at only the *consistency* of these differences. For example, the top voter in Figure 6 achieves its high rank entirely because all but one of the normal cases have AD values lower than any of the myeloma cases. The magnitude of those differences plays no role. *Lesson 3* is that a measure that considers only consistency works quite well, as evidenced by both voting and Bayes nets achieving 100% cross-validation accuracy given AC+AD (and in fact using AD only). One would think that taking magnitude into account as well might be a good idea, especially given the common use of “fold change” in microarray analysis. SVMs provide the capability to consider both consistency and magnitude; magnitude plays a role because large differences in AD values contribute to a wider margin for an SVM. As a result, it was our expectation prior to the experiments that (1) SVMs would perform better than other methods when given AC+AD, and (2) SVMs would perform better given AC+AD than given AC only. Therefore, it was most surprising that including AD values actually decreased SVM performance, and that all other approaches significantly outperformed SVMs given AC+AD. Hence *lesson 4* is that the ability of SVMs to consider the *magnitude* of difference in expression in addition to the *consistency* of difference appears to yield little or no benefit beyond methods that consider only consistency of the difference.

At the end of the previous section, we stated that the primary goal of the analysis was to gain insights into multiple myeloma. Let’s return to the voting approach and examine what the results tell us. The top eight voters were presented in Figures 2 and 6; the appendix presents the full set of the top 70 voters. All are potentially of great interest. APOA2 is Apolipoprotein II, TERT is telomerase reverse transcriptase, UMOD is the uromodulin Tamm-Horsfall, and CDH4 is cadherin-4. These genes merit special attention given that they all relate to known properties of either cancer genetics in general or myeloma genetics in particular. Note that three of these four have *negative* split points and hence would have been missed had we thrown out negative AD values. *Lesson 5* from these results is that throwing out data based on low or negative average difference values, or absent calls (which would have a similar effect), is a mistake. We make this claim only for Affymetrix technology because that is the technology used here, although it may apply to some extent to other technologies. We also note that Affymetrix software recently has been modified to not produce negative AD values, although the same issue still may arise with low values. While *lesson 5* applies clearly to data generated before 2002, it remains to be seen whether some data should be eliminated when using the new Affymetrix software.

## 5. CONCLUSIONS AND FUTURE DIRECTIONS

This paper has reported experiments comparing leading supervised data mining approaches on gene expression microarray data for multiple myeloma. These comparative experiments provide evidence for the following lessons that may prove useful in other applications of supervised data mining to microarray data for studying diseases.

1. A directly comprehensible model, such as a Bayes net, decision tree, or ensemble of voters, has the advantage of exposing “trivially-accurate genes.” These are genes that provide no new insight but are highly-accurate owing to the nature of the disease or of sample collection. If the data mining goal is to obtain insight, and not just an accurate predictor, then trivially-accurate genes should be removed through consultation with a domain expert.
2. Unweighted voting and Bayes net learning provide accuracies at least as high as the other approaches (Figure 1) and arguably provide much more direct insight than they do.
3. Information gain provides a very different sort of insight than the traditional “fold-change” measure for comparing a gene’s expression measurements across different samples. Information gain looks for a high level of *consistency* in differential expression rather than *large* differences between some samples of different classes. Considering *consistency* alone works surprisingly well.
4. The ability of SVMs to consider the *magnitude* of the difference in expression in addition to the *consistency* of the difference appears to yield little or no benefit beyond methods that consider only the consistency of the difference.
5. Throwing out data based on low or negative average difference values or absent calls is a mistake, at least with data generated using Affymetrix technology before 2002.

An obvious direction for future research is to test these lessons on additional, larger microarray data sets when they become available. As mentioned earlier, the multiple myeloma data set is expected to grow to the order of 500 samples in the next year. To further facilitate comparative experiments, we hope other researchers will make their data sets publicly available, particularly those with sample numbers of a hundred or more. In addition to the need for further comparative experiments, three other future research directions are now becoming obvious as well.

The first of these new directions regards multiple myeloma in particular. A benign plasma cell dyscrasia called MGUS (*monoclonal gammopathy of undetermined significance*) appears to cause expression patterns very similar to multiple myeloma, yet MGUS is harmless unless it progresses to multiple myeloma. (About 1% of all MGUS cases progress to multiple myeloma per year.) It is possible that most myelomas progress from the MGUS condition. Therefore, perhaps an even better way to understand myelomagenesis and to identify critical myeloma specific therapeutic targets would be to compare myeloma vs. MGUS, myeloma vs. normal, and MGUS vs. normal. Toward this aim, we have begun collecting a large panel of MGUS samples. An initial six samples are available now at [www.lambertlab.uams.edu/publicdata.htm](http://www.lambertlab.uams.edu/publicdata.htm).

Second, it is surprising that simple unweighted voting performs as well as Bayes net learning, which as noted in Section 4 can be viewed as a sophisticated weighted voting scheme. This comparison should be carried out on larger data sets when they become available, and also on cases such as MGUS vs. myeloma where distinctions are likely to be more difficult. More generally, it will be interesting to repeat the comparison reported in this paper on larger data sets as they

become available and on MGUS vs. myeloma, or other tasks where distinction is likely to be very difficult.

Third, how will the lessons of this paper change as the underlying technology changes? Affymetrix has just introduced a new method for computing AC and AD that, among other differences, results in fewer negative average difference values. We expect the broad lessons of our study are robust enough to hold across such technological changes, but that will need to be tested. Also, the entire field of gene expression microarrays may lose some ground to emerging techniques in proteomics, where the amount of protein product from a gene is measured directly, rather than measuring the amount of mRNA, which serves as a (noisy) surrogate for the amount of protein. We expect that supervised data mining algorithms can be applied in similar ways to proteomics data, and again that our broad lessons will be applicable, but this applicability also will need to be tested.

#### ACKNOWLEDGEMENTS

DP was supported in part by grants from the University of Wisconsin Medical School, Graduate School, and Comprehensive Cancer Center. DP and MW were supported in part by NSF grant 9987841. JS and BB were supported in part by National Cancer Institute, Bethesda, MD grant CA55819.

#### REFERENCES

- Cheng, J., Hatzis, C., Hayashi, H., Krogel, M.-A., Morishita, S., Page, D., Sese, J. (2002). KDD Cup 2001. *SIGKDD Explorations*, to appear January.
- Friedman, N., Linial, M., Nachman, I., Pe'er, D. (2000). Using Bayesian networks to analyze expression data. In *4th Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, ACM-SIGACT, April 2000.
- Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16, 906-914.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomeld, C.D., Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.
- Pe'er, D., Regev, A., Elidan, G., Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. In *International Conference on Intelligent Systems for Molecular Biology (ISMB) 2001*, 215-224.
- Slonim, D., Tamayo, P., Mesirov, J., Golub, T., Lander, E. (2000). Class prediction and discovery using gene expression data. In *Proc. 4th Annual International Conf. on Computational Molecular Biology (RECOMB)*, 263-272, Tokyo, Japan: Universal Academy Press.
- Zhan, F., Hardin, J., 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., Tricot, G., Crowley, J., Barlogie, B., Shaughnessy, J. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance and normal bone marrow plasma cells. *Blood*, in press.

## APPENDIX

Below are the top 70 (by information gain) clones, used in the AC+AD ensemble (unweighted vote). All are in fact AD values; the split point for each is given. The adjusted p-value gives the probability that a gene would look as one-sided as we see; all other columns are as in Figure 2. To compute the p-value, we used the hypergeometric distribution (for each of the possible extreme tables) which gives the probability of such a configuration under the assumption of no association between the two variables disease status and high/low status. We have adjusted our p-values to account for multiple testing procedures. Therefore, this column gives us the probability that any gene (of the 7129) will show such an extreme configuration. Since our p-values are so low, this gives us strong evidence that the assumption of no association between the two variables is false. And in turn implies that genes which high association between the two groups and the numerical value of the gene may be genes that are of interest in researching the disease.

InfoGain	Gene	Accession Number	Split	MH	ML	NH	NL	Adjusted p-value
0.802422	<i>APOA2</i>	X04898	-777	74	0	1	30	1.363E-21
0.735975	<i>HERVK22 pol</i>	K03498	637	3	71	31	0	1.174E-19
0.704489	<i>TERT</i>	AF015950	-1610	70	4	0	31	1.056E-18
0.701219	<i>UMOD</i>	M15881	1119.1	0	74	28	3	1.364E-18
0.701219	<i>CDH4</i>	L34059	-278	74	0	3	28	1.364E-18
0.664859	<i>ACTRIA</i>	Z14978	3400.6	3	71	30	1	7.853E-18
0.664859	<i>MASPI</i>	D17525	-536.6	71	3	1	30	7.853E-18
0.650059	<i>PTPN21</i>	X79510	1256.1	6	68	31	0	4.950E-17
0.650059	<i>TCEB3</i>	L47345	1451.4	6	68	31	0	4.950E-17
0.63397	<i>SDF1</i>	L36033	2178.6	4	70	30	1	6.783E-17
0.625966	<i>TNFRSF7</i>	M63928	8129	7	67	31	0	2.758E-16
0.625966	<i>UROD</i>	M14016	2718.9	7	67	31	0	2.758E-16
0.625966	<i>KIAA0135</i>	D50925	788.7	7	67	31	0	2.758E-16
0.625966	<i>KIAA0133</i>	D50923	-1517	67	7	0	31	2.758E-16
0.612641	<i>PML</i>	M79463	88.7	71	3	2	29	2.657E-16
0.612641	<i>IFNA4</i>	M27318	415	3	71	29	2	2.657E-16
0.612641	<i>PPP2R5D</i>	L76702	-1601.8	71	3	2	29	2.657E-16
0.606152	<i>H2BFQ</i>	X57985	506.3	69	5	1	30	4.821E-16
0.606152	<i>ABL1</i>	X16416	709	69	5	1	30	4.821E-16
0.606152	<i>DCTD</i>	L39874	3545	69	5	1	30	4.821E-16
0.606152	<i>H2AFO</i>	L19779	7368	69	5	1	30	4.821E-16
0.603481	<i>CNGB1</i>	U58837	173.8	8	66	31	0	1.379E-15
0.603481	<i>ADRA1B</i>	HT4369	994	8	66	31	0	1.379E-15
0.603481	<i>RAD23A</i>	D21235	5044	66	8	0	31	1.379E-15
0.582583	<i>NNT</i>	U40490	-2.1	74	0	6	25	5.819E-15
0.582366	<i>DUSP7</i>	X93921	1088	9	65	31	0	6.282E-15
0.58236	<i>MAPK12</i>	X79483	1814.7	4	70	29	2	2.201E-15
0.580711	<i>H326</i>	U06631	1501	68	6	1	30	2.934E-15
0.580711	<i>APOE</i>	M12529	1044	6	68	30	1	2.934E-15

0.580711	<i>SLC34A1</i>	L13258	732.5	6	68	30	1	2.934E-15
0.562437	<i>H2BFH</i>	Z80780	455.7	64	10	0	31	2.639E-14
0.562437	<i>HIF2</i>	X57129	455	64	10	0	31	2.639E-14
0.562437	<i>CTSH</i>	X16832	3458.7	10	64	31	0	2.639E-14
0.562437	<i>RXRβ</i>	U41068	2555	10	64	31	0	2.639E-14
0.562437	<i>CD81</i>	M33680	9422	10	64	31	0	2.639E-14
0.562437	<i>DSC3</i>	D17427	1172.6	10	64	31	0	2.639E-14
0.561439	<i>ABCβ10</i>	U18237	-284	73	1	5	26	1.137E-14
0.557191	<i>SCAP</i>	D83782	-662	67	7	1	30	1.571E-14
0.555139	<i>ITGB2</i>	X64072	230	5	69	29	2	1.501E-14
0.543549	<i>HRAS</i>	V00574	117.7	63	11	0	31	1.031E-13
0.543549	<i>ITS1</i>	U13369	353	11	63	31	0	1.031E-13
0.543549	<i>KIAA00167</i>	D28589	4334	63	11	0	31	1.031E-13
0.537806	<i>SI00A5</i>	Z18954	2142	4	70	28	3	4.810E-14
0.537806	<i>MAS1</i>	M13150	-433	70	4	3	28	4.810E-14
0.537806	<i>PPY</i>	M11726	1366.6	4	70	28	3	4.810E-14
0.535273	<i>OR1D2</i>	X65857	617.6	8	66	30	1	7.557E-14
0.535273	<i>NARS</i>	U79254	4613	66	8	1	30	7.557E-14
0.535273	<i>ATR</i>	U49844	1128.8	66	8	1	30	7.557E-14
0.535273	<i>FCER1G</i>	M33195	805.2	8	66	30	1	7.557E-14
0.535273	<i>TCN2</i>	L02648	1481	8	66	30	1	7.557E-14
0.530287	<i>ALDOC</i>	X05196	3226	6	68	29	2	8.776E-14
0.530287	<i>GNRH1</i>	X01059	297.8	6	68	29	2	8.776E-14
0.530287	<i>GNL1</i>	HT3404	2757	6	68	29	2	8.776E-14
0.525585	<i>MHC2 beta W52</i>	HT3779	757	12	62	31	0	3.782E-13
0.525585	<i>HML2</i>	D50532	1255	12	62	31	0	3.782E-13
0.515204	<i>RFX2</i>	HT3504	-4294.2	74	0	8	23	7.073E-13
0.514718	<i>H4FL</i>	Z80788	-346	65	9	1	30	3.313E-13
0.514718	<i>ABCC1</i>	L05628	-938.8	65	9	1	30	3.313E-13
0.514718	<i>KIAA0084</i>	D42043	1375.7	9	65	30	1	3.313E-13
0.511192	<i>PSEN1</i>	U40380	3746	5	69	28	3	3.147E-13
0.508447	<i>DPYSL2</i>	U97105	305.2	13	61	31	0	1.309E-12
0.508447	<i>CD79A</i>	U05259	9482.1	13	61	31	0	1.309E-12
0.508447	<i>ITGB2</i>	M15395	146	13	61	31	0	1.309E-12
0.507348	<i>ATIC</i>	D82348	3516	67	7	2	29	4.517E-13
0.497793	<i>anonymous gene</i>	L18972	1737.5	4	70	27	4	7.958E-13
0.495344	<i>NAP1L4</i>	U77456	1745	64	10	1	30	1.339E-12
0.495344	<i>ETV4</i>	U18018	1331	10	64	30	1	1.339E-12
0.492055	<i>MYOIF</i>	X98411	1744.8	14	60	31	0	4.301E-12
0.492055	<i>PLXNB1</i>	X87904	-600.5	60	14	0	31	4.301E-12
0.492055	<i>TXN</i>	X77584	11024	60	14	0	31	4.301E-12

