

*University of Texas, MD Anderson Cancer  
Center*

UT MD Anderson Cancer Center Department of Biostatistics  
Working Paper Series

---

*Year 2005*

*Paper 5*

---

An Empirical Study of Univariate and  
GA-Based Feature Selection in Binary  
Classification with Microarray Data

Michael L. LeCocke\*

Kenneth Hess†

\*Rice University, mlecocke@stat.rice.edu

†Department of Biostatistics and Applied Mathematics, UT MD Anderson Cancer Center,  
khess@mdanderson.org

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/mdandersonbiostat/paper5>

Copyright ©2005 by the authors.

# An Empirical Study of Univariate and GA-Based Feature Selection in Binary Classification with Microarray Data

Michael L. Lecoche and Kenneth Hess

## Abstract

Motivation: Feature subset selection is an important aspect of performing binary classification using gene expression data. Once feature subsets are obtained, there is the need to evaluate the various models that are formed. This paper considers both univariate- and multivariate-based feature selection approaches for the problem of binary classification with microarray data. In considering the more sophisticated multivariate approach, the idea is to determine whether it leads to better misclassification error rates because of the greater potential to consider jointly significant subsets of genes than would an approach combining individually predictive genes selected by a univariate approach. Further, we wish to see if the multivariate approaches can perform well without overfitting the data.

Results: An empirical study is presented, in which a 10-fold cross-validation is applied externally to both a univariate-based and two multivariate- (genetic algorithm (GA)-) based feature selection processes. These procedures are applied with respect to three supervised learning algorithms and six published two-class microarray datasets. We find that although the multivariate feature selection approaches in general may have more potential to select jointly significant combinations of genes than would the simpler univariate approach, the 10-fold external cross-validation misclassification error rates between the two approaches for all classifiers and across all subset sizes were actually very comparable. Considering all datasets, learning algorithms, and subset sizes together, the average 10-fold external cross-validation error rates for the univariate-, single-stage GA-, and two-stage GA-based processes are 14.2%, 14.6%, and 14.2%, respectively. Further, we find that a more sophisticated two-stage GA approach did not demonstrate a

significant advantage over a 1-stage approach. We also find that the univariate approach had higher optimism bias and lower selection bias compared to both GA approaches. Finally, considering all datasets, learning algorithms, and subset sizes together, we find that the optimism bias estimates from the GA analyses were half that of the univariate approach, but the selection bias estimates from the GA analyses were 2.5 times that of the univariate results. This higher selection bias suggests that selecting genes in multivariate models using a GA may be more likely to select spurious genes than would be the case with a univariate-based approach.

# An Empirical Study of Univariate and GA-Based Feature Selection in Binary Classification with Microarray Data

Mike Lecocke\* and Kenneth Hess<sup>†</sup>

2nd March 2005

## Abstract

**Motivation:** Feature subset selection is an important aspect of performing binary classification using gene expression data. Once feature subsets are obtained, there is the need to evaluate the various models that are formed. This paper considers both univariate- and multivariate-based feature selection approaches for the problem of binary classification with microarray data. In considering the more sophisticated multivariate approach, the idea is to determine whether it leads to better misclassification error rates because of the greater potential to consider jointly significant subsets of genes than would an approach combining individually predictive genes selected by a univariate approach. Further, we wish to see if the multivariate approaches can perform well without overfitting the data. **Results:** An empirical study is presented, in which a 10-fold cross-validation is applied externally to both a univariate-based and two multivariate- (genetic algorithm (GA)-) based feature selection processes. These procedures are applied with respect to three supervised learning algorithms and six published two-class microarray datasets. We find that although the multivariate feature selection approaches in general may have more potential to select jointly significant combinations of genes than would the simpler univariate approach, the 10-fold external cross-validation misclassification error rates between the two approaches for all classifiers and across all subset sizes were actually very comparable. Considering all datasets, learning algorithms, and subset sizes together, the average 10-fold external cross-validation error rates for the univariate-, single-stage GA-, and two-stage GA-based processes are 14.2%, 14.6%, and 14.2%, respectively. Further, we find that a more sophisticated two-stage GA approach did not demonstrate a significant advantage over a 1-stage approach. We also find that the univariate approach had higher optimism bias and lower selection bias compared to both GA approaches. Finally, considering all datasets, learning algorithms, and subset sizes together, we find that the optimism bias estimates from the GA analyses were half that of the univariate approach, but the selection bias estimates from the GA analyses were 2.5 times that of the univariate results. This higher selection bias suggests that selecting genes in multivariate models using a GA may be more likely to select spurious genes than would be the case with a univariate-based approach. **Availability:** Datasets and *genalg* software are available from the authors upon request. **Contact:** mlecocke@stat.rice.edu and khess@mdanderson.org

---

\*Department of Statistics, Rice University, Houston, Texas 77005

<sup>†</sup>Department of Biostatistics and Applied Mathematics, UT MD Anderson Cancer Center, Houston, Texas 77030

<sup>‡</sup>To whom correspondence should be addressed.

# 1 Introduction

## 1.1 Motivation

DNA microarray technology has greatly influenced the realms of biomedical research, with the hopes of significantly impacting the diagnosis and treatment of diseases. Microarrays have the ability to measure the expression levels of thousands of genes simultaneously. They measure how much a given type of messenger RNA (mRNA) is present in a tissue sample at a given moment. The wealth of gene expression data that has become available for microarray data analysis has introduced a number of statistical questions to tackle. Some questions are targeted towards various preprocessing stages of a microarray experiment such as RNA hybridization to arrays, image processing, and normalization, while others are geared towards assessing differential expression and identifying profiles for classification and prediction. Within the framework of tumor classification, the types of goals that have been explored include discovering or identifying previously unknown tumor classes, classifying tumors into previously known classes, and identifying “marker genes” that characterize various tumor classes.

In standard discrimination problems, the number of training observations  $N$  is usually much larger than the number of feature variables  $p$ . However, in the context of microarrays, the number of tissue samples  $N$  is usually between 10 and 100, significantly smaller than the thousands of genes considered in a typical microarray analysis. This presents a number of problems to a prediction rule in a discriminant analysis setting. The prediction rule may not even be able to be formed using *all*  $p$  variables, as is the case with Fisher’s linear discriminant analysis (Ambroise and McLachlan, 2002). Further, even if all the variables could be taken into account in forming the prediction rule, some of them may possess minimal (individual) discriminatory power, potentially inhibiting the performance of the prediction rule when applied to new (unclassified) tumors. Ultimately, with a collection of genes that have high discriminatory power, an effective prediction rule can be developed based on these genes and used to allocate subsequent unclassified tissue samples as one of two classes such as cancer and normal, or perhaps as one of two subtypes of a particular cancer. Discovery of key genes needed for accurate prediction could pave the way to better understand class differences at the biological level, which could hopefully provide more information about how to select important biomarkers to be used in the development of clinical trials for predicting outcome and various forms of treatment.

## 1.2 Supervised Learning

Gene expression data for  $p$  genes over each of  $N$  mRNA samples can be expressed as an  $N \times p$  matrix  $X = (x_{ij})$ ,  $i = 1, \dots, N$  and  $j = 1, \dots, p$ . Each value  $x_{ij}$  corresponds to the expression level for gene  $j$  in sample  $i$ . Each sample would have associated with it a gene expression profile  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in R^p$ , along with its class designation  $y_i$ . This variable serves as the response, or dependent variable, and can take one of two predefined values from  $\{0, 1\}$ . Using the observed measurements  $X$ , a classifier for two classes is thus a mapping  $G : R^p \rightarrow \{0, 1\}$ , where  $G(\mathbf{x})$  denotes

the predicted class,  $\hat{y} = c$ ,  $c \in \{0, 1\}$ , for a sample with feature vector  $\mathbf{x}$ .

The samples already known to belong to certain classes,  $\mathcal{L} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_{n_L}, y_{n_L})\}$ , constitute the training (or learning) set. The training set is used to construct a classifier, which is then used to predict the classes of an independent set of samples (the test set  $\mathcal{T} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_T}\}$ ). This way, the class  $\hat{y}_i$ , ( $i = 1, 2, \dots, n_T$ ) predictions for each test set expression profile  $\mathbf{x}_i$  can be made. Of course, with the true classes  $y_i$ , ( $i = 1, 2, \dots, n_T$ ) of the test set known, a misclassification error rate (*MER*) can then be computed.

### 1.3 Feature Subset Selection

In general, feature (variable) selection is an important aspect of classification problems, since the features selected are used to build the classifier. Careful consideration should be given to the problem of feature subset selection with high-dimensional data. With respect to microarray data, this of course amounts to reducing the number of genes used to construct a prediction rule for a given learning algorithm. There are several reasons for performing feature reduction. Whereas two variables could be considered good predictors individually, there could be little to gain by combining the two variables together in a feature vector. It has been reported that as model complexity is increased with more genes added to a given model, the proportion of training samples (tissues) misclassified may decrease, but the misclassification rate of new samples (generalization error) would eventually begin to increase; this latter effect being the product of overfitting the model with the training data (Hastie et al., 2001; McLachlan, 1992; Theodoridis, 1999; Xing, 2002; Xiong et al., 2001). Further, if another technology will be used to implement the gene classifier in practice (e.g., to develop diagnostic assays for selected subsets of genes), the cost incurred is often a function of the number of genes. Finally, there is the obvious issue of increased computational cost and complexity as more and more features are included in a model.

### 1.4 Assessing the Performance of a Prediction Rule: Cross-validation

One approach to estimate the error rate of the prediction rule would be to apply the rule to a “held-out” test set randomly selected from among the training set samples. As an alternative to the “hold-out” approach, cross-validation (CV) is very often used, especially when one does not have the luxury of withholding part of a dataset as an independent test set and possibly even another part as a validation set (usually the case with microarray data). Further, the repeatability of results on new data can be assessed with this approach. In general, all CV approaches can fall under the “ $K$ -fold CV” heading. Here, the training set of samples is divided into  $K$  non-overlapping subsets of (roughly) the same size. One of the  $K$  subsets is “held-out” for testing, the prediction rule is trained on the remaining  $K - 1$  subsets, and an estimate of the error rate can then be obtained from applying each stage’s prediction rule to its corresponding test set. This process repeats  $K$  times, such that each subset is treated once as the test set, and the average of the resulting  $K$  error rate estimates forms the  $K$ -fold CV error rate. The whole  $K$ -fold CV process could be repeated multiple times, using different partitions of the data each run and averaging the results, to obtain

more reliable estimates. At the expense of increased computation cost, repeated- (10-) run CV has been recommended as the procedure of choice for assessing predictive accuracy of the classification of microarray data (Braga-Neto and Dougherty, 2004; Kohavi, 1995).

With microarray classification problems, the practice has generally been to perform CV only on the classifier construction process, not taking into account feature selection. The feature selection process is applied to the entire set of data. This approach to CV is referred to as “internal” cross-validation (Ambroise and McLachlan, 2002; Dudoit and Fridlyand, 2003; McLachlan, 1992). Although the intention of CV is to provide accurate estimates of classification error rates, using CV in this manner means that any inference would be made with respect to the classifier building process only. Leaving out feature selection from the cross-validation process will inevitably lead to selection bias, as the feature selection would not be based on the particular training set for each CV run. Hence, overly optimistic error rates would be obtained. To prevent this selection bias from occurring, an “external” cross-validation process (Ambroise and McLachlan, 2002; Dudoit and Fridlyand, 2003; McLachlan, 1992) should be implemented following the feature selection at each CV stage. That is, the feature selection is performed based only on those samples set aside as training samples at each stage of the CV process, external to the test samples at each stage.

Careful consideration should be given to the feature subset selection problem when constructing a prediction rule within the framework of a supervised classification problem. This paper focuses on the implementation of external cross-validation to assess the predictive accuracy of various classification rules. Empirical results based on both univariate and multivariate feature selection procedures, for three different learning algorithms and six published microarray datasets, are presented in this paper. Of particular interest is whether the results based on the more sophisticated multivariate feature selection scheme offer a significant advantage in terms of lower error rates than results based on univariate-based feature selection.

## 2 Methods

### 2.1 Supervised learning methods

In this study, three well known and widely used choices of supervised learning algorithms were implemented: support vector machines (SVM's), DLDA, and  $k$ -NN ( $k=3$  in this study). For more details on each of these classifiers, the reader should refer to Dudoit and Fridlyand (2003) and Dudoit et al. (2000).

### 2.2 Feature subset selection

#### Univariate Feature Selection

A univariate-based means of feature subset selection was used to perform gene selection. Rank-based, unequal variance T-tests were performed on each of the genes from the designated training sets of samples among each of the six datasets. In each training set, this resulted in an ordered

list of “top genes”. This list, ordered according to increasing p-value, was then used in generating various “top gene subset size” models. To obtain a Monte Carlo type of estimate of the 10-fold external CV misclassification error rates, the standard 10-fold process is run 10 separate times, and the average of the resulting ten 10-fold CV MER estimates is recorded. For a given dataset, for each of the classifiers implemented for a given dataset, the same ten training and test set partitions for a given iteration were used to maintain consistency in interpreting the repeated-run 10-fold CV results.

## GA and “GA-GA” Feature Selection

A multivariate-based means of feature subset selection, the genetic algorithm (GA), was used to perform gene selection. For more details on the GA in general, the reader is referred to Freitas (2001); Holland (1975); Mitchell (1997). Both a single-stage and a two-stage GA feature selection process was implemented. Repeated (10) runs of the GA process are used to provide more stable estimates of the 10-fold CV error rates. The repeated runs are implemented via the *genalg* software, which provides the user with the opportunity to run the entire GA multiple (10) times for each training set of data (in the case of external CV, the GA is applied 10 times on each of the 10 training subsets of samples, corresponding to each stage of the 10-fold CV process). It should be noted that to evaluate each candidate  $d$ -gene subset, Mahalanobis distance is used as the objective function for sample classification based on the  $d$  genes. For more information on the *genalg* software, the reader should refer to Baggerly et al. (2003).

For the single-stage GA approach, the GA considers all  $p$  genes of each dataset. The number of  $d$ -gene solutions (“chromosomes”) selected by each implementation of the single-stage GA is 1000, so over 10 iterations, a “superpopulation” of 10000 candidate solutions are obtained. The number of generations to run for each iteration of the GA was set to 250, which we found was large enough to ensure convergence of the 1000 solutions. For the two-stage approach (“GA-GA”), the first stage GA takes into account all  $p$  genes, while in the second stage, the algorithm is applied to a reduced set of genes based on the initial GA’s selection results for each training subset of the datasets. For each training set of data, for a given subset size  $d$ , the union of all genes selected among the final generation’s population of 1000 solutions of  $d$  genes from the first stage of GA, for all 10 iterations of the GA, is retained as the reduced gene pool to use for the second stage of GA. That is, the second stage’s GA procedure uses as its initial gene pool all genes that appeared at least once among the “superpopulation” of 10000 solutions obtained from the initial GA stage. This way, genes that may appear in a small proportion of the 1000 solutions of any iteration, but appear in multiple iterations of the GA for a given set of training data, have a better chance of being considered for use in building classifiers based on a given gene subset size and an appropriate learning method. Thus, the idea behind the second implementation of the GA would be to attempt to select the ‘best of the best’ genes from a given training dataset. It should be noted that for the second stage GA, the number of generations remained at 250, but the number of  $d$ -gene solutions selected by each implementation of the GA was reduced to 500, since the initial gene pool was

reduced considerably.

### 3 Datasets

The following datasets are analyzed in this paper, all of which are from Affymetrix microarrays (Affymetrix, 1999, 2000a,b, 2002). The only preprocessing that was done on each dataset was to standardize the arrays such that they each have zero mean and unit variance (an approach also used in the comparative gene expression classification study of Dudoit et al. (2000)). Standardization of microarray data in this manner achieves a location and scale normalization of the arrays. This was done to ensure that all the arrays of a given dataset were independent of the particular technology used. That is, the standardization was done to take into account the effect of processing artifacts, such as longer hybridization periods, less post-hybridization washing of the arrays, and greater laser power, to name a few. This way, for a given dataset, the values corresponding to individual genes can be compared directly from one array to another. Further, it's been shown that this type of normalization has been effective in preventing the expression values of one array from dominating the average expression measures across arrays (Yang et al., 2001). Currently there is no universally accepted means of normalizing microarray data.

#### **Alon et al. (1999) colon cancer dataset**

This dataset consists of gene expression levels measured from Affymetrix oligonucleotide arrays (HU6000; quantization software uncertain) for 2000 genes across 62 samples. The binary classes used for analysis are normal (22 samples) and tumor (40 samples). As discussed in Li et al. Li et al. (2001), five colon samples previously identified as being contaminated were omitted (N34, N36, T30, T33, and T36), leaving the total sample size for analysis at 57. See Alon et al. (1999) for more details on this dataset.

#### **Golub et al. (1999) leukemia dataset**

This dataset consists of gene expression levels from Affymetrix chips (HuGeneFl). The oligonucleotide arrays have 7129 probe sets over 72 samples. The binary classes used for analysis are acute myeloid leukemia (AML; 25 samples) and acute lymphoblastic leukemia (ALL; 47 samples). See Golub et al. (1999) for more details on this dataset.

#### **Nutt et al. (2003) brain cancer dataset**

This dataset consists of gene expression levels measured from Affymetrix high-density oligonucleotide chips (U95Av2) using the GeneChip software. Each array contains 12625 probe sets over 50 samples. The binary classes used for analysis are glioblastoma (28 samples) and anaplastic oligodendroglioma (22 samples). The downloaded raw expression values were previously normalized by

Research Archive

linear scaling such that the mean array intensity for active (“present”) genes was identical for all the scans. See Nutt et al. (2003) for more details on this dataset.

#### **Pomeroy et al. (2002) brain cancer dataset**

This dataset consists of gene expression levels measured from Affymetrix high-density oligonucleotide chips (HuGeneFl) using the GeneChip software. Each chip contains 7129 probe sets. To facilitate the binary classification framework, dataset ‘A2’ from the project website was used, in which 60 medulloblastoma (MD) samples formed one class and the remaining 30 samples classified as “Other” for the second class (Note: of these 30, there were 10 malignant gliomas (MG), 10 atypical teratoid/rhaboid tumor (AT/RT), 6 supratentorial primitive neuroectodermal tumors (PNET), and 4 normal cerebellum samples). See Pomeroy et al. (2002) for more details on this dataset.

#### **Shipp et al. (2002) lymphoma dataset**

This dataset consists of gene expression levels measured from Affymetrix chips (HuGeneFL) using the GeneChip software. Each oligonucleotide array contained 7129 probe sets over 77 samples. The two classes used for analysis are diffuse large B-cell lymphoma (DLBCL; 58 samples) and follicular lymphoma (FL; 19 samples). See Shipp et al. (2002) for more details on this dataset.

#### **Singh et al. (2002) prostate cancer dataset**

This dataset consists of gene expression levels measured from Affymetrix chips (HU95Av2) using the GeneChip software. The number of arrays available for analysis was 102, with each containing 12600 probe sets. The two classes used for analysis are normal (50 samples) and prostate cancer (52 samples). See Singh et al. (2002) for more details on this dataset.

## **4 Results**

### **4.1 Gene Selection: Univariate vs. Multivariate**

First of all, to get an idea of how effective the two GA-based feature selection processes were at selecting genes that would otherwise not be considered “top genes” from a univariate screening procedure, all three feature selection approaches were implemented in a resubstitution setting, in which all samples were used for each dataset. Table 1 provides a breakdown of the percentage of genes, relative to each gene subset size, among each of the GA-based feature selection processes that were not even among the top 100 univariately significant genes, for all six datasets.

Table 1: % of Genes of Each Subset Size Not Within Top 100 Univariately Significant Genes List

	Alon		Golub		Nutt		Pomeroy		Shipp		Singh	
Size	GA	GAGA	GA	GAGA	GA	GAGA	GA	GAGA	GA	GAGA	GA	GAGA
<b>1</b>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
<b>2</b>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	0.0	0.0	0.0
<b>3</b>	66.7	66.7	33.3	33.3	33.3	66.7	33.3	100.0	100.0	66.7	0.0	33.3
<b>4</b>	50.0	50.0	50.0	25.0	75.0	75.0	50.0	50.0	75.0	100.0	50.0	50.0
<b>5</b>	20.0	60.0	40.0	20.0	100.0	80.0	80.0	60.0	60.0	80.0	60.0	60.0
<b>10</b>	70.0	70.0	60.0	60.0	70.0	90.0	70.0	80.0	70.0	80.0	60.0	60.0
<b>15</b>	73.3	66.7	60.0	80.0	93.3	86.7	66.7	60.0	93.3	80.0	73.3	73.3
<b>20</b>	75.0	75.0	70.0	80.0	95.0	95.0	85.0	80.0	85.0	85.0	85.0	95.0
<b>25</b>	84.0	84.0	76.0	72.0	96.0	92.0	76.0	84.0	96.0	96.0	84.0	76.0

From Table 1, one can note that for all datasets except the Golub one in the case of the 2-stage GA process, for gene subset sizes of 4 or more, both the single-stage and the two-stage GA approaches generated final gene subsets in which the majority of the genes of each subset size were not among the top 100 from that dataset’s univariately significant genes. This finding was especially true for subset sizes of 10, 15, 20, and 25. Whether or not this translates to much improved misclassification error rates, however, is the topic of Section 4.3.

## 4.2 Optimism Bias, Selection Bias, and Total Bias

In a previous study, both 10-fold external and internal CV was performed, using both univariate- and GA-based feature subset selection. The idea was to consider the problem of how best to evaluate prediction rules formed from models such that the effects of optimism bias, selection bias, and “total” bias are properly taken into account, where these bias estimates are defined as follows:

$$\widehat{ob} = MER_{IntCV} - MER_{Resub} \quad (1)$$

$$\widehat{sb} = MER_{ExtCV} - MER_{IntCV} \quad (2)$$

$$\widehat{tb} = \widehat{sb} + \widehat{ob} \quad (3)$$

$$= MER_{ExtCV} - MER_{Resub} \quad (4)$$

In these analyses, the empirical results were based on the same six datasets and same three learning algorithms as this study. Considering all datasets, learning algorithms, and gene subset sizes together, we found that for the results based on univariate feature selection, the average

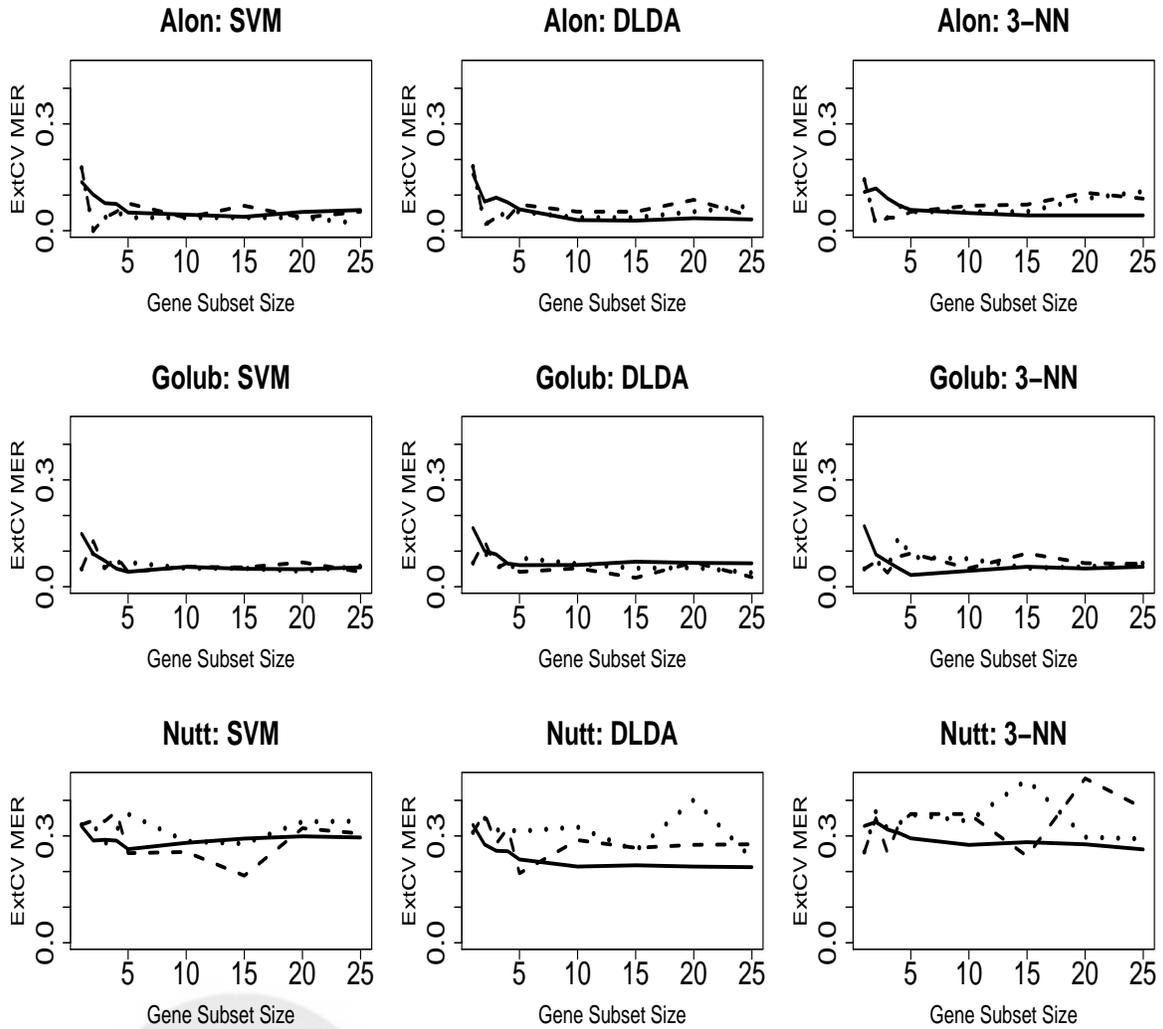
optimism, selection, and total bias estimates were only 4%, 3%, and 7%, respectively. The average optimism, selection, and total bias estimates for the GA-based results were 2%, 8%, and 10%, respectively, and those for the “GA-GA”-based results were 2%, 7.5%, and 10%, respectively. Hence, the optimism bias, incurred from using the same data to both train the classifier and estimate the classifier’s performance, from each of the GA-based analyses was half that of the univariate-based results. However, the selection bias, incurred from using the same data to both select the gene subsets and estimate the classifier’s performance, from each of the GA-based analyses was 2.5 times that of the univariate-based results.

### 4.3 External Cross-Validation

The 10-run external CV results for each dataset and classifier combination, across a number of gene subset sizes, are shown in Figures 1 and 2 below. Within each graph, three curves are shown, corresponding to the external CV error rates based on univariate, GA, and “GA-GA” feature selection.



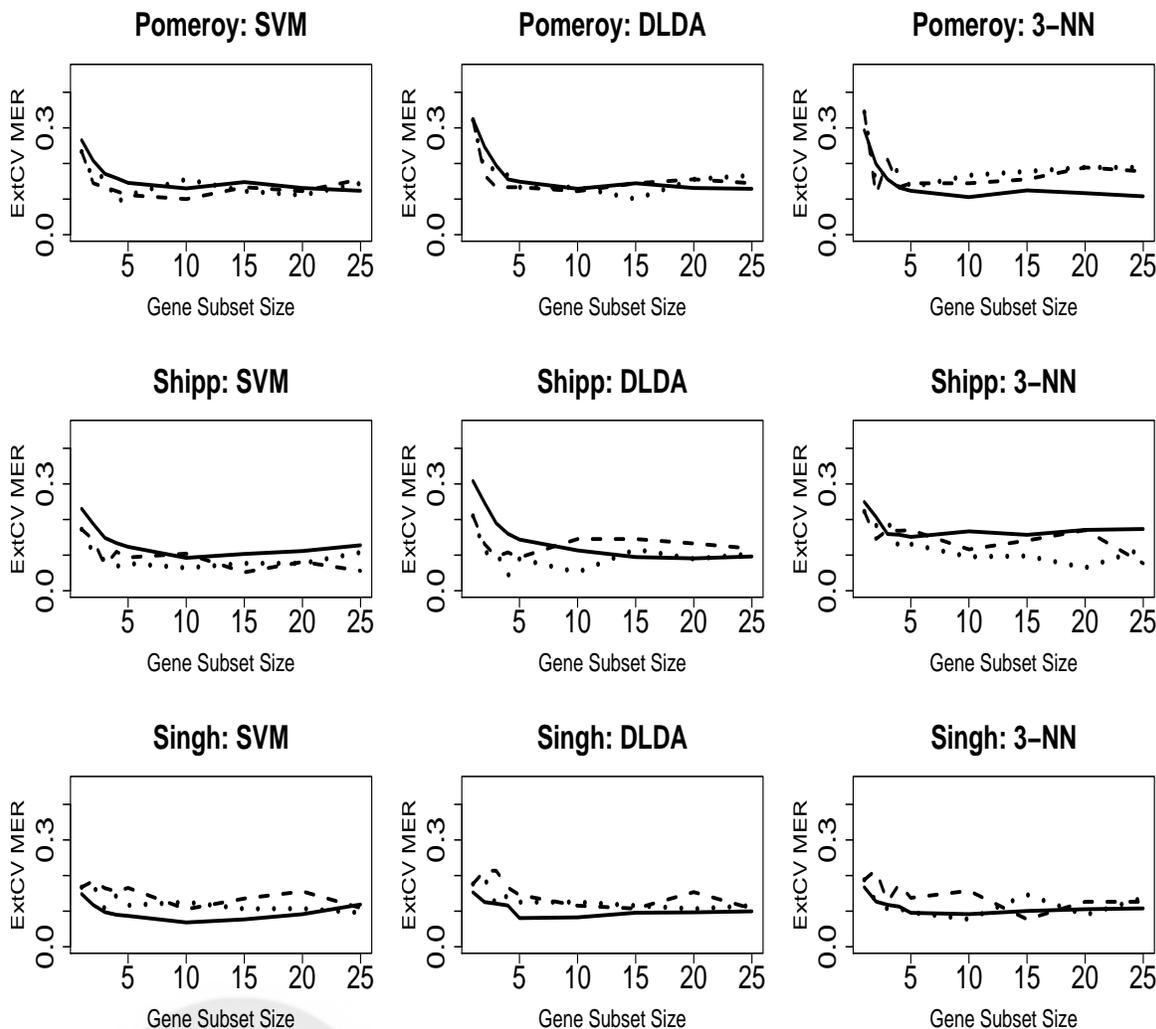
Figure 1: 10-Fold External CV; MER vs. Gene Subset Size: Alon, Golub, Nutt Datasets



(a) Solid: Univariate FSS, Dashed: GA FSS, Dotted: GA-GA FSS



Figure 2: 10-Fold External CV; MER vs. Gene Subset Size: Pomeroy, Shipp, Singh Datasets



(a) Solid: Univariate FSS, Dashed: GA FSS, Dotted: GA-GA FSS

Several observations should be noted from Figures 1 and 2. First, in comparing the datasets, in general the Alon and Golub datasets had the lowest MER values for all the classifiers across the gene subset sizes, followed by the Pomeroy, Shipp, Singh, and finally Nutt datasets. Among the three learning algorithms, no single one emerged across all datasets as the best in terms of lowest average error rates. Among the two GA-based feature selection approaches, the more complicated two-stage one did not offer a significant advantage in terms of lower average error rates over the simpler single-stage one. Further, it should be noted that for all datasets and classifiers, although

Table 2: 10-Fold External CV MER's Across All Gene Set Sizes

Dataset	Classifier	Univariate: Mean (SD)	GA: Mean (SD)	GAGA: Mean (SD)
Alon	SVM	0.071 (0.031)	0.060 (0.049)	0.048 (0.050)
	DLDA	0.066 (0.043)	0.064 (0.048)	0.060 (0.048)
	3-NN	0.070 (0.030)	0.070 (0.039)	0.070 (0.039)
Golub	SVM	0.068 (0.034)	0.064 (0.028)	0.060 (0.016)
	DLDA	0.083 (0.034)	0.059 (0.031)	0.062 (0.018)
	3-NN	0.069 (0.041)	0.068 (0.019)	0.072 (0.030)
Nutt	SVM	0.291 (0.017)	0.301 (0.058)	0.318 (0.031)
	DLDA	0.246 (0.040)	0.286 (0.046)	0.314 (0.048)
	3-NN	0.298 (0.027)	0.333 (0.072)	0.336 (0.058)
Pomeroy	SVM	0.165 (0.046)	0.140 (0.039)	0.143 (0.042)
	DLDA	0.178 (0.067)	0.162 (0.062)	0.172 (0.063)
	3-NN	0.151 (0.061)	0.179 (0.069)	0.190 (0.061)
Shipp	SVM	0.140 (0.044)	0.098 (0.040)	0.093 (0.034)
	DLDA	0.160 (0.076)	0.131 (0.035)	0.102 (0.048)
	3-NN	0.177 (0.032)	0.153 (0.041)	0.137 (0.053)
Singh	SVM	0.099 (0.025)	0.149 (0.027)	0.127 (0.029)
	DLDA	0.108 (0.023)	0.155 (0.041)	0.135 (0.030)
	3-NN	0.114 (0.023)	0.148 (0.041)	0.119 (0.034)
Alon Grand Avg		0.069 (0.035)	0.065 (0.046)	0.059 (0.046)
Golub Grand Avg		0.074 (0.036)	0.064 (0.026)	0.065 (0.021)
Nutt Grand Avg		0.278 (0.028)	0.307 (0.058)	0.323 (0.046)
Pomeroy Grand Avg		0.165 (0.058)	0.160 (0.057)	0.168 (0.055)
Shipp Grand Avg		0.159 (0.051)	0.128 (0.039)	0.110 (0.045)
Singh Grand Avg		0.107 (0.024)	0.151 (0.037)	0.127 (0.031)
All Data Grand Avg		0.142 (0.038)	0.146 (0.044)	0.142 (0.041)

the GA feature selection methods may have greater potential to select combinations of genes that are jointly discriminatory than would an approach that combines individually predictive genes, there was not a clear advantage for either of the more sophisticated GA-based methods over the univariate-based method.

Concluding this section is a table summarizing the means and standard deviations of the 10-fold external CV misclassification error rates averaged across gene subset sizes, for each dataset and classifier combination, for each of the three approaches to feature subset selection. The empirical grand means and standard deviations for each dataset across all subset sizes and classifiers, as well as the empirical grand means and standard deviations across all datasets, subset sizes, and classifiers, for each of the three feature selection methods, are provided in the last seven rows of Table 2. Overall, considering all datasets, classifiers, and gene subset sizes together, the average 10-fold external CV error rates are very comparable – 14.2%, 14.6%, and 14.2%, respectively. It should be noted that if the Nutt data were excluded, these averages become 11.5%, 11.2%, and 10.6%, respectively.

## 5 Discussion

Dudoit et al. (2000) provide an in-depth comparative study of several supervised learning methods for tumor classification based on filtered sets of genes from several published microarray datasets.

The learning algorithms used in their study were linear discriminant analysis (LDA), diagonal LDA (DLDA), quadratic LDA (DQDA), classification trees, and  $k$ -NN. The gene selection method implemented was to select the  $p$  genes with largest ratio of between to within-sum-of-squares. In this study, repeated (150) runs of training/test set partitions were performed, with feature selection done only on each training set. The ratio of training to test set samples was 2:1. No cross-validation study was performed. More recently, Dudoit and Fridlyand (2003) applied univariate screening with both a simple t-test and a rank-based t-test (Wilcoxon Test) to analyze a couple of published two-class microarray datasets. The classification schemes they used were  $k$ -NN, DLDA, boosting with trees, random forests, and SVM's. In this study, they applied external and internal CV, but only using leave-one-out (LOO) CV. For both studies, the general conclusion was that the simpler classification methods such as DLDA performed better than the more complicated ones such as  $k$ -NN and SVM. In the more recent study, the authors found that the internal LOO CV led to misclassification error rates that were severely biased downward compared to the external CV approach.

In the study of Xiong et al. (2001), there were two multivariate feature selection methods used – a Monte Carlo method and a stepwise forward selection method. Three binary classification datasets were used in this study: The results are based only on using Fisher's LDA as the classification mechanism. Also, these authors used a "holdout" method to evaluate the performance of the selected genes, dividing the data into a training and test set in the following proportions: (50%, 50%), (68% and 32%), and (95% and 5%), respectively, and then averaged the results of 200 runs of each of these approaches. In this study, it was found that both multivariate methods performed better than the univariate-based T-test and prediction strength statistic (Golub et al., 1999) methods. However, the accuracy of classification criterion for forming gene subsets was based on the total collection of tissue samples, which allows for the presence of selection bias. In addition, the only subset sizes considered in this study were 1, 2, and 3.

External CV was implemented on two published datasets in the study of Ambroise and McLachlan (2002). The samples were randomly divided into 50 different training and test set partitions, with the CV performed only on the training data. They used two schemes for multivariate feature selection and classification – backward selection with SVM and forward selection with LDA. No univariate-based approach to perform the feature selection was implemented. They considered the effect of selection bias by performing external 10-fold CV and internal LOO CV. Unfortunately, no internal 10-fold and external LOO results were provided in the study. The average values of the error rate estimates across the multiple runs were obtained for both approaches for each dataset. They found that the internal LOO CV led to overly optimistic error rates compared to the external 10-fold CV process, for both classification schemes and datasets.

With respect to GA's, there has been some work with them in the context of classification of microarray data. In the study of Li et al. (2001), a GA was implemented on a training set from the colon cancer data of Alon et al. (1999) to select a number of 50-gene subsets that discriminate between normal and tumor tissue samples. 3-NN was used as the objective function within the

GA. Once a large enough number (6348) of these 50-gene solutions were obtained, the solutions were pooled together such that a frequency count could be performed. That is, using all the genes comprising these solutions, a ranked list of the most often selected genes was formed. From this list, the top  $D$  genes were used to classify test set samples using the 3-NN classifier. The authors found that the test set predictions stabilized when as few as 25 and up to 110 top genes were used. As more top genes were included, the number of unclassifiable samples increased. The same GA/ $k$ -NN method was used for training of 38 samples from the leukemia dataset of Golub et al. (1999). Using the top 50 most frequently selected genes among the 50-gene subsets generated by the GA/3-NN method, they correctly classified 33 of the 34 test samples.

The current research builds on the findings of the studies by Xiong et al. (2001), Ambrose and McLachlan (2002), and Dudoit and Fridlyand (2003), in the sense that 10-fold external CV was implemented to take into account selection bias when estimating the misclassification error of a classification rule based on microarray data. However, in this research, the external CV is performed in conjunction with both univariate- and multivariate GA-based feature selection to assess the performance of various prediction rules across multiple two-class microarray datasets. The current research also extends on the analyses of Li et al. (2001) in that the GA is actually incorporated into each stage of a 10-fold (external) CV procedure, rather than have the data split into a training and test set. It also builds on these results in that once subsets of genes are selected by the GA (single-stage approach), they are not then re-pooled together such that the final gene subsets used for modeling are actually selected from a new pool of genes based on frequency of selection among the final gene subsets selected by the GA procedure – ultimately an inherently univariate notion of feature selection. Instead, in this research the GA-selected gene subsets are left alone and not further formed based on frequency of selection among all subsets. Also, a simpler and less computationally intensive objective function than  $k$ -NN, Mahalanobis distance, is employed in the GA algorithm implemented in this research.

Repeated- (10-) run 10-fold external cross-validation was applied to each of six datasets, using each of three different learning algorithms. With the external CV approach, the feature selection was performed at each stage of the CV process, based only on the training set partitions of each stage and hence external to the test sets used to evaluate the models. The average error rates across all the classifiers and gene set sizes were very comparable among the univariate and two multivariate feature selection approaches, as they were all 14%. In terms of datasets, only the Nutt dataset had noticeably higher error rates across classifiers and subset sizes than those of the other datasets, as they were, on average, above 27% for each of the three feature selection approaches. Ultimately, the misclassification rates did not vary significantly by dataset, for five of the six datasets at least, suggesting that these results should generalize well to other clinical microarray datasets. The same generalization ability should hold with respect to classifiers, since the three classifiers used function in different ways, and since there is no clear reason to suspect that the results are connected to the method of classification.

For each of the six datasets used in this study, we have shown that although a multivariate-

based approach to feature subset selection may have greater potential to select combinations of genes that are jointly discriminatory than would a method that combines individually predictive genes, there was no clear advantage of the more computationally intensive GA approaches over the simpler univariate ones in estimating the prediction error for a classification rule constructed from a selected subset of genes from microarrays. Considering all classifiers, subset sizes, and learning algorithms, we found that the optimism bias estimates from the GA analyses were half that of the univariate approach, while the selection bias estimates from the GA analyses were 2.5 times that of the univariate results. This higher selection bias suggests that selecting genes in multivariate models using a GA may be more likely to select spurious genes than would be the case with a univariate-based approach. This finding makes sense in that since the selection bias measures the bias in the estimate of CV prediction error due to feature selection, one would suspect that it would be higher with the multivariate feature selection approach since this approach searches a much higher dimensional model space when finding the features. Thus, with the multivariate feature selection approach, it would naturally be more possible to include spurious genes in candidate models, which can be seen as overfitting the data. That is, considering the notion of overfitting to mean that too much flexibility is allowed in the model space, such that the models trace the data too closely, likely select spurious features of the given data set, and hence do not accurately generalize to independent validation data, it would be safe to say that the GA-based methods tend to overfit the data. Ultimately, whether a univariate or a GA-based feature selection approach is used, the presence of optimism and selection bias should be taken into account through the use of external CV.

## Acknowledgments

The authors sincerely thank Jeff Morris for careful reading of the manuscript and helpful suggestions. The authors also wish to thank James Martin for his assistance in implementing the GA software for this research.



## References

- Affymetrix (1999). Genechip analysis suite. User guide, version 3.3, Affymetrix.
- Affymetrix (2000a). Expression analysis technical manual. Technical report, Affymetrix.
- Affymetrix (2000b). Genechip expression analysis. Technical manual, Affymetrix.
- Affymetrix (2002). Statistical algorithms description document. Technical report, Affymetrix.
- Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.*, pages 6745–6750.
- Ambrose, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the USA*, 99(10):6562–6566.
- Baggerly, K., Morris, J., Wang, J., Gold, D., Xiao, L., and Coombes, K. (2003). A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics*, 3(9):1667–72.
- Braga-Neto, U. and Dougherty, E. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380.
- Dudoit, S. and Fridlyand, J. (2003). *Classification in Microarray Experiments*, chapter 3, pages 93–158. Chapman and Hall/CRC. Appearing in 'Statistical Analysis of Gene Expression Microarray Data' (ed. Terry Speed).
- Dudoit, S., Fridlyand, J., and Speed, T. (2000). Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576, University of California, Berkeley, Dept. of Statistics.
- Freitas, A. (2001). *A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery*. Springer-Verlag. Appearing in 'Advances in Evolutionary Computing' (eds. A. Ghosh and S. Tsutsui).
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
- Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, MI.

- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143, Montreal, Canada. in 'Proceedings of the 14th International Joint Conference on Artificial Intelligence' (IJCAI-95).
- Li, L., Darden, T., Weinberg, C., Levine, A., and Pedersen, L. (2001). Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinatorial Chemistry and High Throughput Screening*, 4(8):727–739.
- McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.
- Mitchell, M. (1997). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.
- Nutt, C., Mani, D., Betensky, R., Tamayo, P., Cairncross, J., Ladd, C., Pohl, U., Hartmann, C., McLauhlin, M., Batchelor, T., Black, P., von Deimling, A., Pomeroy, S., TR, T. G., and Louis, D. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, 63:1602–1607.
- Pomeroy, S., Tamayo, P., Gaasebeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E., and Golub, T. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442.
- Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutor, J., Aguiar, R., Gaasenbeer, M., Angelo, M., Reich, M., Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, A., Mesirov, J., Neuberg, D., Lander, E., Aster, J., and Golub, T. (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74.
- Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T., and Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209.
- Theodoridis, S. (1999). *Pattern Recognition*. Academic Press, San Diego.
- Xing, E. (2002). *Feature Selection in Microarray Analysis*, chapter 6, pages 110–131. Kluwer Academic Publishers. Appearing in 'A Practical Approach to Microarray Data Analysis' (eds. D. Berrar, W. Dubitzky, and M. Granzow).
- Xiong, M., Fang, X., and Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Research*, 11:1878–1887.
- Yang, Y., Speed, T., Dudoit, S., and Luu, P. (2001). Normalization for cdna microarrya data. In Bittner, M. et al., editors, *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proc. SPIE*, pages 141–152.