RICE UNIVERSITY

An Empirical Study of Feature Selection in Binary Classification with DNA Microarray Data

by

Michael Louis Lecocke

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:

Dr. Rudy Guerra, Chairman
Professor of Statistics
Rice University

Dr. Kenneth Hess, Adviser
Associate Professor of Biostatistics
UT MD Anderson Cancer Center

Dr. Jeff Morris
Assistant Professor of Biostatistics
UT MD Anderson Cancer Center

Dr. David W. Scott
Noah Harding Professor of Statistics
Rice University

Dr. Devika Subramanian
Professor of Computer Science
Rice University

Houston, Texas

May, 2005

Abstract

An Empirical Study of Feature Selection in Binary Classification with DNA Microarray Data

by

Michael Louis Lecocke

Motivation: Binary classification is a common problem in many types of research including clinical applications of gene expression microarrays. This research is comprised of a large-scale empirical study that involves a rigorous and systematic comparison of classifiers, in terms of supervised learning methods and both univariate and multivariate feature selection approaches. Other principle areas of investigation involve the use of cross-validation (CV) and how to guard against the effects of optimism and selection bias when assessing candidate classifiers via CV. This is taken into account by ensuring that the feature selection is performed during training of the classification rule at each stage of a CV process ("external CV"), which to date has not been the traditional approach to performing cross-validation. **Results:** A large-scale empirical comparison study is presented, in which a 10-fold CV procedure

is applied internally and externally to a univariate as well as two genetic algorithm- (GA-) based feature selection processes. These procedures are used in conjunction with six supervised learning algorithms across six published two-class clinical microarray datasets. It was found that external CV generally provided more realistic and honest misclassification error rates than those from using internal CV. Also, although the more sophisticated multivariate FSS approaches were able to select gene subsets that went undetected via the combination of genes from even the top 100 univariately ranked gene list, neither of the two GA-based methods led to significantly better 10-fold internal nor external CV error rates. Considering all the selection bias estimates together across all subset sizes, learning algorithms, and datasets, the average bias estimates from each of the GA-based methods were roughly 2.5 times that of the univariate-based method. Ultimately, this research has put to test the more traditional implementations of the statistical learning aspects of cross-validation and feature selection and has provided a solid foundation on which these issues can and should be further investigated when performing limited-sample classification studies using high-dimensional gene expression data.

Acknowledgements

I would like to first of all thank my adviser, Dr. Hess, for guiding me along this thesis path for the past couple of years, and helping me stay focused on not missing the forest for the trees throughout this process. From my literature review(s) on GBM's to the wonderful world of microarrays, it's been quite a ride. I would like to thank the rest of my thesis committee – Dr. Guerra, Dr. Morris, Dr. Scott, and Dr. Subramanian – for allowing me to pursue this type of large-scale empirical comparison study, as it was extremely interesting and proved to be a very useful project with what I feel are very practical and insightful results to contribute to the microarray classification literature! Jeff, my fellow Eagles fan and great friend, thank you especially for all your insights on everything from the GA to job thoughts. I would also like to thank Dr. Baggerly, Dr. Coombes, and especially James Martin for all their help in my understanding and implementation of the genetic algorithm. I would like to acknowledge my officemates over the years, from Rick to the "Sweatshop Crew of DH1041" – Ginger, Gretchen, Jason, and Chris – your patience has been tested and proven! I'd especially like to thank Rick for helping me keep my head above water during my first couple of years as a graduate student in the world of sigma fields, and for being an incredible friend. HG and Chris, the same goes to both of you as great friends and fellow survivors, especially the past couple of years. Finally, and most importantly: I'd like to thank my parents for giving me the wonderful

opportunities over the years to be where I am today, my sister for the countless email exchanges that simply cannot be duplicated and that helped the days go by much better, and last but absolutely, positively, NEVER least, my wife Meredith – her love and support (and patience!) have been beyond what I could have ever imagined before, and without her, this road would have been immeasurably tougher and much less pleasantly traveled.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Microarrays Overview

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 being made in a tissue sample at a given moment, which gives a good idea of how much of a corresponding protein is produced. Hence, a "signature" of a tumor can be obtained from the readings of mRNA abundance in the tumor cells. 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 hy-

bridization 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. The focus of this research is targeted not towards the statistical issues involved during various preprocessing stages of a microarray experiment, but instead towards the issue of feature subset selection (i.e., variable selection) – in particular, feature subset selection within the framework of a binary classification problem.

1.2 Motivation

This research is composed of a large-scale empirical analysis focused on the comparison of several popular supervised learning techniques in conjunction with several feature (gene) subset selection approaches within the context of binary classification of microarray data. The motivation behind this research is to obtain a comprehensive understanding of a variety of popular supervised learning methods as well as both univariate and multivariate feature selection methods, as applied in a binary classification setting with gene expression data - a type of comprehensive analysis that has not been conducted to this extent, and that would offer valuable insights regarding how to most effectively and honestly conduct microarray analysis research (namely

feature selection and classification).

With respect to the feature selection aspect of this research, several issues should be noted. First off, the prediction rule may not even be able to be formed using all p variables (e.g., if using Fisher's linear discriminant analysis). 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. Also, 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 [19, 26, 36, 40, 41]. The motivation for performing a multivariate feature selection technique (namely, a genetic algorithm (GA)-based approach) is grounded in the fact that the merits of implementing multivariate feature selection in the context of microarrays in general have been given relatively little attention compared to the much more widespread use of univariate approaches such as the simple T-test. In terms of ease of implementation and computation, T-tests applied on a gene-by-gene basis have of course been preferred over multivariate feature selection approaches. However, there are other considerations that should be given more attention than has been given in the past with respect to feature subset selection within the context of microarrays. Because univariate approaches can only consider a single gene at a

time, the possibility of detecting sets of genes that together jointly discriminate between two classes of patients is greatly reduced. After all, the gene subsets formed from a ranked list of the top X univariately significant genes may not include genes that are *not* discriminatory in a univariate sense yet still offer independent prognostic information when considered jointly with other genes. In implementing a GA-based search technique, the potential to select combinations of genes that are jointly discriminatory would be greater than if one combined individually predictive genes from a univariate screening method. Part of this research includes a study on how effective the more sophisticated GA-based feature selection approaches really are in detecting discriminatory genes that would be otherwise undetected among the top X genes selected by univariate screening methods. Discovery of key genes needed for accurate prediction could pave the way to better understand class differences at the molecular 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.

Ultimately, with a collection of genes that has 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 (e.g., cancer or normal, or perhaps one of two subtypes of a particular cancer). Regarding the formation of prediction rules, aside from selecting an appropriate feature selection approach and classification technique, there is also the need to assess the candidate prediction rules

in an effective and honest manner. A customary approach to estimate the error rate of a prediction rule would be to apply the rule to a "held-out" test set randomly selected from among a training set of samples. However, with microarray data, one usually does not have the luxury of withholding part of a dataset as an independent test set, as a result of the small number of samples (usually between 10 and 100, significantly smaller than the thousands of genes involved). As an alternative, cross-validation (CV) is very often used. With microarray classification problems, the practice has generally been to perform CV only on the classifier construction process, not taking into account feature selection. Leaving out feature selection from the CV process will inevitably lead to problems with selection bias (i.e., with overly optimistic error rates), as the feature selection would not be based on the particular training samples used for each CV stage. To prevent this from happening, the feature selection should be 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. This issue constitutes a prominent area of investigation within this research.

Overall, as a result of this research conducted over multiple published microarray datasets, one may be able to determine whether the success of the results obtained is really a product more of the structure of the data or of the classification process itself - a question that today remains unresolved. To address this overriding question, however, other statistical questions involved in microarray classification research, which to date remain largely unsettled, are investigated. Although many of these were dis-

cussed above and in the previous section, the following section outlines them within the framework of two "research phases."

1.3 Areas of Investigation

- *Phase I: Exhaustive and compreshensive analysis of multiple public datasets*
  - How to go about a systematic comparison of feature selection techniques and learning methods for 2-class microarray datasets?
  - Implementation of univariate (rank-based, unequal variance T-test) and multivariate (two variations of GA) approaches to feature subset selection (FSS)
  - Implementation of various learning algorithms in conjunction with FSS (e.g., SVM, DLDA, k-NN ($k = 1, 3, 7, 15$))
  - Performance evaluation: 10-fold cross-validation
    - With respect to FSS, consider classification *and* FSS performed per CV run (inclusion of FSS in CV process, serving as safeguard against selection bias)
    - Consider both single-run *and* repeated runs of CV
- *Phase II: Reflection & Interpretation*
  - Modularizing – What is the best marriage (if any) among FSS, learning algorithm, and gene subset size?
  - Should a univariate or multivariate feature subset selection (FSS) approach be implemented? In a resubstitution setting (i.e., training set only), do the

more sophisticated GA-based approaches actually detect discriminatory genes that would be otherwise undetected among the top X genes of a univariate screen? Can one deduce from the data structure which type of feature selection approach to use?

- What type of supervised learning algorithm would be best suited to a particular dataset? Can one deduce from the data structure which type of approach to use?
- Is there a gene subset size that leads to the smallest predictive errors of a given classification process across a series of datasets, or perhaps on a dataset-by-dataset basis?
- Is there a particular combination of learning algorithm and feature selection that consistently works best among a series of datasets, or does the best combination depend heavily on the particular dataset (where the notion of "best" refers to lowest error rates, based on gene subset sizes that are as minimal as possible)?
- What effect does building the feature selection process into each stage of a 10-fold CV approach to assessing the predictive accuracy have? Is there a selection bias incurred from not building the feature selection into the CV process?

- *Plan:*
  - Develop strategies for a rigorous & systematic comparison of supervised learning algorithms and feature selection techniques
  - Develop classification schemes that are efficient, accurate, and honest (i.e., consider the effect of selection bias, if possible)
  - Reach conclusions regarding prediction rules that could be generalizable to other microarray datasets
  - Provide fertile ground upon which a number of other interesting research problems can be investigated in the future

Chapter 2

Background

2.1 Introduction

First of all, a brief explanation of some key underlying aspects of this research will be discussed. In particular, some space will be given to summarizing both the notion of supervised learning in general as well as several popular techniques of supervised learning that are used in this research (results of which are discussed in Chapter 4). Next, some comments on feature (variable) subset selection (FSS) are provided. Following this is some general information on two general approaches to FSS, namely univariate ("filter") methods and multivariate methods. It is in this discussion of multivariate FSS methods that some of the basic ideas behind genetic algorithms are provided. Concluding the background material of this research are sections that discusses the importance of cross-validation as a means of assessing prediction rules

formed for performing classification using microarray data.

2.2 Supervised Learning: A General Overview

Although unsupervised learning techniques (clustering) has been among the most widely applied methods of analyzing gene expression data classification problems, supervised learning approaches have become increasingly popular in recent years. Some basic notions and several popular techniques of supervised learning are discussed in this section.

To begin with, gene expression data for p genes over each of N mRNA samples can be expressed as an N x p matrix $X = (x_{ij})$ ($i = 1, ..., N$ and $j = 1, ..., p$). Each value x_{ij} corresponds to the expression level for gene j in sample i. Each sample has associated with it a gene expression profile $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in R^p$, along with its class designation y_i (response, or dependent variable), which is one of K predefined and unordered values among $\{k_1, k_2, \ldots, k_K\}$; for this study, the setting is binary classification, so $y_i \in \{0, 1\}$. Using the observed measurements X, a classifier for K classes is thus a mapping $G : R^p \rightarrow \{0, 1, ..., K-1\}$, where $G(\mathbf{x})$ denotes the predicted class, $\hat{y} = k$, for a sample with feature vector $\mathbf{x}$.

The samples already known to belong to certain classes,
$\mathcal{L} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_{n_L}, y_{n_L})\}$, 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 $\mathcal{T} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, , y_2), \ldots, (\mathbf{x}_{n_T}, y_{n_T})\}$).

This way, the class $\hat{y}_i, \ \ (i = 1, 2, ..., n_T)$ predictions for each test set expression profile $\mathbf{x}_i$ can be made. Of course, with the true classes $y_i, \ \ (i = 1, 2, ..., n_T)$ of the test set known, a misclassification error rate (MER) can then be computed.

2.3 Supervised Learning: Some Popular Existing Methods

This section includes discussion of several popular types of supervised learning techniques that are implemented in this reserach. These are also methods that have been widely used not only with respect to microarray classification, but also in many other applications of statistical learning.

2.3.1 Standard Discriminant Analysis

Fisher's technique of linear discriminant analysis (LDA) [16] was one of the earliest formal statistical methods to ever be developed, and is still widely used today, some 69 years later. Fisher's LDA merely searches for a "sensible" rule to discriminate between classes, by searching for the linear discriminant function **a'x** that maximizes the ratio of the between-groups sum of squares to the within-groups sum of squares. This ratio is given by $\mathbf{a'}B\mathbf{a}/\mathbf{a'}W\mathbf{a}$, where B and W represent the $p \ x \ p$ matrices of between-groups and within-groups sum of squares, respectively. If **a** is the vector that maximizes the above ratio, the function **a'x** is known as Fisher's linear discriminant

function, or the first canonical variate. Mardia et al. [25] show that the vector **a** in Fisher's linear discriminant function is the eigenvector of $W^{-1}B$ corresponding to the largest eigenvalue. In general, the matrix $W^{-1}B$ has no more than $m = \min(p, K-1)$ non-zero eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, with corresponding linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ denoting the first, second, and subsequent *canonical variates*.

For gene expression levels $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, the (squared) Euclidian distance is given below, in terms of the *discriminant variables* $u_l = \mathbf{x}\mathbf{v}_l,\ l = 1, 2, \ldots, m$, from the *1 x p* vector of averages $\bar{\mathbf{x}}_k$ (for class k), for the training set $\mathcal{L}$.

$$d_k(\mathbf{x}) = \sum_{l=1}^{m} ((\mathbf{x} - \bar{\mathbf{x}}_k)\mathbf{v}_l)^2 \tag{2.1}$$

The predicted class for expression profile $\mathbf{x}$ is the class with mean vector nearest $\mathbf{x}$ in the discriminant variables space, and is described below:

$$\mathcal{C}(\mathbf{x}, \mathcal{L}) = argmin_k\, d_k(\mathbf{x})\,. \tag{2.2}$$

One should refer to Mardia et al. [25] to see how Fisher's discriminant function can also arise in a parametric setting. In particular, for $k = 2$ classes, Fisher's LDA results in the same classifier as that derived from the *maximum likelihood discriminant rule* for multivariate normal class densities with equal covariance matrices.

Dudoit et al. [15] provide some information on classification rules when the class conditional densities $Pr(\mathbf{x}|y = k)$ are already known. In a situation of this

nature, there is no need for a training set, and the class of an expression profile $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ is predicted as shown below by the maximum likelihood discriminant rule, in which the predicted class is the one that gives the biggest likelihood to $\mathbf{x}$.

$$\mathcal{C}(\mathbf{x}) = argmax_k \, Pr(\mathbf{x}|y = k) \, . \tag{2.3}$$

It should be noted that a training set could be necessary even if the distributional forms are known, to estimate the parameters of the class conditional densities. In this case, the rule becomes the *sample maximum likelihood discriminant rule*, and a training set is used to obtain the sample mean vectors and covariance matrices. That is, $\hat{\mu}_k = \bar{\mathbf{x}}_k$ and $\hat{\Sigma}_k = S_k$. If a constant covariance matrix is used, the pooled estimate is used as follows:

$$\hat{\Sigma}_k = \sum_k \left\{ \frac{(n_k - 1) \, S_k}{(n - K)} \right\} \, . \tag{2.4}$$

The maximum likelihood discriminant rule for multivariate normal class conditional densities $(\mathbf{x}|y = k) \sim N(\mu_k, \Sigma_k)$ is a quadratic discriminant rule, as shown below:

$$\mathcal{C}(\mathbf{x}) = argmin_k \left\{ (\mathbf{x} - \mu_k) \, \Sigma_k^{-1} \, (\mathbf{x} - \mu_k)' + \log |\Sigma_k| \right\} . \tag{2.5}$$

Several special cases of the multivariate normal rule are given by Dudoit et al. [15]. Each is based on a particular choice of covariance matrix for the class conditional

densities.

- If the densities have an identical covariance matrix, Σ, the rule is linear and based on the square of the Mahalanobis distance [25]

$$\mathcal{C}(\mathbf{x}) = argmin_k \left(\mathbf{x} - \mu_k\right) \Sigma^{-1} \left(\mathbf{x} - \mu_k\right)' . \tag{2.6}$$

- If the densities have the same diagonal covariance matrix $\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$, the rule is known as the *diagonal linear discriminant rule (DLDA)*:

$$\mathcal{C}(\mathbf{x}) = argmin_k \sum_{i=1}^{p} \left\{ \frac{(x_i - \mu_{ki})^2}{\sigma_{ki}^2} \right\} . \tag{2.7}$$

- If the densities have diagonal covariance matrices $\Sigma_k = diag(\sigma_{k1}^2, \sigma_{k2}^2, \ldots, \sigma_{kp}^2)$, the rule is known as the *diagonal quadratic discriminant rule (DQDA)*:

$$\mathcal{C}(\mathbf{x}) = argmin_k \sum_{i=1}^{p} \left\{ \frac{(x_i - \mu_{ki})^2}{\sigma_{ki}^2} + \log \sigma_{ki}^2 \right\} . \tag{2.8}$$

2.3.2 k-Nearest Neighbors

The k-nearest neighbors methodology is based on a distance function that describes the "closeness" of training points to a particular observation in the test set. The general idea, in a very simple case, is that for any test point x_0, one finds the k training points $x_{(i)}$, $i = 1, 2, \ldots, k$ closest in distance to the test point and then makes a classification based on majority vote among the k-nearest neighbors [19].

The notion of "closeness" implies some sort of metric, so for simplicity it can be Euclidian, Mahalanobis, or really any distance metric in the feature space:

$$d_{eucl}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_g w_g (x_{gi} - x_{gj})^2} \tag{2.9}$$

where $w_g = 1$ for (unstandardized) Euclidean distance, $w_g = \frac{1}{s_g^2}$ for standard deviation-standardized distance, and $w_g = \frac{1}{R_g^2}$ for range-standardized distance.

$$d_{maha}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j) S^{-1} (\mathbf{x}_i - \mathbf{x}_j)'} \tag{2.10}$$

where S is any $p\,x\,p$ positive definite matrix (usually the sample covariance matrix of the p variables); if $S = I_p$, $d_{maha} = d_{eucl}$. Dudoit et al. [15] describe this methodology for microarray data based on a different distance function, $1-$*the correlation*. That is, for two gene expression profiles, $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ and $\mathbf{x'} = (x'_1, x'_2, \ldots, x'_p)$, their correlation is given by

$$r_{\mathbf{x},\mathbf{x'}} = \frac{\sum_{i=1}^{p} (x_i - \bar{x})(x'_j - \bar{x}')}{\sqrt{\sum_{i=1}^{p} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{p} (x'_i - \bar{x}')^2}} \tag{2.11}$$

Again, one first determines the k nearest observations within the training set and then makes the class prediction based on which class is the most common (e.g., has the highest proportion) out of the k closest observations. Cross-validation can be used to determine k. For example, with leave-one-out CV, the distance from each of the training set observations to the *remainder* of the training set observations is computed, and a class prediction can be made for each observation. With the true

classes known, one can compare the prediction with truth, obtain a cross-validation error rate, and ultimately keep the value of k that yields the smallest error rate. More on cross-validation can be found in Section 2.5. Another issue to keep in mind is that standard k-NN algorithms equally weight all the k neighbors. However, by assigning distance weights to each of the neighbors based on their distance from the test sample (where weighting is done inversely proportional to distance from the test sample), a more sensitive rule can be obtained [14]. As discussed in Ripley [32], however, this type of weighting scheme has proven to be controversial. Overall, several choices must be made with respect to k-NN, including what distance function to use, what number of neighbors to use, whether or not to weight the votes based on distance, and of course what features to include in the classifier.

2.3.3 Support Vector Machines

The power of support vector machines (SVM's) [37] lies in their ability to map input vectors into a (possibly) higher dimension, in which the data can be separated in linear (or nonlinear) fashion using a separating hyperplane. SVM's always look for a global optimized solution and avoids over-fitting, so they potentially have the advantage of being able to deal with high-dimensional datasets. An important feature of SVM's is that the separating hyperplane can be determined without defining the actual feature space. In the binary classification framework with training data $\mathcal{L} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_{n_L}, y_{n_L})\}$, with class labels $y_i \in \{-1, 1\}$, $i = 1, 2, ..., n_L$

($\{-1, 1\}$ just used as an example for binary labels), the idea is to use a maximum margin separating hyperplane between the positive and negative samples in a higher-dimensional feature space. The "margin" is defined as the minimum distance from the hyperplane to the nearest data instance of each class. Hence, the "maximum margin" separating hyperplane is that which maximizes the margin and can be completely defined by a linear combination of the input vectors, each of which is multiplied by some weight. The hyperplane is defined by

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{2.12}$$

where the vector w defines a direction perpendicular to the hyperplane and b is the bias of the plane from the origin (i.e., varying b moves the hyperplane parallel to itself). The hyperplane of Equation 2.12 satisfies the following conditions:

$$\mathbf{x}_i \cdot \mathbf{w} + b > 0 \text{ if } y_i = 1 \quad \textit{and} \quad \mathbf{x}_i \cdot \mathbf{w} + b < 0 \text{ if } y_i = -1 \quad i = 1, 2, ..., n_L \tag{2.13}$$

So, combining the two equations above, an equivalent decision surface can be obtained as

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad i = 1, 2, ..., n_L \tag{2.14}$$

The hyperplane that optimally separates the data into two classes can be shown to be the one that minimizes the functional $\frac{\|\mathbf{w}\|^2}{2}$ (where $\|\mathbf{w}\|^2$ represents the Euclidian

norm of $\mathbf{w}$), so the optimization can be reformulated into an equivalent unconstrained problem using Lagrangian multipliers. As a quadratic optimization problem, then, the functional would be the following, with the $\alpha_i's$ the Lagrange multipliers:

$$L(\mathbf{w}, b, \alpha) = \frac{\|\mathbf{w}\|^2}{2} - \sum_{i=1}^{n_L} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{n_L} \alpha_i \tag{2.15}$$

Minimizing with respect to $\mathbf{w}$ and b, the solution can be shown to be $\mathbf{w}_0 = \sum_{i=1}^{n_L} y_i \alpha_i \mathbf{x}_i$, and inserted into Equation 2.15 yields the following, which has to be maximized with respect to the constraints $\alpha_i \geq 0$:

$$W(\alpha) = \sum_{i=1}^{n_L} \alpha_i - \frac{1}{2} \sum_{i=1}^{n_L} \sum_{j=1}^{n_L} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{2.16}$$

Once the solution has been found (i.e., $\alpha^0 = (\alpha_1^0, \alpha_2^0, ..., \alpha_{n_L}^0)$), the optimal separating hyperplane can be found to be $\mathbf{w}_0 = \sum_{support\, vectors} y_i \alpha_i^0 \mathbf{x}_i$ and $b_0 = -\frac{1}{2}\mathbf{w}_0 \cdot (\mathbf{x}_r + \mathbf{x}_s)$, where $\mathbf{x}_r$ and $\mathbf{x}_r$ are any support vectors from the two classes. Finally, the classifier can be constructed as shown below (note that only the vectors $\mathbf{x}_i$ which lead to non-zero Lagrangian multipliers α_i^0 are referred to as "support vectors"):

$$f(\mathbf{x}) = sign(\mathbf{w}_0 \cdot \mathbf{x} + b_0) = sign\left(\sum_{support\, vectors} y_i \alpha_i^0 (\mathbf{x}_i \cdot \mathbf{x}) + b_0 \right) \tag{2.17}$$

In the event the data are not separable, slack variables ξ_i can be introduced to measure the amount by which the constraints are violated. Again the margin is maximized, now taking into account a penalty proporional to the amount of constraint

violation. Formally, $\frac{\|\mathbf{w}\|^2}{2} + C(\xi_i)$ is minimized with respect to

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i \quad i = 1, 2, ..., n_L, \quad \xi_i \geq 0 \tag{2.18}$$

where C is a parameter chosen a priori, defining the cost of constraint violation. As before, the Lagrangian is formed as follows:

$$L(\mathbf{w}, b, \alpha) = \frac{\|\mathbf{w}\|^2}{2} + C\left(\sum_{i=1}^{n_L} \xi_i\right) - \sum_{i=1}^{n_L} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{n_L}(\alpha_i - \xi_i) - \sum_{i=1}^{n_L} \xi_i \tag{2.19}$$

where α_i and ξ_i are associated with the constraints in Equation 2.14 and $0 \leq \alpha_i \leq C$. The solution is determined by the saddle point of this Lagrangian in a similar fashion as before.

Finally, in the event that the decision surface is non-linear, SVM's can perform non-linear mapping of the input vectors into a higher-dimensional space by specifying a non-linear mapping a priori. The extension to non-linear boundaries is achieved through the use of kernel functions (rather than dot products between two data instances as before). Popular choices of kernel functions are [19]:

- d^{th}-degree polynomial: $K(\mathbf{x}, \mathbf{x}_i) = 1 + (\mathbf{x} \cdot \mathbf{x}_i)$
- radial basis: $K(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\|\mathbf{x} - \mathbf{x}_i\|^2\right)$
- neural network: $K(\mathbf{x}, \mathbf{x}_i) = \tanh(\kappa_1(\mathbf{x} \cdot \mathbf{x}_i) + \kappa_2$, where κ_1 and κ_2 are user-defined.

Using Lagrange multiplier analysis similar to Equations 2.15 and 2.16 , replacing the dot product with a given kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, the classifier is given by,

$$f(\mathbf{x}) = sign\left(y_i \alpha_i^0 K(\mathbf{x}_i, \mathbf{x}) + b_0\right) \tag{2.20}$$

For more details on SVM theory or further references, the reader is referred to [13] or [37].

2.4 Feature Subset Selection (FSS)

In general, feature (variable) selection is an extremely 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. To borrow terminology from the machine learning literature, there are two basic methodologies for the problem of feature subset selection – a "wrapper" (multivariate) approach and a "filter" (univariate) approach. In the former, the feature selection criterion depends on the learning algorithm used to build the prediction rule, while in the latter, the selection criterion is independent of the prediction rule. One should note that although wrapper methods could likely perform better than straightforward univariate approaches, they could do so at the risk of eventually finding a gene subset that performs well on the test set by chance alone. There is also

of course the negative aspect of significantly larger amounts of computation time being used with wrapper approaches (especially if used in a cross-validation setting). Kohavi and John provide more extensive insights on these these two approaches [22].

There are several reasons for performing feature reduction. First of all, whereas two variables could be considered good predictors *individually*, there could be little to gain by combining more than one variable together in a feature vector, as a result of potential high correlation between the variables. 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 [19, 26, 36, 40, 41]. 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. Along these lines, one should keep in mind that genes selected for their discriminatory power among two classes should later be validated for biological relevance through further experimental studies. In this way, the process of feature selection can function as an excellent way to begin identifying differentially expressed subsets of genes for future clinical research. Finally, there is the obvious issue of increased computational cost and complexity as more and more features are included.

2.4.1 Univariate Screening ("Filter") Approach to FSS

Filter approaches have for years been the most common methodology used in statistics. The features are selected based solely on the training data during a preprocessing stage (i.e., prior to running the learning algorithm). In doing so, they don't take into account any biases resulting from the learning algorithm. A couple of filtering feature selection methods were implemented on various gene expression datasets in the study of Xiong et al. [41]. One of the approaches that has been used extensively has been the simple t-test used to measure the degree of gene expression difference between two types of samples. In general, the top K genes in terms of T-statistics are retained for use in the discriminant analyses. Another type of filtering performed in the study of Xiong et al. [41] was based on the prediction strength statistic first proposed by Golub et al. [18]. Using the means and standard deviations of the (log) expression levels of each gene g in the cancerous and normal tissue samples, the K genes with highest (i.e., most informative) correlation strength statistic given by

$$P(g) = \frac{\mu_1(g) - \mu_2(g)}{s_1(g) + s_2(g)} \tag{2.21}$$

were retained for use in the discriminant analyses. In the comparison study of discrimination methods by Dudoit et al. [15], another filtering scheme used was based on the ratio of the between-groups to within-groups sum of squares of the genes. This statistics was used to take into account a large number of genes that exhibited nearly constant expression levels across observations (samples). That is, for gene j,

$$\frac{BSS(j)}{WSS(j)} = \frac{\Sigma_i \Sigma_k I(y_i = k)(\bar{x}_{kj} - \overline{x}_{.j})^2}{\Sigma_i \Sigma_k I(y_i = k)(x_{ij} - \overline{x}_{kj})^2} \tag{2.22}$$

where $\overline{x}_{.j}$ denotes the average gene expression level of gene j across all samples and $\overline{x}_{kj}$denotes the average gene expression level of gene j across samples belonging to class k. This gene selection method operates by considering only the p genes that have the largest BSS/WSS statistic for use in the classification algorithm. They also used a variant of Golub's PS statistic to compare with their filtering scheme (for more details on this variant, see [15]). The important thing to keep in mind, however, is that with these filter-based methods the effectiveness of the selected feature subsets is not directly measured by the learning algorithm to which they are applied. Furthermore, any possible interaction effects among combinations of genes are not able to be captured with this methodology.

2.4.2 Multiple Comparisons

In performing microarray analyses in which univariate-based feature selections are implemented to identify genes that are significantly differentially expressed between two conditions (say, normal and tumor), one important issue to keep in mind is that of multiple comparisons. Perhaps the best way of discussing this is in terms of multiple hypothesis testing, in which hypothesis tests are to be performed on all genes of a given microarray to determine whether each one is differentially expressed or not. As in classical hypothesis testing, a null hypothesis would be formed to compare with an

alternative hypothesis. In gene expression studies, the null hypothesis would be that there is no change in expression levels between two (or more) experimental conditions, and the alternative would then be that there is a change. If the test statistic falls into some pre-selected rejection region, then the null hypothesis would be rejected in favor of the alternative. When testing each individual gene for differential gene expression, two types of errors could be committed. When the test statistic is significant, but the gene is not truly differentially expressed, a Type I error would be committed (false positive). When the test statistic is not significant, but the gene is truly differentially expressed, a Type II error would be committed (false negative). In the context of multiple hypothesis testing, though, the situation is very complicated, since each gene would have potential Type I and II errors. Further, one should determine how to measure the overall error rate.

More precisely, take the significance level α to be the acceptable probability of a Type I error. When a t statistic for a gene is more extreme than the threshold t_α, the gene would be called differentially expressed. However, if it occurred just as a result of random effects, a Type I error would be committed (with probability α). If no mistake is made, though, the correct conclusion for that given gene would occur with probability:

$$Prob(correct) = 1 - p \tag{2.23}$$

Now, taking into account the multiple comparisons since there are, say, G total

genes being tested, the ultimate goal would be to make the correct conclusion for all the genes. The probability of making the correct decision for all the genes (assuming the events are independent) would thus be:

$$Prob(all\ correct) = (1-p)\cdot(1-p)\cdot\cdots\cdot(1-p) = (1-p)^G \tag{2.24}$$

Hence, the probability of being wrong for at least one of the genes would be the complement of the above probability:

$$Prob(wrong\ somewhere) = 1-(1-p)^G \tag{2.25}$$

This would also represent the significance level of the entire experiment, often referred to as the family-wise error rate (FWER). Rewriting Equation 2.25, one has:

$$\alpha_{Gl} = 1-(1-\alpha_g)^G \tag{2.26}$$

where α_{Gl} represents the probability of a Type I error at the overall (global) level and α_g represents the Type I error at the individual gene level. Ultimately one wants to determine the value of α_g such that the global Type I error is no bigger than α_{Gl}. A traditional correction that has been used for multiple corrections is that of Bonferroni [10, 11]. For small α_g, he noted that the first two terms of the binomial expansion of $(1-\alpha_g)^G$ could be used to approximate Equation 2.26:

$$\alpha_{Gl} = 1-(1-\alpha_g)^G = 1-(1-G\cdot\alpha_g+\cdots) \approx G\cdot\alpha_g \tag{2.27}$$

Hence, the Bonferroni correction for multiple comparisons would be given as:

$$\alpha_g = \frac{\alpha_{Gl}}{G} \tag{2.28}$$

However, the Bonferroni correction is unsuitable for microarray analyses because of the large number of genes involved, which causes the required individual gene significance level to decrease at a very fast rate (see Table 2.1).

Table 2.1: Bonferroni Significance Levels Needed at Individual Gene Level to Ensure Overall Significance Level of 0.05

Gene Subset Size	**Bonferroni**
1	0.05
10	0.005
25	0.002
50	0.001
100	0.0005
1000	0.00005
10000	0.000005
15000	0.0000033

Clearly, if a gene is still significant after the Bonferroni correction, then it is truly a differentiallyexpressed gene between two groups. However, if a gene is not significant after this correction, it could still be truly differentially expressed.

The Holm step-down group of methods are less conservative than the Bonferroni approach. Here, the genes are arranged in order of increasing p-value (arising from a simple T-test between two groups such as normal and tumor), and successive smaller adjustments are made on a gene-by-gene basis. That is, the threshold is not unique for all genes, a la the Bonferroni correction. Each gene has its own p_i value, corresponding

to the probability that its test statistic occurred by chance alone (under a true null hypothesis that the average expression between normal and tumor patients is the same). The adjusted p-values depend on each gene's position in the ordered list of raw (uncorrected) p-values. The thresholds are given as $\frac{\alpha_{Gl}}{G}$ for the first gene, $\frac{\alpha_{Gl}}{G-1}$ for the second gene, and so on until $\frac{\alpha_{Gl}}{1}$ for the last gene (with highest raw p-value). The null hypotheses H_i, $i = 1, 2, ..., k$ are rejected, where k is the largest i for which the following holds:

$$p_i < \frac{\alpha_{Gl}}{G - i + 1} \tag{2.29}$$

Both Bonferroni and Holm-based corrections for multiple comparisons assume the genes are independent, however. To take into account the often complex dependencies among genes in an organism, the false discovery rate (FDR) procedure of Benjamini and Yekutieli [9] allows for some dependencies among genes. Similar to the Holm step-down method, the FDR approach again orders the genes in order of increasing p-values. However, now the thresholds are based also on the proportion $p_0$of null hypotheses H_i that are actually true. Of course, this is not known, so $p_0$can be conservatively estimated as 1 (meaning all null hypotheses are actaully true no differentially expressed genes exist). The thresholds are given as $p_i < \frac{i}{G}\frac{\alpha_{Gl}}{p_0}$ for the i^{th}p-value. Hence, they would be (assuming $p_0 = 1$) as follows: $\frac{1}{G}\alpha_{Gl}$ for the first gene, $\frac{2}{G}\alpha_{Gl}$ for the second gene, and so on until α_{Gl} for the last gene (with highest raw p-value). The null hypotheses for those genes with p-value lower than their threshold

would be rejected. Thus, the null hypotheses H_i $i = 1, 2, ..., k$ are rejected, where k is the largest i for which $p_i < \frac{i}{G}\alpha_{Gl}$.

Finally, a more general permutation-based method of adjusting for multiple comparisons was proposed by Westfall and Young [39]. Their approach fully takes into account all dependencies among genes, which of course is important for highly correlated genes. The method proceeds by initially randomly changing the measurements between the two groups (or, by randomly permuting the labels). New p-values are computed based on the new arrangements, and the values are corrected using the Holm step-down procedure discussed previously. This procedure of re-labeling and testing is repeated thousands to tens of thousands of times. A final p-value for gene i is given as the proportion of times the t-statistic based on the original labels, t_i, is less than or equal to the test statistic from a random permutation, as shown below:

$$p - value(gene\, i) = \frac{Number\ of\ permutations\ for\ which\ u_j^{(b)} \geq t_i}{Total\ number\ of\ permutations} \tag{2.30}$$

where $u_j^{(b)}$ are the corrected values as done in Holm's method for permutation b.

Although this approach takes into account dependencies among genes, its main disadvantage is that it is an empirical process that requires a very large amount of computation and time to run, especially as the number of permutations are increased (a minimum of 1000 are often suggested).

2.4.3 Multivariate Approach to FSS

The idea behind this methodology is to incorporate feature selection more directly into a supervised learning algorithm and use the marriage of these two tasks as a more accurate tool for biomarker identification and the creation of an optimal classification rule. All genes are potential candidates, such that one isn't working from a filtered list of genes based on univariate screening. Instead, the feature selection can be accomplished in such a way that all individual genes as well as potentially all combinations of genes can be considered and evaluated in terms of how well they classify within the context of the particular learning algorithm. Multivariate approaches to feature selection are often referred to as "wrapper" approaches in the machine learning literature, since the result of the learning algorithm is used by the feature selection method to assess the effectiveness of the feature subset. The feature selection method generates feature subsets in an iterative manner, and these subsets are considered "candidate" solutions. Selected qualities of the candidate solutions are then evaluated, and the iterative process continues until a prespecified termination criterion is reached. Of course, since the learning algorithm is asked to be run with *all* sets of features considered, there is a trade-off in that computation time could become very intensive with multivariate feature selection methods.

Two multivariate methods were implemented in the Xiong et al. study [41]. The first one was a Monte Carlo (MC) method in which n randomly selected subsets of size K were obtained (where n was taken to be 200000 and K started at 1). K

was increased by 1 each time, with the MC process continuing until a prespecified classification accuracy threshold was reached, or until a prespecified subset size K was reached. The second type of method implemented in this study was a stepwise forward selection (SFS) procedure, in which all possible combinations of two genes were initially considered, with the top pair in terms of classification accuracy being retained. Next, the classification accuracy was determined based on these two genes and each of the remaining genes (hence composing a triplet of genes now). The gene that leads to the highest classification accuracy is then included in the steadily-increasing optimal subset of genes, and the process continues by adding one gene at a time until a prespecified classification accuracy threshold was reached, or until a prespecified subset of K genes is obtained. A modification of this stepwise algorithm, the sequential floating forward selection algorithm (SFFS), attempts to take into account the "nesting effect" problem, where once a particular feature is included in the optimal subset, it cannot be discarded later. For more details, SFFS is discussed in Xiong et al. [41] and further in Pudil et al. [31]. It should be noted that both backward and forward selection wrapper approaches were also implemented in the study conducted by Ambroise and McLachlan [7].

One multivariate feature selection method that has received quite a bit of attention, and one I intend to investigate as an integral part of this research, is an evolutionary algorithm known as a genetic algorithm [17], first proposed by John Holland [20]. Genetic algorithms are discussed in much more detail in Section 2.4.4.

Ultimately, the hope with multivariate feature selection methods is to avoid the limitations of univariate approaches; namely, that among some chosen gene subsets from univariate screening, there could be many genes that happen to be highly correlated with one another, offering little relevant information. Further, the selection of these genes could prevent the inclusion of other genes that may have lesser individual significance in terms of being differentially expressive, but when considered together with other genes, form a group that is significantly differentially expressed between two classes. Some genes may be highly differentially expressed within a particular subcategory of tumor but not in another. That is, samples of one class may be at different developmental stages of a cancer, which could cause some (very informative) t-statistics to be quite small and go unnoticed. Multivariate feature selection methods such as genetic algorithms could pick these up and hence provide very informative insights about the predictive structure of the data, let alone improve classifier performance. A solid understanding of the predictive structure of the data would greatly help to better understand class differences at the biological level, which could hopefully provide valuable information regarding the selection of important biomarkers to be used in the development of clinical trials for predicting outcome and various forms of treatment.

2.4.4 A Modular Multivariate Approach in an "Evolutionary" Way

Evolutionary algorithms, and in particular, genetic algorithms (GA's), apply the principles of natural evolution and selection as a means of determining an optimal solution to a feature subset selection problem. Biologically speaking, the characteristics of an organism are determined by its genetic information, which is stored in chromosomes. This information can of course be passed onto future generations with selection depending on fitness. However, this information can also be altered along the way, as a result of genetic functions such as crossover and mutation. In essence, mutation in GA's is inspired by the role played by mutation within an organism's DNA in natural evolution, as the GA periodically makes mutations in at least one member of the population of chromosomes. Crossover in GA's is analogous to the role played by sexual reproduction in the evolution of living organisms, or more specifically, by the crossover of DNA strands that occurs during reproduction. This genetic function seeks to combine elements of existing solutions together and hence form a new solution (or "offspring") with some features from each "parent." Finally, a third parallel between GA's and natural evolution is the selection process. Coinciding with the "survival of the fittest" notion of evolution, GA's perform a selection process in which the "most fit" members of the population survive, whereas the "least fit" ones are dismissed.

In the context of GA's as an optimization problem within the setting of a classi-

fication problem based on gene expression data, the "chromosome" is represented by a set of genes, and each of these chromosomes (among a population of possible chromosomes) is considered a candidate solution to apply to the classification problem. Each chromosome can be thought of as a point in the high-dimensional search space of candidate solutions. The chromosomes are often encoded as bit strings (where each locus has two possible alleles: 0 and 1). Whether or not a chromosome is passed onto the next generation depends on its fitness. That is, its passage depends on the closeness of its particular properties to those which are desired, where the notion of "closeness" depends on the fitness function chosen for use in the GA (in practice the fitness function being some type of supervised learning algorithm). As one would expect, the better the fitness, the greater the chance a given chromosome has to be selected and passed on. As discussed above, there can occur random combinations and/or changes among the passed chromosomes, which would of course induce variations in later generations of "offspring." Ultimately, optimal (or as close to optimal as possible) candidate solutions are generated after evolving through many generations. By implementing an algorithm such as a GA, one can take into account the discrimination capabilities of not only individual genes, but also combinations of genes – as mentioned in Section 2.4.3, a very important characteristic of multivariate feature selection methods. Consequently, as an inherently multivariate method, it is quite possible that GA's could find that whereas certain genes *individually* may not have significant discriminatory power, when considered in *combination* with other genes,

are significant in terms of how well they discriminate classes. A potential drawback to keep in mind with GA's is that a solution is considered better only with respect to other presently known solutions, and hence no single optimal solution is ever really attained. Further, as is the case with other multivariate feature selection methods, there is the issue that a GA never really knows when to stop iterating, aside from being provided with a prespecified number of iterations, time allotment, and/or candidate solutions to reach. More details on GA's being used in conjunction with a supervised learning algorithm (e.g., k-NN) are provided in Section 3.4.2. For more information on GA's in general, including theory behind them as well as various applications of them, the reader is referred to [27].

2.5 Assessing the Performance of a Prediction Rule: Cross-Validation

A rather simple (and perhaps ideal) 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. This "holdout" approach is preferred over using the whole training set to perform feature selection, build the classifier, and estimate the classification error rates (resubstitution (training error) estimation). This latter type of error estimation may decrease as the complexity of the classifier increases, but the generalization error would at some point then begin to increase since the model

would have adapted itself too closely to the training data (i.e., overfit the data) [19]. However, it should be noted that the "holdout" approach has the obvious disadvantage that the size of the training set, from which the prediction rule is generated, is reduced. Often, especially when the training set size is small to begin with (as is usually the case with microarray data), this is not a desirable approach since one would ideally like to use as much of the available samples as possible for feature selection and construction of the prediction rule.

As an alternative to the "holdout" approach, cross-validation 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. Cross-validation can come in a number of different flavors. In general, however, 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, while the prediction rule is trained on the remaining $K-1$ subsets. Thus, an estimate of the error rate can be obtained from applying the prediction rule to the test set. This process repeats K times, such that each subset is treated once as the test set. Ultimately, the average of the resulting K error rate estimates forms the K-fold CV error rate. Equation 2.31 below describes the standard K-fold CV misclassification error rate (MER) calculation (where the two class labels have unit difference, and the

predictions $\hat{y}_i$, $(i = 1, 2, ..., n_{test})$ take on these values; e.g., 0 and 1).

$$MER_{K-fold} = \frac{1}{K}\sum_{k=1}^{K} MER_k\,, \qquad \text{where} \tag{2.31}$$

$$MER_k = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} |y_i - \hat{y}_i|\,, \qquad \text{and} \tag{2.32}$$

$$n_{test} = \frac{N}{K} \tag{2.33}$$

Leave-one-out CV (LOO) occurs when $K = N$, where K represents the total number of available samples. In this case, each sample serves as its own test set. However, in the case of LOO CV, although it is nearly unbiased, it is generally highly variable and requires considerable computation time. In general, there is a bias-variance tradeoff with the selection of K, as larger values of K generally yield smaller bias but less stability (higher variance) than smaller values of K. It should also be noted that the CV process could be repeated multiple times, using different partitions of the data each run and averaging the results, to obtain more reliable estimates. Overall, Hastie et al. [19] suggest that 5- or 10-fold CV is a good compromise [19]. Further, at the expense of increased computation cost, repeated- (10-) run CV has been recommended as the procedure of choice for assessing predictive accuracy limited sample classification processes in general, including those based on gene expression data [12, 21].

2.6 Two Approaches to Cross-Validation: Internal and External CV

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 ("internal" cross-validation [7, 14, 26]). Although the intention of CV is to provide accurate estimates of classification error rates, using CV in this manner means that any inference made would be made with respect to the classifier building process only. However, because the significant genes are usually unknown to begin with, the idea is to make inference taking into account the feature selection process also. Leaving out feature selection from the cross-validation process will inevitably lead to a problem with selection bias (i.e., overly optimistic error rates), as the feature selection would not be based on the particular training set for each CV stage. To prevent this selection bias from occurring, an "external" cross-validation process should be implemented following the feature selection at each CV stage [7, 14, 26]. 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. Unfortunately, one cannot guarantee that the same subset of genes chosen per CV stage would be the same as originally obtained when all the training samples were considered. Hence, with external CV, a final model can be specified, but not in terms of which subset of

genes in particular make up the model. Rather, one would have the best model in terms of what size subset of genes yielded the lowest CV misclassification error rate; plus of course a model based on a CV approach that takes into account the effect of selection bias. This notion of selection bias, as well as another type of bias that is investigated, optimism bias, are discussed in the following section.

2.7 Optimism Bias and Selection Bias

Several measures of bias are considered in this study. First of all, the difference between the internal CV misclassification error rate (MER) and the resubstitution (training) MER, for any given subset size of genes, is referred to as the optimism bias:

$$\widehat{ob} = MER_{IntCV} - MER_{Resub}. \tag{2.34}$$

This estimate represents the bias incurred from using the same data to both train the classifier and estimate the performance of the classifier. Feature subset selection for both MER's used in this computation is based on using all samples of each dataset. Positive values of this estimate signify that the MER's based on internal CV were higher than those based on using all the data to both train and test the classification rule, and vice-versa for negative values.

Taking this a step further, another bias estimate of great interest used in this study is the selection bias, given by:

$$\widehat{sb} = MER_{ExtCV} - MER_{IntCV}. \tag{2.35}$$

This estimate represents the bias incurred from using the same data to both select the gene subsets and estimate the performance of the classification rule based on these subsets. Positive values of this estimate signify that the MER's based on including the feature subset selection in the CV process were higher than those based on performing the feature selection outside the CV process using all the samples of a given dataset, and vice-versa for any negative values.

Finally, consider the selection and optimism bias estimates of Eqns. 2.34 and 2.35, respectively, as the two components that comprise a third bias estimate – measure of total bias:

$$\widehat{tb} = \widehat{sb} + \widehat{ob} \tag{2.36}$$

$$= MER_{ExtCV} - MER_{Resub} \tag{2.37}$$

This estimate represents the bias incurred from using the same data to select gene subsets, train the classifier, and estimate the performance of the classifier. It also takes into account the bias from using the same data to select the gene subsets and estimate the performance of the classification rule based on these subsets. In a sense then, computing this difference between the 10-fold external CV error and the resubstitution error can also provide one with a measure of the degree of overfitting, where overfitting refers to the situation when a model fits training data well, but not

so well with respect to independent test data. In the case of total bias, the external CV MER would be treated as the test set performance, while the resubstitution MER would of course be treated as the training set performance.

Chapter 3

Previous Work & Results

3.1 Introduction

This chapter includes a review of the principle work done in the fields of feature selection and supervised learning, as applied to gene expression analysis using microarrays. With respect to feature selection in particular, there is discussion of both univariate methods and multivariate methods that have been applied, the latter including some initial work implementing the genetic algorithm. Also included in this chapter are summaries of the published microarray datasets that have been analyzed in the studies discussed (all publically available for download). Of these datasets, there are six that are also analyzed in this research. The analyses and results obtained from these six datasets are discussed in Chapters 4, 5, and 6.

3.2 Published Dataset Descriptions

- ***Alizadeh et al. lymphoma dataset [5]***

  http://www-genome.wi.mit.edu/cancer

This dataset is composed of gene expression levels measured by a specialized cDNA microarray, the Lymphochip, which contain genes that are differentially expressed in lymphoid cells or known to be immunologically or oncologically important. There are 4026 genes over 47 mRNA samples (24 germinal center B-like diffuse large B-cell lymphoma (DLBCL) and 23 B-like DLBCL). See [5] for more details on this dataset.

- ***Perou et al. breast cancer dataset [29]***

  http://genome-www.stanford.edu/sbcmp

This dataset consists of gene expression levels from cDNA microarrays containing 5776 human sequences over 27 samples (14 human mammary epithelial cells and 13 breast tumors). See [29] for more details on this dataset.

- ***West et al. breast cancer dataset [38]***

  http://data.cgt.duke.edu/west.php

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 (including 6817 human genes) over 49 breast tumor mRNA samples. Because it is believed that both estrogen receptor (ER) status and lymph node status

are important prognostic factors for the development of breast cancer, two different two-class problems were analyzed in the original study: ER+ (25 samples) vs. ER- (24 samples) and lymph node status (i.e., affected node present or node+ (25 samples) vs. affected node absent or node- (24 samples)). See [38] for more details on this dataset.

Public Datasets Analyzed in this Research

- ***Alon et al. colon cancer dataset [6]***

  http://microarray.princeton.edu/oncology/affydata/index.html

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 colon tumor samples). As discussed in [24], 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 [6] for more details on this dataset.

- ***Golub et al. leukemia dataset [18]***

  http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43

This dataset consists of gene expression levels (presumably measured from the GeneChip software) 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 [18] for more details on this dataset.

- ***Nutt et al. brain cancer dataset [28]***

  http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=82

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 linear scaling such that the mean array intensity for active ("present") genes was identical for all the scans. See [28] for more details on this dataset.

- ***Pomeroy et al. brain cancer dataset [30]***

  http://www.broad.mit.edu/mpr/CNS

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 [30] for more details on this dataset.

- ***Shipp et al. lymphoma dataset [33]***

  http://www.broad.mit.edu/mpr/lymphoma

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 [33] for more details on this dataset.

- ***Singh et al. prostate cancer dataset [34]***

  http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=75

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 [34] for more details on this dataset.

3.3 Univariate Screening

There has been a large amount of work done with respect to univariate feature selection in conjunction with classification of microarray data. As discussed throughout Section 2.3, Dudoit et al. [15] provide an in-depth comparative study of several supervised learning methods (LDA, DLDA, DQDA, CART, k-NN) for tumor classification using microarray data based on filtered (univariately screened) sets of genes. As discussed in Section 2.4.1, the gene selection method implemented by Dudoit et al. [15] operates by considering only the p genes that have the largest ratio of between to within-sum-of-squares for use in the classification algorithm. A similar type of univariate screening also compared in this comparative study was the PS statistic [18]. More recently, Dudoit and Fridlyand [14] also apply univariate screening via both t-test (using the expression values) and a rank-based t-test, the Wilcoxon Test, to analyze the datasets of West et al. [38] and Pomeroy et al. [30]. The classification schemes they used were k-NN, DLDA, DQDA, boosting with trees, random forests, and SVM's.

Starting with the more recent results of Dudoit and Fridlyand [14], for the k-NN analyses, no value of k above 5 was considered. Leave-one-out cross-validation was used in obtaining these classifieres. Also, for each CV training set, feature selection was performed, to ensure that it was taken into account when evaluating the classifiers' performances. In general, there appeared to be no significant advantage to using more complicated classification algorithms that require more tuning parameters (i.e.,

boosting trees, random forests, and SVM's versus simpler algorithms such as DLDA or k-NN), especially considering the very poor results obtained with random forests. With respect to k-NN, boosting, and SVM's, the number of neighbors, boosting iterations, and combination of cost parameter and kernel choice, respectively, did not have a significant impact on the results. As far as how many genes to include in the classifiers, there was little significant change in general as more genes were added to the models, especially for gene sets of 100 or less. However, for the gene sets of 500 or greater, there did appear to be slight increases in the misclassification results. The boosting and random forest classifiers were generally insensitive to the number of features used to build them, as expected since they have their own built-in feature selection abilities. As far as comparing the rank-based Wilcoxon test statistic and the regular t-statistic for feature selection, there seemed to be less change across subset sizes with the Wilcoxon test results than with the t-test results.

As an additional study, Dudoit and Fridlyand compared the effect of feature selection performed on the entire training set (*internal* LOO CV) and on each individual training set of a LOO CV (*external* LOO CV). As discussed in Section 2.5, the former does not lead to as realistic and honest estimates of the generalization error, since they generally tend to be overly optimistic. This statement holds true in their results, as the internal LOO CV method did indeed lead to misclassification results that were severely biased downward compared to the external approach.

Some earlier studies by Dudoit, et al [15] on the leukemia dataset of Golub (dis-

tinguising ALL from AML [18]) show similar types of results to the above. They implemented k-NN, LDA, DLDA, DQDA, classification trees (CART), bagging with trees (variations including parametric multivariate normal (MVN), non-parametric (standard bagging), and convex pseudo-data (CPD)), and boosting with trees. In this study, repeated (150) runs of training/test set partitions were performed, with feature selection performed only on each training set. The ratio of training to test set samples was 2:1. The data were preprocessed such that there were 3571 genes to be analyzed for each of the 72 samples (see [15] for more details on the preprocessing). The dataset was already divided into a training set of size 38 and test set of size 34. For this dataset, the top 40 genes in terms of the largest BSS/WSS ratio (see Section 2.4.1) were retained for inclusion in the classifiers. Aggregating methods performed were boosting (using 50 "pseudo" training sets) and CPD bagging. Increasing the number of bagging or boosting iterations didn't have much impact on the classifiers' performance. Relatively low test set error rates are obtained for each of 150 different runs, where a random sampling scheme was presumably used to re-create additional training and test sets for the remainder of the runs. The number of neighbors for k-NN was selected by cross validation (details of which are unclear), and for about half of the runs k was 1 or 2, and in general less than 7. DLDA yielded much better error rates than did LDA and DQDA. Hence, it could be said that for this dataset at least, better error rates were obtained by ignoring correlations between genes.

Overall, the highest error rates were found when LDA was used, while DLDA

(median error rate of 0) and k-NN provided the best error rates, followed by the best of the aggregating classifiers (boosting CART). As discussed in [15], the poor performance of LDA could be attributed to the fact that this classifer borrows strength from the majority of the data (unlike the more local k-NN method), so some samples may not be well represented by the discriminant variable of this dataset. Also, because of the "large p, small N" situation, the BSS and WSS matrices may not provide good estimates of the corresponding population quantities. The authors also briefly investigated the effect of increasing the number of features to include and reported that increasing the number of features to 200 did not have a significant impact on the performances of the classifiers.

For the above study and the previous one of Dudoit and Fridlyand [14], the general conclusion was that the simpler classification methods performed better than the more complicated ones. However, as discussed in [15], a couple of important factors other than generalization misclassification error that should be considered when choosing a classifier are simplicity and insight about the predictive structure of the data itself. For example, although DLDA is simple and generated low error rates, it does not take into account the effect of gene inter-correlations, which are important biological factors that should not be disregarded lightly.

The feature selection aspect of microarray classification is the stage where investigation of gene interactions could (and probably should) be conducted. The above results were all based on univariate screening as the feature selection mechanism,

which as discussed in Section 2.4.1, is not conducive to detecting groups of significant genes. Also, they aren't able to avoid the possibility of including redundant sets of genes among a list of genes (i.e., genes that are highly correlated with one another).

3.4 Multivariate Feature Selection

This section includes discussion on two general approaches to multivariate feature selection – a Monte Carlo approach and a couple of variations of the traditional stepwise forward selection approach. Their application in previous studies and the results of these studies are provided in this section as well.

3.4.1 MC and SFS Approaches

As discussed in Section 2.4.3, there were two multivariate feature selection methods, an MC method and the SFS methods, implemented in the Xiong et al. study [41]. There were three binary classification datasets used in this study: the colon dataset of Alon et al. [6], the leukemia dataset of Golub [18], and a breast cancer dataset of Perou et al. [29]. The authors found that using optimal or near-optimally selected *subsets* of genes can generate very high classification results (i.e., low misclassification rates). These results were based only on using LDA for the classification of tumor and normal samples. It should also be noted that these authors used a "holdout" method to evaluate the performance of the selected genes within the LDA analyses. An interesting caveat of this study was that the authors divided 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. That is, no cross-validation was implemented to assess the predictive accuracy of the classification processes. In this study, it was found that both multivariate methods performed better than the univariate-based T-test and PS statistic methods that had been used previously. It was also found that the stepwise method required less computation time than did the MC method. It should be noted that the accuracy of classification for forming gene subsets (only sizes 1,2, and 3 considered) was based on the total collection of tissue samples, which allowed for the presence of selection bias [7].

On the other hand, external CV was implemented on two published datasets in Ambroise and McLachlan [7]. 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 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 (although unfortunately no internal 10-fold and external LOO results were provided). 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.

The focus of the multivariate-based feature selection aspect of this research, however, will be on the use of genetic algorithms, which have received very limited use in the context of binary classification with gene expression data. An example of how a GA has been used in this context is presented in the following section.

3.4.2 Genetic Algorithms: GA + kNN

3.4.2.1 Procedures

The GA was used in conjunction with the k-NN supervised learning technique in a couple of studies [23, 24]. For the Li et al. study [24], the authors looked at the binary classification problems with the colon microarray data of Alon et al. [6] and the lymphoma data of Alizadeh et al. [5]. One should recall from Section 3.2 that five samples were omitted from the colon data, and both datasets were divided into a training set and test set (although the exact breakdown of numbers of particular classes within the training and test sets was not made clear). The basic GA/k-NN process is discussed next.

An initial population of chromosomes was created, in which each "chromosome" consisted of a prespecified number of randomly selected (distinct) genes from a pool of 2000 genes (choices of chromosome length d were 5, 10, 20, 30, 40, and 50). Since the number of possibilities of selecting 50 genes, for example, from a pool of 2000 is about 10^{100}, investigating all these is clearly is not a practical approach. Sub-

populations ("niches") were created, and typically there were 10 of these allowed to evolve in parallel per run, with 150 chromosomes in each niche. Each of the sub-populations were evolved independently. However, at each generation, the top chromosome (one from each of them) were identified and combined to replace the 10 least fit chromosomes in each niche in the following generation. Hence the best chromosomes were preserved at each generation.

The notion of "best" was determined by the fitness function used, the k-NN method. That is, the fitness, or merit, of each chromosome was determined by its ability to accurately classify the training set samples according to the k-NN method. For this fitness function, the class of each chromosome selected was compared to that of its three nearest neighbors, in terms of Euclidean distance in d-dimensional space. Their choice of k was chosen large enough to ensure tight clusters could be formed and to reduce computation time that would have otherwise been more intense with larger values of k. They employed an "all or nothing" (or "concensus") voting approach, in which a score of 1 was given to the particular sample only if all four chromosomes belonged to the same class. The scores were then summed across samples to determine the goodness-of-classification of each chromosome and hence form a sum they refer to as the "cross-validation R^2". Thus, the maximum value of this statistic would be the number of training samples, n_{train}. A proportion of correctly classified samples could then be given by the ratio $\frac{R^2}{M}$. A fitness score was given to each of the chromosomes based on this classification measure. When no

chromosomes among the first population achieve a targeted R^2 (at least 31/34 for the lymphoma data and 38/40 for the colon data), a second population of chromosomes was generated based on the survival-of-the-fittest principle. The best chromosome from each niche is passed onto the subsequent niche deterministically, and the other 149 probabilistically based on the fitness score assigned to them.

The probabilistic way in which mutation was performed was such that each chromosome was selected from its parent niche based on a probability proportional to its fitness score rank. If a chromosome were selected for transmission, between 1 and 5 of its genes were randomly selected for mutation with probabilities 0.53125, 0.25, 0.125, 0.0625, and 0.03125, respectively. Hence, a single-gene mutation is assigned the highest probability. With this mutation number determined, genes outside the chromosome replace those selected to be changed.

The entire procedure was repeated until the targeted cross-validated R^2 for the training set was achieved in any of the 10 niche runs (typically 10-50 generations needed per run), at which point that chromosome was saved and the GA/k-NN process restarted. The procedure ended when a pre-specified large number of chromosomes (not necessarily distinct) was reached (10000 for this study). With 10000 "large-R^2" chromosomes now saved, the frequency with which genes were selected is investigated; the reason being that these genes were part of chromosomes that discriminated fairly well between the two classes of a given dataset. To validate a set of top (i.e., most frequently selected) genes, the set of genes is used for classification

of the test set samples. For each d, each of the test set samples were classified using sets of these top-ranked genes. A sample would be classified as germinal center B-like DLBCL if all of its 3 nearest training set neighbors were of the same class (consensus rule), and similarly for the other class. Otherwise, the sample would be given an "unclassifiable" label.

For the lymphoma data, a systematic difference based on t-statistics was noted between the training and test sets from the *original* assignment of [5], in which the *first* 34 (of 47) samples were assigned to the training set. Hence, for this "new" original assignment, the lymphoma samples were randomly reshuffled before assignment. Two other types of assignments were also carried-out for both datasets: a *random* one in which n_{train} samples were randomly selected from the entire data set to be the training set, and a *discrepant* one in which the *last* n_{train} samples were assigned to the training set. Each of the three training sets of each dataset were used in the GA/k-NN process described above, and the genes selected for each were compared to estimate the dependence of gene selection on training set makeup. For this stability study, the chromosome length was 40. This length was chosen based on sensitivity and reproducibility studies also performed in this report, where it was shown that the selection of optimal genes was insensitive to the choice of d (see [24] for more details on these two studies).

3.4.2.2 Results

For the lymphoma data, the 50 most frequently selected genes from each of the three types of training sets for each dataset were used to classify the corresponding test set samples. Classification was done based on majority, not concensus, rule (i.e., 2 or 3 neighbors must agree with the test case). The only discrepancies found were in the reshuffled training sample assignment, where only two samples were misclassified. For the colon data, only one sample was misclassified across all three types of training assignments.

Tuning parameter issues involved in this process included the choice of chromosome length d, termination criterion R^2, number of near-optimal chromosomes, and number of top genes for test set classification. For choices of d, smaller values (5˜10) led to faster computational time than did larger values (up to 50). However with only a small chromosome length, a few genes dominated the selection process. As d increased, the gene selection process stabilized. For sample classification, the choice of d had little impact on the classification results, although $d = 2050$ provided the best overall results.

For the choice of R^2, it was reported that preliminary studies showed that gene selection is more sensitive to the choice of termination criterion than is test set classification. A less stringent criterion such as $R^2 = (n_{train} - 2)/n_{train}$ *or* $(n_{train} - 1)/n_{train}$ may cause the relative rank order of genes to vary, but had little impact on the selection of the most important genes. A less stringent statistic could have helped in

terms of computation time, as well as allowing for a higher probability of genes of predictive importance being retained even if they may have failed for a few samples.

In choosing the number of top genes for classification, too many genes could add irrelevant information to the classification. Using a concensus rule of classification for k-NN, the authors found that anywhere between 50 and 200 top genes led to the best classification of the test data. This number could change with choice of classification algorithm. This report made clear that not all genes contain relevant information, since including all 4026 genes with a concensus rule k-NN classification yielded only 31% classification accuracy on the test set. With a majority rule, only 61% accuracy was obtained. In previous work [23], similar results were found for the colon [6] and leukemia [18] datasets (2-class case between ALL and AML samples).

Finally, for the previous GA/k-NN study of Li et al. [23] looking at the colon and leukemia datasets, the same basic GA/k-NN process was used. For the colon data, an initial population of chromosomes was created, in which each "chromosome" consisted of 50 randomly selected selected (distinct) genes from the pool of 2000 genes. Sub-populations ("niches") were created, typically 10 per run, with 150 chromosomes per niche. Each of the sub-populations were evolved independently. At each generation, the top genes (one from each of them) were identified and combined to replace the 10 least fit chromosomes in each niche in the following generation. Hence the best chromosomes were preserved at each generation. For the colon dataset, the GA/k-NN method led to 6348 chromosomes selected (implementing a 3-NN scheme with

concensus rule). Again taking the 50 most frequently selected genes, these genes were used to classify 20 test set samples (first 42 samples being the training set and the other 20 the test set; both with ratios 2:1 of tumor:normal samples). When only the top gene was used for classification, there were 7 misclassifications and 2 samples that were unclassifiable. When using between 25 and 110 of the top genes, the predictive ability stabilized, but adding too many genes beyond 110˜120 only led to an increased number of unclassifiable samples as a result of high-dimensional noise being introduced. Including all 2000 genes, for example, led to 8 of the 20 samples being unclassifiable, again showing that not all gene expression data is necessary for discriminating between the normal and tissue samples.

For comparison purposes, the same GA/k-NN method was applied to the leukemia dataset (recall from Section 3.2 the assignment of samples to training and test sets). Here, a different set of top 50 genes was selected than that found in the original study of Golub et al. [18] (which used the univariate PS statistic approach discussed in Section 2.4.1). Test set classification using 3-NN and concensus rule showed only one sample (of 34) was misclassified. Another analysis was done using 5-NN (again concensus rule), which led to similar results, with the exception of now incurring one unclassifiable sample. For more details on these analyses, refer to Li et al. [23].

3.5 What's Next: A Large-Scale Investigation

Much of the work done to date with respect to binary classification of microarray data has been based on univariate-based feature selection approaches. Of course, if among the thousands of features (genes), there exist several dominant individually predictive genes, then perhaps simple univariate gene ranking approaches would suffice. Nonetheless, however, based on the results of Section 3.4 above, I believe there is definitely cause to further investigate the merits of a multivariate, modular approach to feature subset selection for binary classification with high-dimensional data from microarrays. In doing so, currently existing methods of performing feature selection in this context should be rigorously assessed. Biologically, it is known that genes work together, so it seems reasonable to further investigate performing variable selection in a multivariate manner. At the same time, it should also be noted that biological interaction among genes does not necessarily imply that the genes will be jointly predictive in a classification problem. An evolutionary algorithm such as a GA, which has had relatively little use in the context of microarrays, has the advantage of being able to find at least near-optimal solutions (i.e., subsets of jointly significant genes) to use for the classification of microarray data in a "large p, small N setting", in which there are simply too many genes to find truly optimal combinations of genes through a completely exhaustive search of the high-dimensional feature space. Further, there are solutions that otherwise would have very likely not been found by combining individually predictive genes found from univariate screening (as discussed in Sections

2.4.1 and 3.3).

The work of Li et al. [24] provided some promising results on the use of GA's in conjunction with k-NN as the fitness function. However, I believe some modifications to their method are in order. First of all, with respect to the final solution selected, they essentially ignore all the candidate chromosomes (i.e., solutions, or combinations of genes) generated by the GA process, since they resorted back to a rank-based selection approach by choosing the top genes from among the most frequently selected genes among all the candidate chromosomes. With respect to the k-NN fitness function, the authors described a cross-validated R^2 as a goodness-of-classification measure to describe the number of correctly classified samples, based on a concensus rule. However, for some datasets one might incur a more serious problem with unclassifiable samples than was the case with their analyses if the k nearest neighbors of a sample don't satisfy the concensus vote required for classification. Overall, a simpler and less computationally intense fitness function should be investigated.

Ultimately, on a much broader scale, is the question of whether the success of the classification accuracy results from a given prediction rule is really a product more of the structure of the data or of the classification process itself – a question that today unfortunately remains unresolved. To address this question, however, a number of other very important and currently open-ended issues should be investigated in a larger-scale empirical comparative study than any that has been undertaken to date. First of all, not only should one seek to determine what type of feature selection ap-

proach should be used, but also what type of learning algorithm and what gene subset size are best suited for performing binary classification based on gene expression data from a given dataset. With respect to the notion of "best" predictor gene sets, the hope is that the final group of genes selected are strongly correlated with (predictive of) class distinction as much as possible, but also as uncorrelated as possible with each other. Of course, based on this final subset, there is then the question regarding how successful they are in discriminating between two classes when applied to an independent test set. Cross-validation is an important technique for assessing the predictive accuracy of classification processes with microarray data, where one usually does not have the luxury of withholding samples as test and/or validation sets. Other key CV-related issues that warrant further investigation include the effect external vs. internal CV when assessing the predictive accuracy of a given prediction rule and how effective external CV really is in taking into account selection bias, for a number of different types of classifiers across a number of two-class gene expression datasets (i.e., for univariate- and GA-based feature selection techniques, various supervised classification algorithms, and various gene subset sizes). On the gene detection front, another issue to investigate is how effective a GA-based feature selection approach really is in detecting genes that would be otherwise undetected by univariate screening methods. As a result of conducting a large-scale empirical study over multiple published two-class microarray datasets – a type of comparative study that to this date has not been undertaken – the hope is that more insights into all these issues

can be obtained.

Chapter 4

Univariate-Based FSS Results

4.1 Introduction

This is the first of three chapters presenting the results obtained in this research. This chapter includes results obtained using a univariate-based means of feature subset selection in conjunction with six different types of classifiers. To assess the predictive accuracy of the various models, both external and internal single- and repeated- (10-) run 10-fold cross-validation was used. All analyses of this chapter were performed using the *R* statistical software package [35] on a Red Hat Linux machine (dual Intel(R) 3.06 GHz processors and 4 GB memory).

4.1.1 Supervised Learning Methods

In this study, three well known and widely used choices of supervised learning algorithms were implemented – support vector machines (SVM's; linear kernel), diagonal linear discriminant analysis (DLDA), and k-nearest neighbors (k-NN). Each of these three methods have been shown to be successful learning algorithms for the problem of classification using gene expression data. Also, each of them function in inherently different ways. For these reasons, they have been selected for use in this large-scale empirical study. For k-NN, it should be noted that four different variations are implemented (k=1, 3, 7, 15, to cover a wide range of k; Euclidean distance metric). Hence, the total number of different learning algorithms studied was actually six. For more details on each of these classifiers, the reader should refer to [14] and [15].

4.1.2 Feature Subset Selection

Rank-based, unequal variance T-tests were performed on each of the genes from the designated training sets of samples selected from the datasets being studied. In each training set, this resulted in an ordered (by increasing p-value) list of "top genes" for use in generating various "top gene subset size" models for each of the six classifiers implemented. Note that the training set could either be all N samples of a given dataset if "internal" 10-fold CV is being used, or 90% of the N samples if "external" 10-fold CV is being used (see Section 4.1.3).

For the "internal" CV results, in which all samples were used in performing the

T-tests, the final list of ordered genes, along with their (raw) p-values, are provided in Tables A.1 – A.6 of Appendix A. Note that these lists include only up to the top 25 genes. It should also be noted that the first column of these tables, "Gene Index", refers to the row (or column, depending on how one imports the dataset) number of that particular gene, so the reader is referred to the original dataset for cross-referencing any gene of interest (see Section 3.2 for more information pertaining to each dataset's website where the data was downloaded). Included in each of these tables are the adjusted p-values from the Benjamini-Yekutieli ("BY") FDR adjustment, Holm step-down adjustment, Bonferroni correction, and finally the Westfall-Young ("WY") permutation-based correction. As expected, the importance of taking into account multiple comparisons was evident, as the p-values were not as small once they were adjusted, especially as one proceeds down the list of top genes for each of the datasets. For each dataset, it should be noted that the top 25 genes across all methods still maintained significant p-values (at the 0.05 significance level); only the last 5 of the 25 genes listed from the Nutt dataset had permutation-based p-values slightly above the 0.05 level. If one were to actually report particular top gene subsets for model specification and their associated p-values, it would be more accurate to report results from one of the adjusted p-value approaches. For more details on these corrections for multiple comparisons, refer to Section 2.4.2.

4.1.3 Internal and External CV

To assess the predictive accuracy of the classification processes, a classic 10-fold cross-validation approach was used. Although the commonly used leave-one-out approach of CV is nearly unbiased, it can provide highly variable MER estimates. In addition, because repeated-run CV is implemented in this research, the computation burden of leave-one-out CV would be quite heavy compared to a lower-fold variation of CV. For this research, 10-fold CV, which yields a slightly more biased but less variable estimate of the misclassification error rate (MER), is used. 10-fold CV has also been shown to be a reliable approach for assessing the predictive accuracy for limited sample classification problems in general, including those based on gene expression data [12, 19, 21]. By "classic" CV, the key is that each of the ten test set partitions are mutually exclusive to each another. This chapter includes results of both internal and external CV, as well as resubstitution error rate results (i.e., "apparent", or training, error rate; the estimation of misclassification error based on using *all* samples to both build each model as well as evaulate each model). With respect to the resubstitution errors, one would expect these curves to all have very low, if not perfect, MER's. Refer to Section 2.5 for more on cross-validation in general.

4.1.4 Single and Repeated- CV runs

Also included in this chapter are results based on either the standard single-run 10-fold CV process, or a repeated- (10-) run process of the 10-fold CV. The latter refers

to running the standard 10-fold CV process 10 separate times and averaging the resulting ten misclassification rates to obtain the final misclassification error rate, providing for a more stable, Monte Carlo type of estimate. Equation 2.31 describes the standard (single-run) K-fold CV MER calculation (for this study, $K = 10$), where the two class labels have unit difference, and the predictions $\hat{y}_i$, $(i = 1, 2, ..., n_{test})$ take on these labels; e.g., 0 and 1).

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 for each dataset. That is, iteration 1 of 10 would use the same set of ten training and mutually exclusive test set partitions across all the classifiers. Iteration 2 of 10 would do the same, followed by iterations 3, 4,..., and 10 of 10. There could of course be some overlap among training and/or test set samples from iteration to iteration, but this is inevitable for the repeated-run analyses.

4.1.5 Plot Breakdowns

This chapter begins with plots of misclassification error rates (MER's) vs. top gene subset size, for four different "flavors" of 10-fold CV process, with discussion following the first two plots (internal CV; Figures 4.2 and 4.3) and the second two plots (external CV; Figures 4.4 and 4.5). The "flavors" are as follows:

- *single-run* 10-fold *internal* CV (Figure 4.2)

- *repeated* (10)-run 10-fold *internal* CV (Figures 4.3)
- *single-run* 10-fold *external* CV (Figures 4.4)
- *repeated* (10)-run 10-fold *external* CV (Figures 4.5)

Other repeated-run-based plots are also included in this chapter. These include resubstitution error plots, external CV vs. internal CV plots, and finally plots looking at the effect of both optimism and selection bias incurred from performing resubstitution error assessment over internal CV and from performing internal CV over external CV, respectively. For each of the above analyses, there are six individual plots corresponding to the six datasets, each of which contains six MER curves corresponding to each of the six classifiers considered across the selected "top" gene subset sizes. From these plots, any effects of the datasets, classifiers, as well as of the gene subset sizes on the misclassification rates should be seen. The choice of which top gene subset size to use for each dataset was made ensuring that a full range of the possible magnitudes of subset sizes was taken into account for each dataset. Since some datasets were bigger than others, their plots obviously included larger-sized gene subsets.

4.2 Preprocessing of Datasets

As discussed in Section 3.2, there were 6 datasets analyzed in this research, all of which are from Affymetrix microarrays [1, 2, 3, 4]:

- Alon et al. colon cancer dataset (2000 genes; 57 samples: 20 (35%) normal and 37 (65%) tumor) [6]
- Golub et al. leukemia dataset (7129 genes; 72 samples: 25 (35%) AML and 47 (65%) ALL)[18]
- Nutt et al. brain cancer dataset (12625 genes; 50 samples: 28 (56%) glioblastoma and 22 (44%) anaplastic oligodendroglioma) [28]
- Pomeroy et al. brain cancer dataset (7129 genes; 90 samples: 60 (67%) MD and 30 (33%) other) [30]
- Shipp et al. lymphoma dataset (7129 genes; 77 samples: 58 (75%) DLBCL and 19 (25%) FL) [33]
- Singh et al. prostate cancer dataset (12600 genes; 102 samples: 50 (49%) normal and 52 (51%) tumor) [34]

The only preprocessing that was done on each of the datasets was to normalize the arrays such that they each have zero mean and unit variance (an approach also used in the comparative gene expression study of Dudoit et al. [15]). 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 (i.e., reduce the effect of processing artefacts, 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 [42]. Currently there is no universally accepted means of normalizing microarray data.

4.2.1 An Initial Glimpse of the Datasets: Unsupervised Learning via Multimensional Scaling

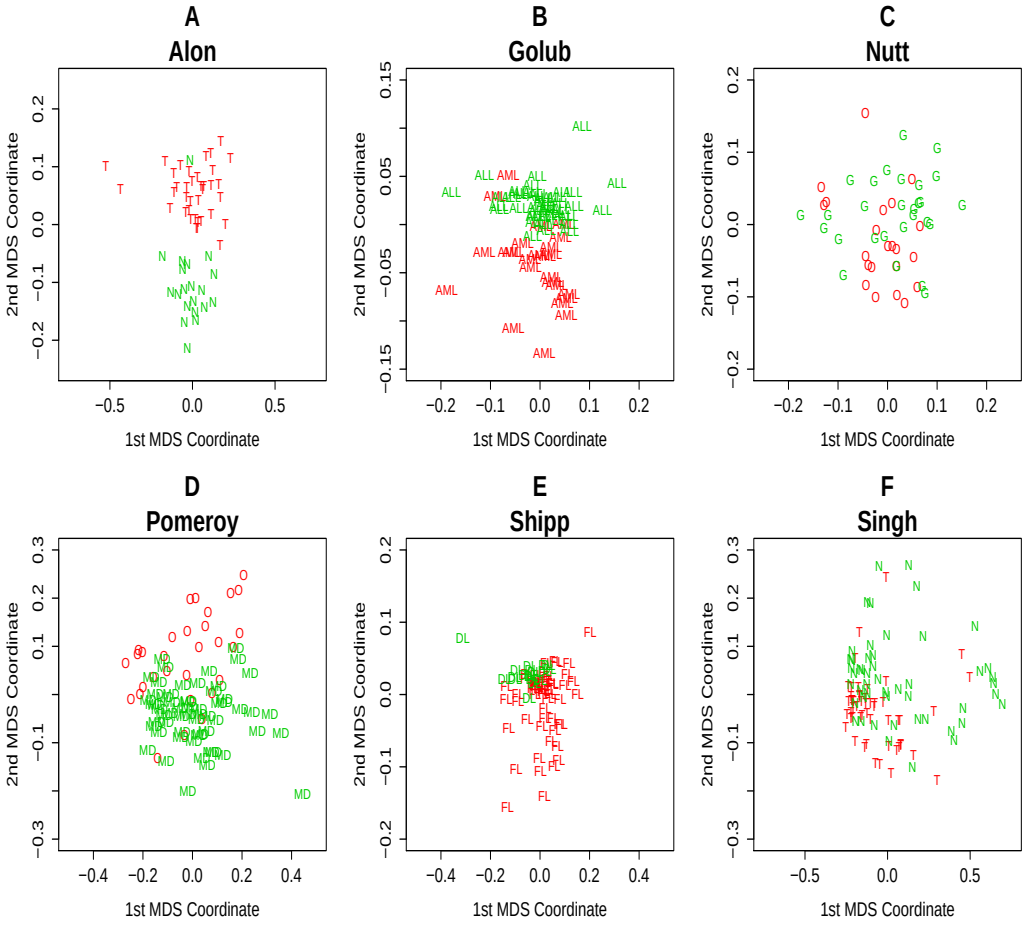
To get an initial idea of how difficult the classification task will be for each dataset, a popular type of unsupervised learning technique, multidimensional scaling (MDS), is applied to each of the datasets, using all samples and all genes in each case. By unsupervised, the goal is to be able to visualize groups (clusters) among the data without the use of any classes defined a priori, as is done with supervised learning methods. The technique of MDS is a method that maps data points in R^p to a lower dimension manifold. For binary classification with gene expression data then, MDS takes each gene expression profile $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in R^p$, $i = 1, ..., N$ and maps this data to two-dimensional space. The general idea of MDS is to represent the expression profiles as points in some lower-dimensional space than the original space such that the distances between points in the reduced space correspond to the dissimilarities between points in the original space. If acceptably accurate representations can be obtained in say, two dimensions, MDS can serve as a very valuable means to

gain insight into the structure of the dataset. Of particular interest in this research would be to obtain a visualization of the expression data which can then be used to give a potentially meaningful interpretation of the relationships between the samples in the data. The distances need not be based on Euclidean distance (Equation 2.9) and can actually represent many types of dissimilarities between sample points (e.g., $1 - correlation$ measure, as discussed in Section 2.3.2). The distance information is provided by an $N\,x\,N$ distance matrix. If a dataset can be easily classified, one would expect to see clear separation between two groups of sample points (corresponding to the two classes). Because this technique is being used merely for the purpose of getting some initial visual ideas of the ease of separability of the two classes for each dataset, further details on MDS are not discussed here. For more details, the reader is referred to [19, 25].

The plots presented in Figure 4.1 show the first two MDS coordinates for each of the six datasets (all genes and all samples used). The results are based on using $1 - correlation$ as the same measure on the normalized expression values. The reader is referred to Section 3.2 for more clarification of what the two classes represent in each of the plots. As seen in Figure 4.1, it would seem to be a more difficult classification task for the Nutt dataset since the two groups demonstrate considerable overlap. On the other hand, one would expect the classification task for the Alon and Golub datasets to be considerably less difficult since they show much better group separation. To see if these conjectures are true, the remainder of this chapter is devoted to the

construction and predictive accuracy assessment of the various classifiers considered for each of the six datasets.

Figure 4.1: MDS Plots for Each Dataset (Measure=1-corr)



4.3 Internal CV Results

This section includes results based on a) running the standard 10-fold internal CV process once and b) running the standard 10-fold internal CV process 10 separate times and averaging the resulting ten misclassification rates, to obtain the final misclassification error rate.

Figure 4.2: 1×10-Fold CV; Univ FSS Based on All Samples

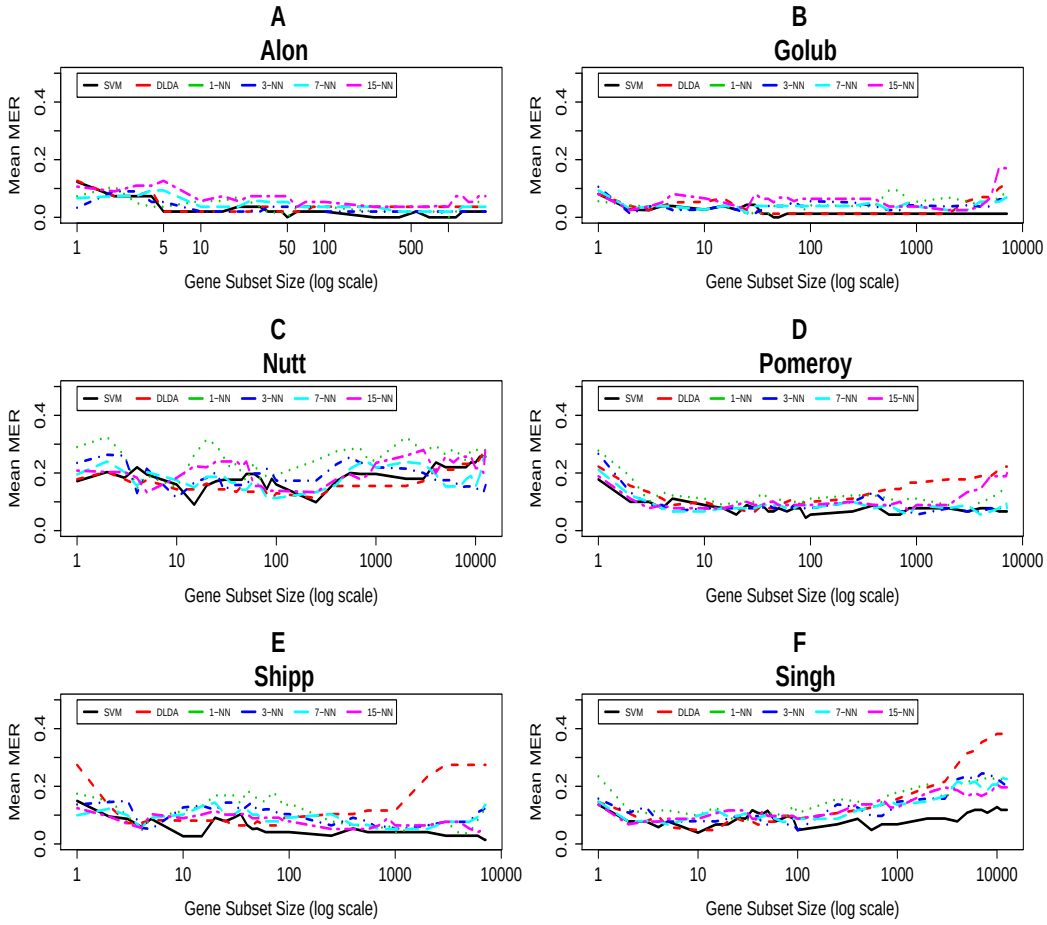
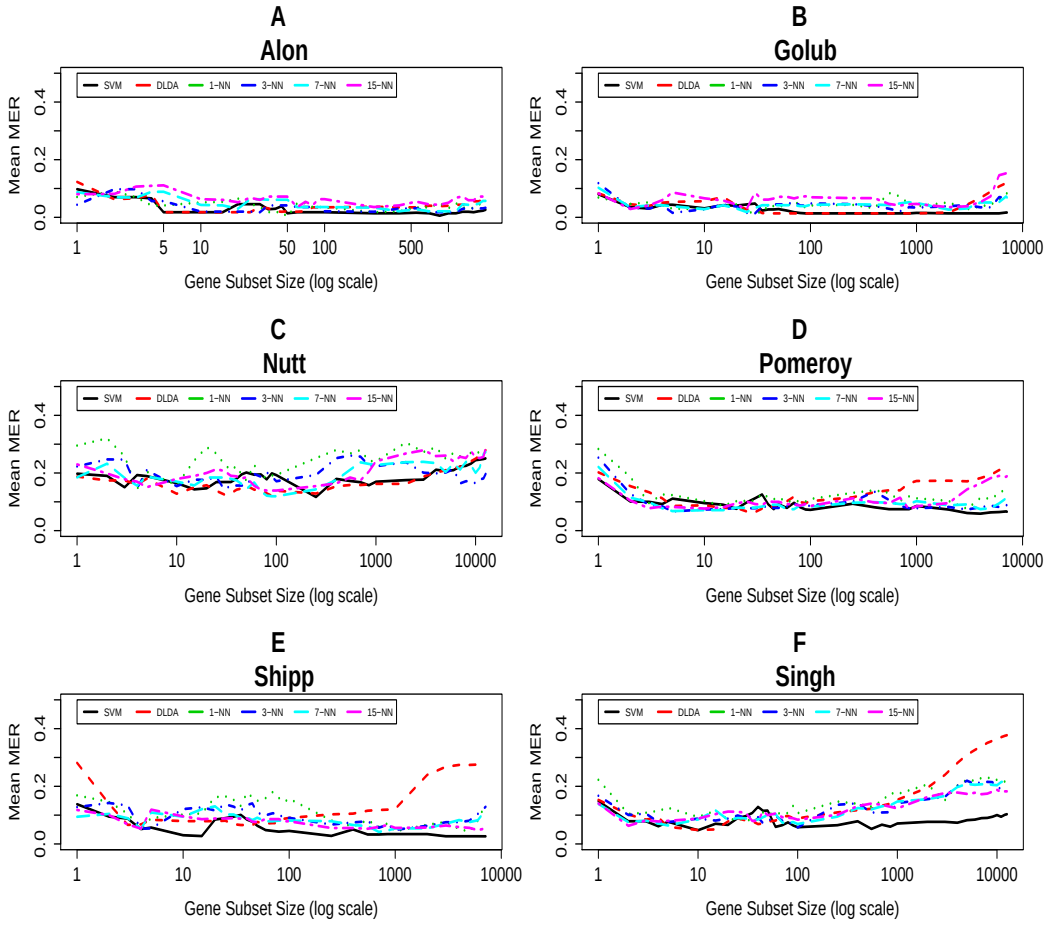


Figure 4.3: 10×10-Fold CV; Univ FSS Based on All Samples



With respect to the internal 10-fold CV process, in comparing the single- and repeated-run results shown in this section, the repeated runs approach of the CV process yielded slightly smoother curves overall. The Alon and Golub datasets (Figures 4.2 A, B and 4.3 A, B) had the lowest MER's in general for all the classifiers (all ≤ 0.10), followed by the Pomeroy, Shipp, and Singh datasets (Figures 4.2 D,E,F and 4.3 D,E,F). The MER's of these latter three were all generally between 0.10 and 0.20, except for the DLDA curves in each, which were marked by a relatively large increase for gene subset sizes of 1000 and greater. A similar large increase in error rates for DLDA was also present with the Shipp data (Figures 4.2 E and 4.3 E), again taking place around gene subset size 1000. On the other hand, the Nutt dataset (Figures 4.2 C and 4.3 C) was marked by much higher MER's (mostly in the range of 0.2 and 0.3) than the other five datasets across all the gene subset sizes. Also, there seemed to be less variation among the classifer curves for the Alon, Golub, and Pomeroy datasets than in the other three. It was not immediately clear from these plots that one particular classifier was obviously better than any of the others across all the datasets. One should note that the SVM classifier for the Shipp and Singh results seemed to be significantly better than the other datasets for gene subset sizes (≥ 100), while the DLDA classifier results for the Pomeroy, Shipp, and Singh datsets seemed to be significantly worse than the rest for even bigger gene subset sizes (≥ 1000). The SVM classifier was also the least variable among the six classifiers in all the datasets. SVM seemed to be consistently one of the best classifiers in each of the datasets. No

gene subset size stood out as the best, as none of the classifiers' curves showed any dominant upward or downward trend with increasing gene subset size.

There appeared to be a dataset-learning algorithm interaction (i.e., for any given number of genes, the rank ordering of the learning algorithms varies from dataset to dataset), as well as some interaction between learning algorithm and gene subset size (i.e., for any dataset, the rank ordering of learning algorithms varies from subset size to subset size). The results also suggested that minimal interaction between dataset and gene subset size was present (i.e., for any learning algorithm, the rank ordering of the datasets varies only slightly from gene subset size to gene subset size). Looking at all three of these experimental parameters together, it seem that a three-way interaction among them exists, since the effect of the number of genes used for classification seemed to be dependent on both the classfier and the dataset. The situation was further complicated because not only does the magnitude of the effect of the number of genes used change, but also the nature of the effect (i.e., monotonic or not) changes with the choice of both dataset and learning algorithm. Interactions aside, it appeared that dataset had the biggest main effect on the error rates, followed by the size of the top gene subset, with learning algorithm having the least effect (although the SVM curves of Figures 4.2 and 4.3 at least, tended to be among the best across gene subset sizes, for all datasets except the Nutt one).

4.4 External CV Results

This section includes results based on a) running the standard 10-fold external CV process once and b) running the standard 10-fold external CV process 10 separate times and averaging the resulting ten misclassification rates, to obtain the final misclassification error rate.

Figure 4.4: 1×10-Fold CV with Univ FSS Built-in

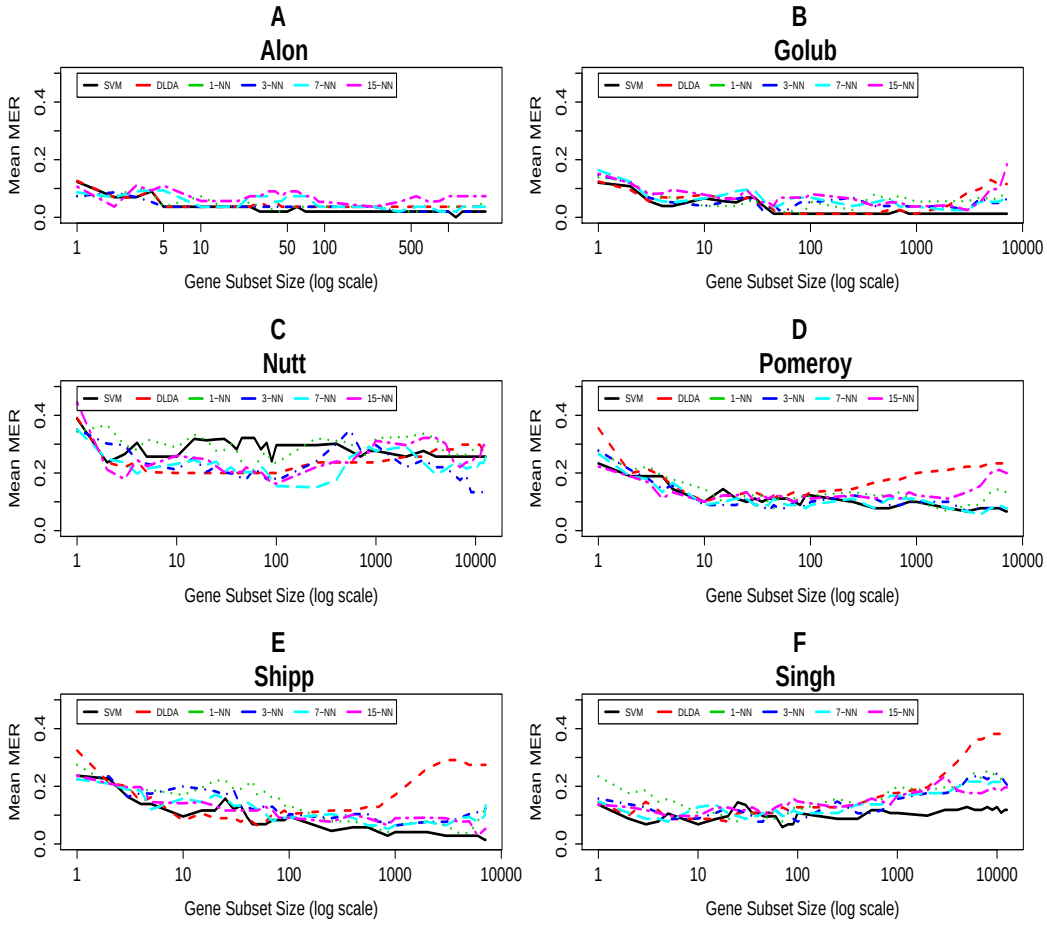
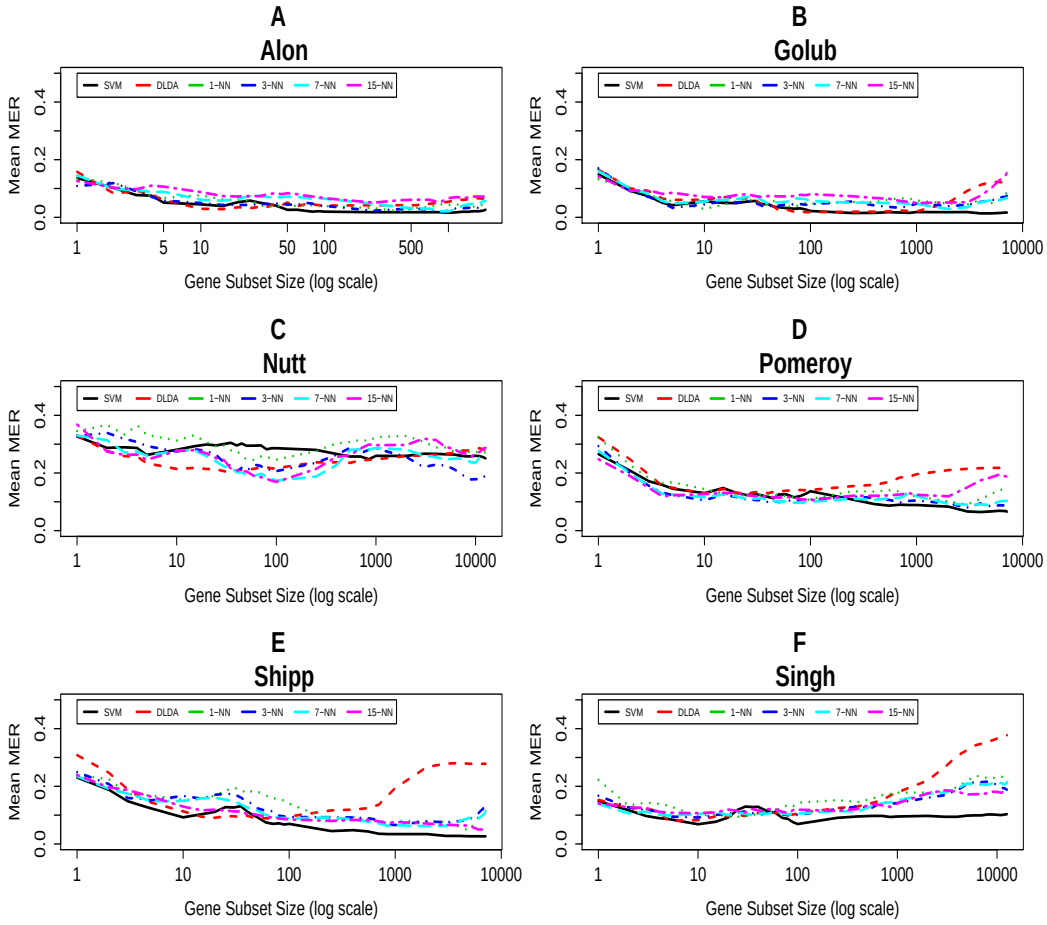


Figure 4.5: 10×10-Fold CV with Univ FSS Built-in



With respect to the external 10-fold CV process, in comparing the single- and repeated-run results shown in Figures 4.4 and 4.5, one should immediately note that performing repeated runs of the CV process led to smoother curves overall. The repeated-run MER curves for these external CV results were marked by more improvement in stability over the corresponding single-run results than was the case with the internal CV repeated-run results' improved stability over their corresponding single-run results (discussed in Section 4.3). This was expected with taking the average MER's from multiple runs of the process, rather than just taking a single run's results. Otherwise, the same general conclusions found with the internal CV analyses are reached and discussed here. Figures 4.4 and 4.5 show that the classifiers' MER curves, except for DLDA in a few datasets, all seemed to follow a general decreasing (or constant) pattern with increasing subset size, especially with respect to subset sizes ≤ 5. The Nutt results (Figures 4.4 C and 4.5 C) were marked by much higher MER's (mostly in the range of 0.20 and 0.30) in general across all the gene subset sizes. The Alon and Golub results (Figures 4.4 A, B and 4.5 A, B) possessed the lowest MER's in general (mostly in the range of 0.05 and 0.10) across subset sizes, followed by the Pomeroy, Shipp, and Singh datasets (generally between 0.1 and 0.2) (Figures 4.4 D,E,F and 4.5 D,E,F). Also, there seemed to be less variation among the classifer curves for the Alon, Golub, and Pomeroy datasets than in the other three. In general, it was not totally clear from these plots that one particular classifier was obviously better than any of the others across all the datasets, as they were all quite

similar to each other. However, as was the case with the corresponding internal CV plots of Section 4.3, SVM at least seemed to consistently be one of the best classifiers in each of the datasets (aside from the Nutt data, where it was generally unusually higher than the other classifiers for a large number of subset sizes). The SVM curves were generally marked by a typical pattern of slightly improved performance with increasing number of top genes. The Singh and Nutt data were exceptions to this trend with SVM, as the curves in these cases showed no dominant upward or downward trend across subset sizes. With respect to the DLDA curves it was interesting to note that they seemed to decrease in the range of the top 1 to 10 genes, followed by a steady (or slight increase) from 10 to 100 genes, and then quickly increased beyond the top 100 genes. A final note on these plots is that the SVM classifier curves for the Shipp and Singh datasets seemed to be significantly better than the other classifiers' curves for some of the larger gene subset sizes ($\geq$ 100), while the DLDA classifier results for the Pomeroy, Shipp, and Singh datsets seemed to be significantly worse than the rest for even bigger gene subset sizes ($\geq$ 1000). The SVM classifier was also the least variable among the six classifiers in all the datasets. As far as what size of gene subset was best, this was not immediately clear, as it varied from classifier to classifier across top gene subset sizes, for each dataset.

Overall, similar to what was discussed in Section 4.3, Figures 4.4 and 4.5 suggest that there appeared to be a dataset-learning algorithm interaction, as well as some interaction between learning algorithm and gene subset size. The results also suggested

that an interaction between dataset and subset size is present, albeit very slight. Finally, taking all three parameters together, the presence of a three-way interaction among them was evident (as discussed in Section 4.4). Considering each parameter individually, it appears that dataset had the biggest main effect on the error rates, followed by the size of the top gene subset, with learning algorithm apparently having the least effect (although the SVM curves of Figures 4.4 and 4.5, at least, tended to be among the best across gene subset sizes, for each of the datasets except the Nutt one).

4.5 Resubstitution, Internal & External CV, & Selection & Optimism Bias: A Closer Look at the Repeated-Run CV Approach

4.5.1 Resubstitution, Internal CV, and External CV MER's

A closer look at the repeated-run 10-fold CV results is provided in Figures 4.6 – 4.11. Included in each plot are the internal CV and external CV MER curves together for a given classifier and dataset. Also included in each plot are the resubstitution error rates.

Figure 4.6: 10×10-Fold CV; Internal CV vs. External CV; Univ FSS; Alon Data

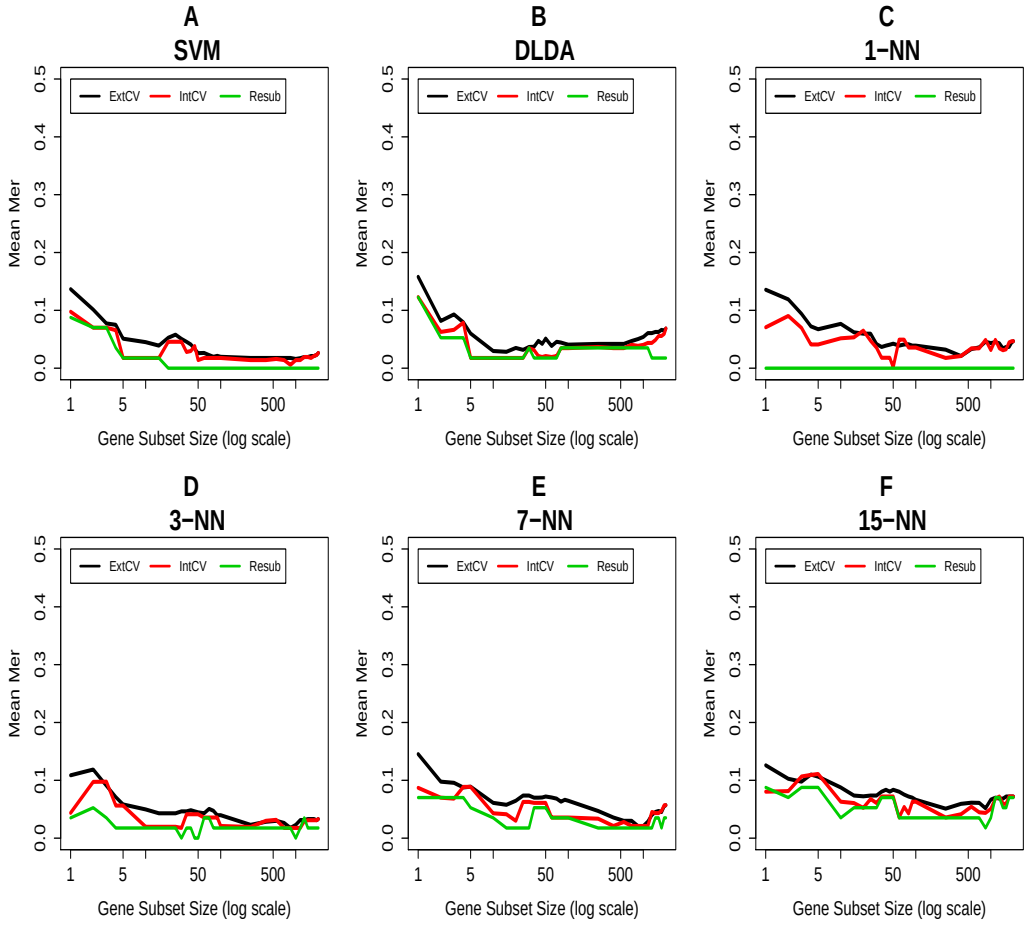


Figure 4.7: 10×10-Fold CV; Internal CV vs. External CV; Univ FSS; Golub Data

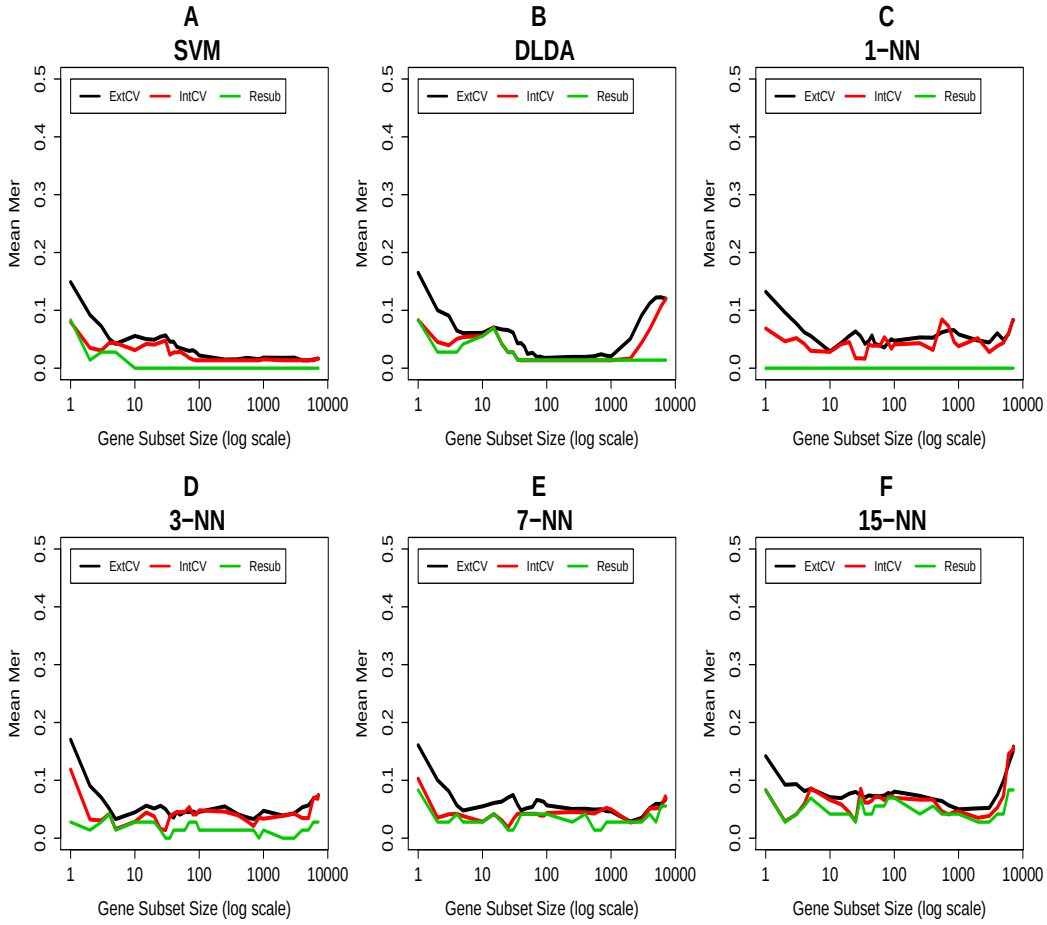


Figure 4.8: 10×10-Fold CV; Internal CV vs. External CV; Univ FSS; Nutt Data

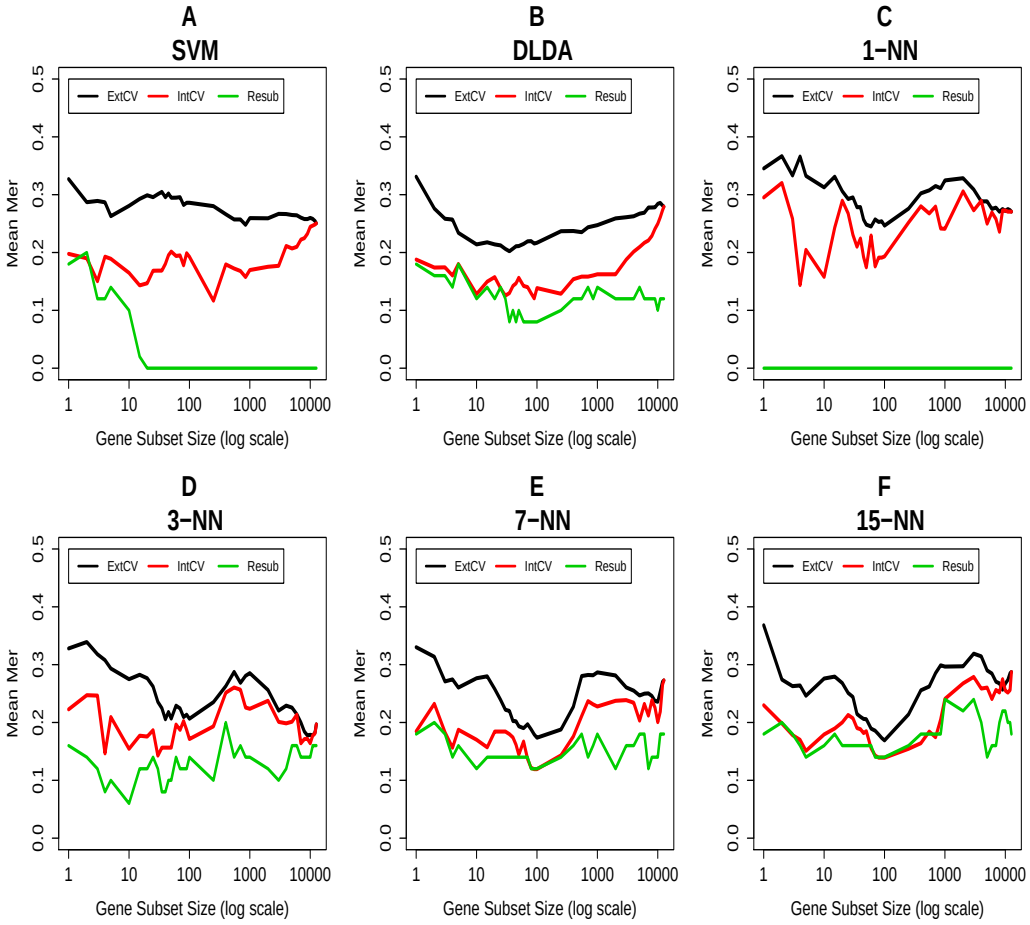


Figure 4.9: 10×10-Fold CV; Internal CV vs. External CV; Univ FSS; Pomeroy Data

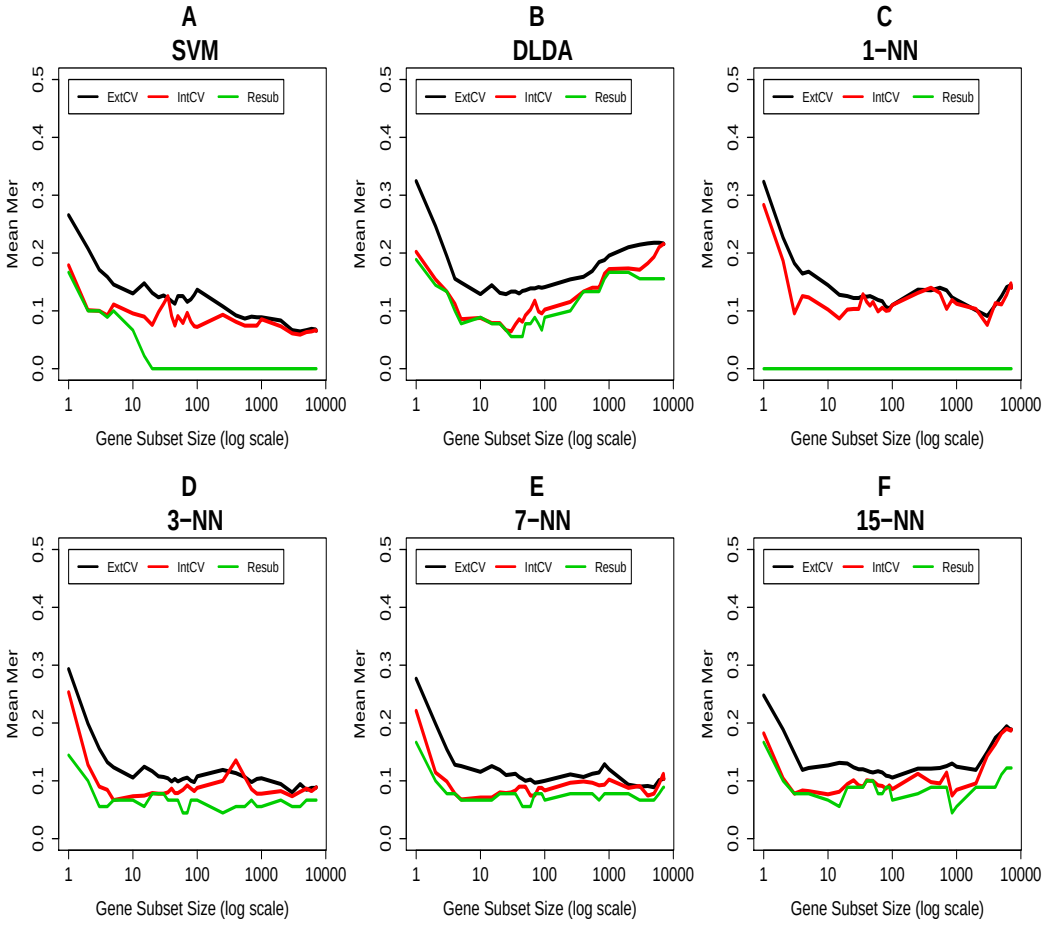


Figure 4.10: 10×10-Fold CV; Internal CV vs. External CV; Univ FSS; Shipp Data

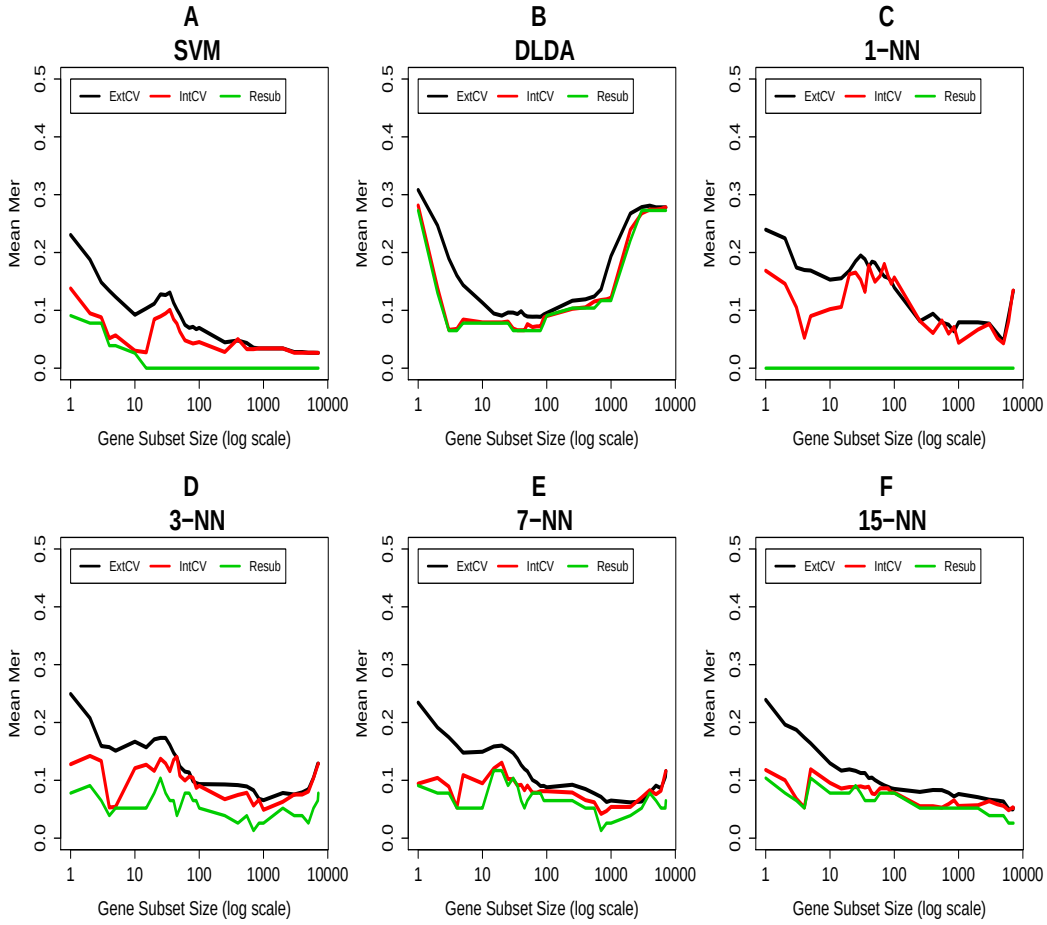
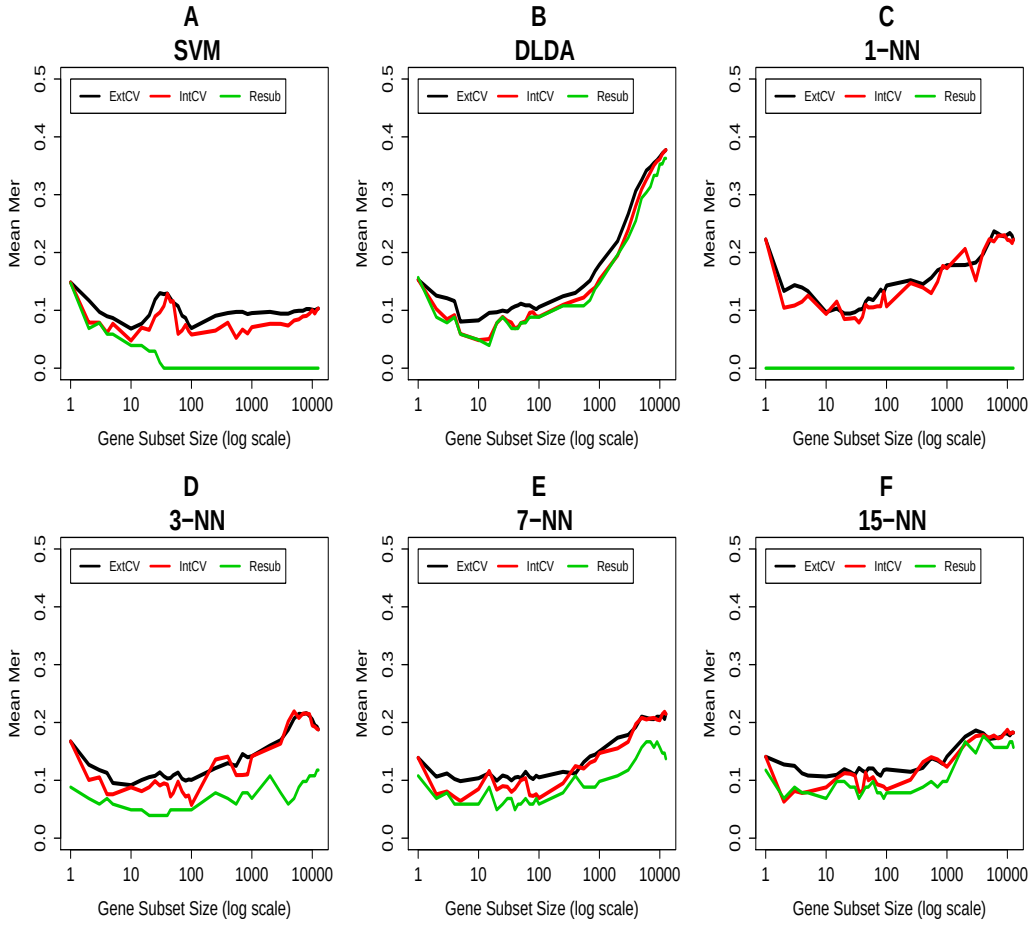


Figure 4.11: 10×10-Fold CV; Internal CV vs. External CV; Univ FSS; Singh Data



Several things should be noted from these plots. First off, with respect to the resubstitution error curves, the ones for the 1-NN classifier were always 0, as one would expect. In general most of the classifiers' resubstitution error rates across all datasets (but the Nutt one) were relatively stable, except perhaps for the DLDA ones from the Shipp and Singh datasets, for gene subset sizes greater than 1000. The Nutt dataset was marked by unstable resubstitution curves for all classifiers (except 1-NN). As discussed in Sections 4.3 and 4.4, the Alon and Golub datasets had the lowest MER curves for all the classifiers across the gene subset sizes for both the external CV and internal CV analyses, followed by the Pomeroy, Shipp, Singh, and Nutt datasets. Finally, one can see that the MER curves from the Shipp, Singh, and especially Nutt datasets were in general more variable than those of the other three datasets, for each of the classifiers.

Regarding the internal and external CV curves, as expected, the internal and external CV results for any given classifier and dataset combination were the same for the maximum possible number of top genes model. On the whole, it was interesting to see that the external CV results appeared to be slightly less variable than the internal ones across the subset sizes, for all classifiers and datasets. Also, one can see that the external CV error rates were greater than the internal ones, as the former were based on CV in which the feature selection process was also included in the CV procedure. It should be noted, though, that for all datasets but the Nutt one, the discrepancy between the external CV results and internal CV results for all classifiers

was not too substantials. Further, with external CV, selection bias was taken into account. This issue is discussed along with optimism bias further in the following section.

4.5.2 Optimism Bias, Selection Bias, and Total Bias

Figures 4.12, 4.13, and 4.14 illustrate how the optimism bias, selection bias, and "total bias" (optimism bias + selection bias) estimates, respectively, vary by classifier within dataset across gene subset sizes (see Equations 2.34, 2.35, and 2.37).

Figure 4.12: 10×10-Fold CV; Univ FSS; Optimism Bias vs. Gene Subset Size

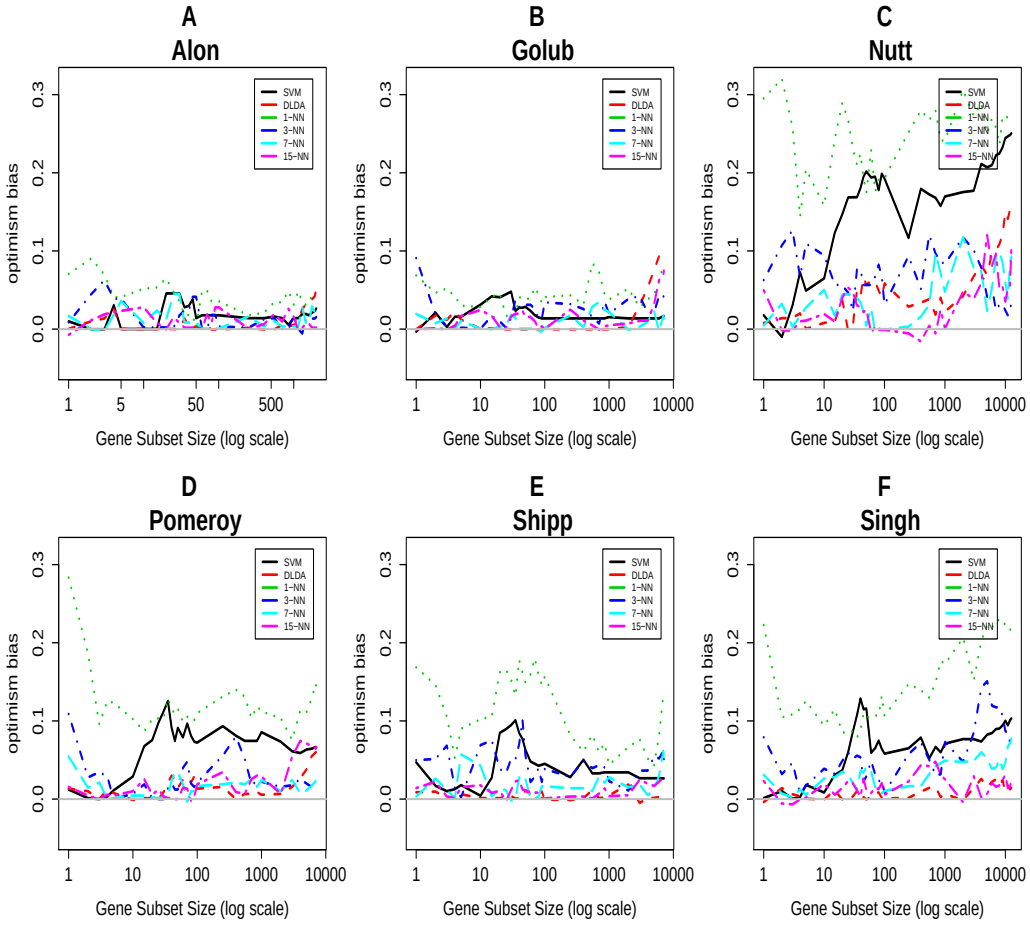


Figure 4.13: 10×10-Fold CV; Univ FSS; Selection Bias vs. Gene Subset Size

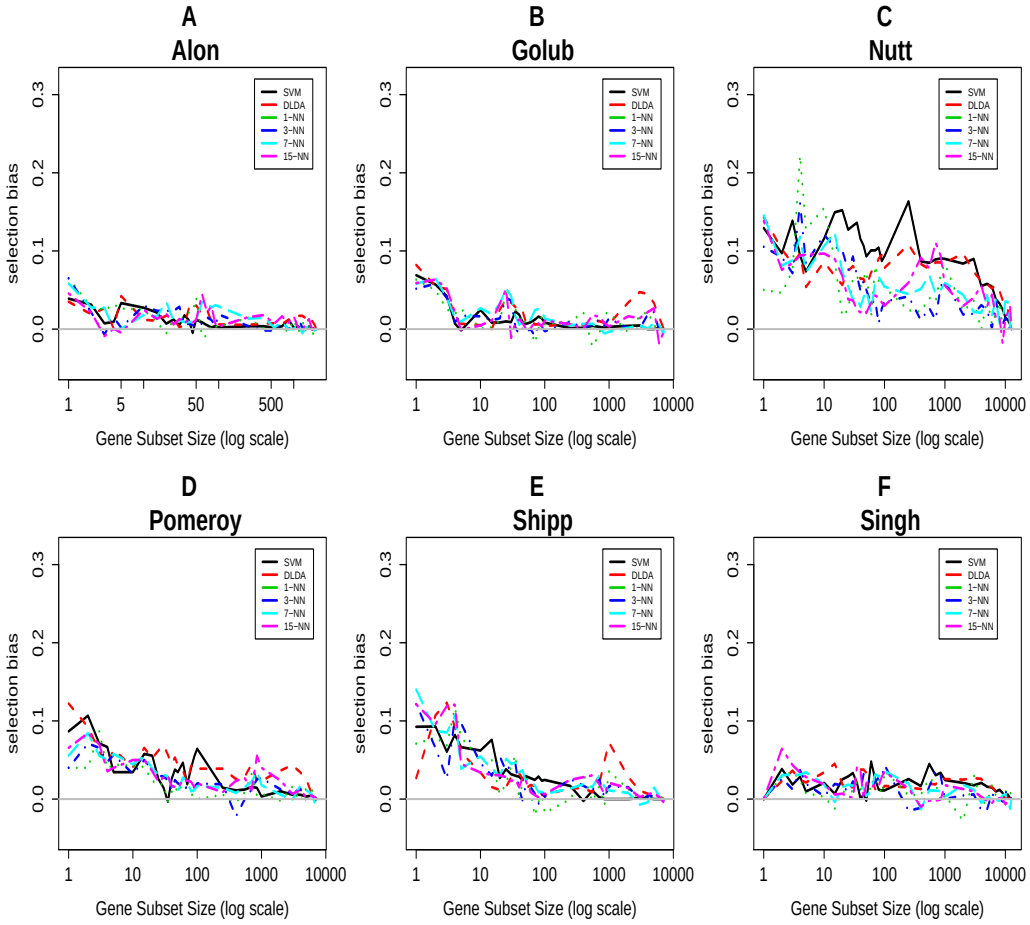
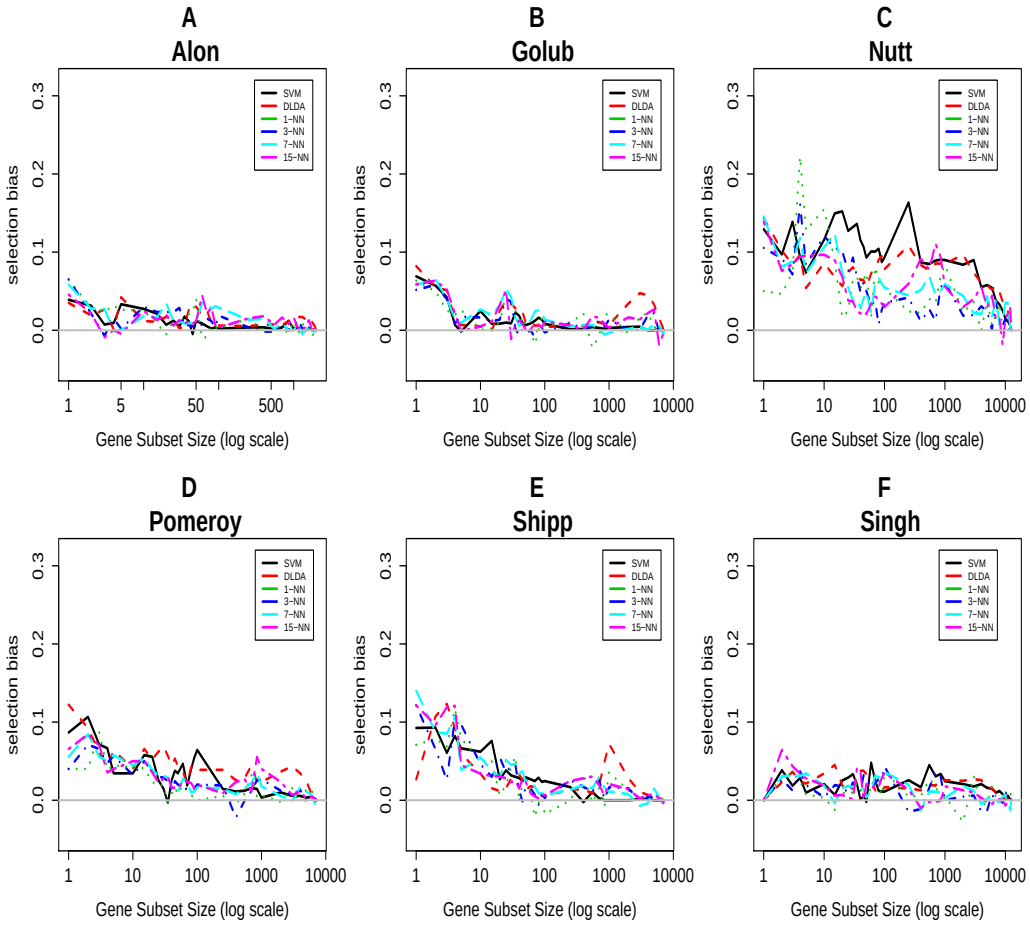


Figure 4.14: 10×10-Fold CV; Univ FSS; Total (Sel + Opt) Bias vs. Gene Subset Size



It should be observed from the plots in Figure 4.12 – 4.14 that the Alon and Golub datasets generally had smaller optimism, selection, and total bias values across the subset sizes than the other datasets, for all classifiers, while the Nutt dataset generally had the highest amount of all three bias values among all the classifiers, across subset sizes. No subset size emerged with significantly better (or worse) bias values across learning algorithms and datasets. With respect to the optimism bias plots in Figure 4.12, among the learning algorithms, DLDA yielded the lowest optimism bias values for all six datasets, while 1-NN generally led to the highest bias values. Since all the curves were predominantly positive across subset sizes and datasets, it was clear that there was at least some penalty in terms of higher MER when not using all the samples to both build the classifier and evaluate the classifier. With respect to the selection bias plots in Figure 4.13, it should be observed that the Alon, Singh, and Golub datasets were all marked by little selection bias across all subset sizes. The Nutt, Shipp, and Pomeroy datasets all had slightly higher selection bias, especially the Nutt dataset, whose selection bias curves for all classifiers were also much more variable than those of the other datasets. No particular learning algorithm emerged with significantly better (or worse) bias values across the subset sizes and datasets. In addition, the fact that all the curves were predominantly positive across subset sizes and datasets indicated there was at least some penalty in terms of higher MER when performing external 10-fold CV instead of the internal CV approach. Finally, looking at the "total bias" plots in Figure 4.14, the Nutt dataset again had the largest

total bias values. Overall, there was clearly some bias present for all the classification schemes, across all the datasets. To the extent that "total bias" measures overfit, the results indicate that overfitting was not a consistent function of the number of genes included.

Concluding this section is a table summarizing the means and standard deviations of each of the three bias estimates across gene subset sizes, for each dataset and learning algorithm combination. The empirical grand means across all gene subset sizes, datasets, and learning algorithms for each of the three bias estimates are provided in the last row of Table 4.1. Overall, considering all datasets, classifiers, and gene subset sizes together, the average optimism, selection, and total bias estimates were only 4.7%, 2.6%, and 7.3%, respectively. It should be noted that if the Nutt data were excluded, these averages became 3.6%, 1.9%, and 5.5%, respectively.

4.6 Final Thoughts

To begin with, for both the internal and external CV analyses, the repeated-run analyses generally led to less variable MER curves than did the single-run analyses. This finding was more evident with the external CV results than with the internal CV results. As far as whether any set of MER curves from the external CV results of Section 4.4 were lower than their corresponding results from the internal CV results of Section 4.3, the latter were perhaps only slightly better across gene subset sizes than their external CV counterparts, but not by a substantial amount at all. Also, it

should be noted that the internal CV MER curves were less stable than those from the external CV approach. Hence, having the feature selection process built into the CV procedure helped avoid the effect of selection bias, and not at the expense of significantly higher MER's. For both implementations of CV, the SVM classifier generally yielded the lowest and least variable MER curves across the datasets. With respect to datasets, the Alon and Golub datasets generally yielded the lowest error rates across all the learning algorithms, while the Nutt dataset was marked by the highest ones. This reinforces what was suggested in the MDS plots in Section 4.2.1. With respect to classifiers, the SVM classifier for the Golub, Shipp, and Singh results seemed to be significantly better than the other datasets for the largest subset sizes, while the DLDA classifier results for the Pomeroy, Shipp, and Singh datasets seemed to be significantly worse than the other datasets for the largest subset sizes. Based on the results of both the external and internal CV analyses, there appeared to be a dataset-learning algorithm interaction, as well as some interaction between learning algorithm and subset size and between dataset and subset size. Taking all three parameters together, the presence of a three-way interaction among them was evident. Considering each parameter individually, it appeared that dataset had the biggest main effect on the error rates, followed by the size of the top gene subset, with learning algorithm having the least effect.

Directly comparing the external and internal CV approaches, it was found that the selection bias estimates across the majority of the gene subset sizes were positive,

indicating there was at least some penalty, albeit not substantial, in terms of higher MER when performing external 10-fold CV instead of the internal CV approach, as expected. In addition, there seemed to be a bigger selection bias for the smaller sized gene subsets (i.e., especially sizes ≤ 10) than for the bigger sized subsets – an observation consistent from classifier to classifier. Only the Nutt dataset had noticeably higher selection bias estimates than those of the other datasets, but even then, on average it was 10%. Similarly, it was found for all datasets that there was a fair amount of optimism bias present among the classification rules used, as a result of using the same samples to both build the classifier as well as estimate the performance. Again though, the optimism bias estimates across all the learning algorithms and subset sizes were very small. Only the Nutt dataset had higher optimism bias estimates, especially with the 1-NN classifier (on average across all classifiers, though, only 6% for the Nutt data). To the extent that the "total bias" estimates measure overfit, the results indicate that overfitting is not a consistent function of the number of genes included in a given model.

Table 4.1: Optimism, Selection, & Total Bias Across All Subset Sizes; Univ FSS

Dataset	Learning Algorithm	Opt.Bias Mean (SD)	Sel.Bias Mean (SD)	Tot.Bias Mean (SD)
Alon	**SVM**	0.019 (0.013)	0.009 (0.011)	0.027 (0.013)
	DLDA	0.010 (0.015)	0.014 (0.010)	0.024 (0.015)
	1-NN	0.041 (0.018)	0.011 (0.016)	0.051 (0.025)
	3-NN	0.016 (0.016)	0.011 (0.014)	0.026 (0.019)
	7-NN	0.013 (0.013)	0.014 (0.015)	0.026 (0.016)
	15-NN	0.010 (0.010)	0.012 (0.012)	0.022 (0.014)
Golub	**SVM**	0.020 (0.011)	0.012 (0.016)	0.032 (0.018)
	DLDA	0.017 (0.033)	0.021 (0.020)	0.037 (0.037)
	1-NN	0.045 (0.018)	0.013 (0.019)	0.058 (0.020)
	3-NN	0.025 (0.018)	0.012 (0.017)	0.036 (0.024)
	7-NN	0.009 (0.010)	0.016 (0.018)	0.025 (0.020)
	15-NN	0.016 (0.019)	0.014 (0.019)	0.030 (0.019)
Nutt	**SVM**	0.166 (0.066)	0.087 (0.042)	0.253 (0.052)
	DLDA	0.056 (0.045)	0.070 (0.027)	0.126 (0.029)
	1-NN	0.243 (0.043)	0.050 (0.046)	0.293 (0.033)
	3-NN	0.066 (0.031)	0.047 (0.040)	0.113 (0.056)
	7-NN	0.043 (0.034)	0.052 (0.033)	0.095 (0.032)
	15-NN	0.031 (0.033)	0.050 (0.035)	0.081 (0.036)
Pomeroy	**SVM**	0.067 (0.032)	0.030 (0.028)	0.097 (0.026)
	DLDA	0.018 (0.017)	0.043 (0.024)	0.061 (0.021)
	1-NN	0.120 (0.036)	0.017 (0.020)	0.138 (0.043)
	3-NN	0.026 (0.023)	0.022 (0.020)	0.048 (0.028)
	7-NN	0.015 (0.012)	0.025 (0.021)	0.040 (0.022)
	15-NN	0.021 (0.024)	0.027 (0.021)	0.049 (0.023)
Shipp	**SVM**	0.043 (0.026)	0.028 (0.029)	0.071 (0.037)
	DLDA	0.004 (0.005)	0.028 (0.031)	0.032 (0.031)
	1-NN	0.114 (0.045)	0.024 (0.033)	0.138 (0.052)
	3-NN	0.044 (0.021)	0.027 (0.032)	0.071 (0.034)
	7-NN	0.019 (0.017)	0.031 (0.034)	0.050 (0.034)
	15-NN	0.012 (0.009)	0.029 (0.035)	0.041 (0.034)
Singh	**SVM**	0.068 (0.034)	0.017 (0.013)	0.085 (0.032)
	DLDA	0.008 (0.008)	0.019 (0.012)	0.027 (0.012)
	1-NN	0.151 (0.054)	0.010 (0.014)	0.161 (0.049)
	3-NN	0.060 (0.036)	0.012 (0.015)	0.072 (0.026)
	7-NN	0.034 (0.019)	0.012 (0.014)	0.046 (0.013)
	15-NN	0.017 (0.013)	0.012 (0.017)	0.029 (0.013)
Grand Avg		**0.047 (0.024)**	**0.026 (0.023)**	**0.073 (0.028)**

Chapter 5

Multivariate-Based FSS Results

5.1 Introduction

Much of the work done to date with respect to binary classification of microarray data has been based on univariate-based feature selection approaches. Of course, simple univariate gene ranking approaches would suffice if, among the thousands of genes, there exist a number of dominant (individually predictive) genes. However, based on the results discussed in Section 3.4, I believe there is definitely cause to further investigate the merits of a multivariate, modular approach to feature subset selection for binary classification with high-dimensional data from microarrays. Biologically, it is known that genes work together, so it would at least seem reasonable to consider performing feature selection in a multivariate manner. At the same time, though, it should be noted that biological interaction among genes does not necessarily imply

that the genes will be jointly predictive in a classification problem. An evolutionary algorithm such as the genetic algorithm has the advantage of being able to find at least near-optimal solutions (i.e., subsets of jointly significant genes) to use for the classification of microarray data in a setting in which there are simply too many genes to find optimal combinations of genes through a completely exhaustive search of the high-dimensional feature space. Further, with univariate screening to form a list of genes based on their individual predictive power, it is suspected that there would be solutions that would not be found by combining top individually predictive genes formed from the list (as discussed in Sections 2.4.1 and 3.3).

This chapter includes misclassification error rates obtained using a multivariate-based means of feature subset selection, the genetic algorithm, in both a single- and two-stage setting, in conjunction with six supervised learning algorithms (SVM, DLDA, and k-NN ($k = 1, 3, 7, 15$)). Both internal and external 10-fold cross-validation were used to estimate the predictive accuracy of the various classification processes. These analyses are done with respect to various gene subset sizes: 1, 2, 3, 4, 5, 10, 15, 20, and 25). The same six published microarray datasets and preprocessing used for the CV analyses with univariate-based feature subset selection are used in this chapter (refer to Section 4.2 for more details on the datasets). All analyses of this chapter were performed using the *R* statistical software package [35] on a Red Hat Linux machine (dual Intel(R) 3.06 GHz processors and 4 GB memory).

5.1.1 Internal CV, External CV, and Repeated Runs with the GA

One of the more interesting aspects with respect to the results of this chapter is that half of them are based on implementing the GA in a manner not before investigated in the microarray classification and feature sleection litearture. That is, the GA feature selection processes are built into each stage of a 10-fold CV process (external CV) such that the effect of selection bias can be taken into account, as was successfully done with the univariate feature selection approach of Chapter 4. The other half of the results are based on an internal CV approach, in which the GA is performed 'up-front', prior to the CV process, using all the samples of each datset. For more motivation on these two approaches to CV, refer to Section 4.1.3. 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 (discussed further in Section 5.1.3), which provides the user with the opportunity to run the entire GA multiple (10) times for each microarray dataset. In the case of external CV, the GA is applied 10 times for each of the 10 training subsets of samples corresponding to each stage of the 10-fold CV process. The same 10 training subsets for each of the datasets were used across learning algorithms to maintain consistency in interpreting the 10-fold CV results among the classifiers.

5.1.2 Single- and Two-Stage GA-Based Approaches

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 [17, 20, 27]. Both a single-stage and a two-stage GA feature selection process was implemented.

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 was large enough to ensure convergence of the 1000 solutions. For the two-stage approach ("GA-GA"), the first GA stage 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 d-gene solutions selected by each implementation of the GA was reduced to 500, since the initial gene pool was reduced considerably (number of generations remained at 250).

5.1.3 *Genalg* Files and Parameterization

The code used is a *C++* program named *genalg* from the Department of Applied Mathematics and Biostatistics at the University of Texas M.D. Anderson Cancer Center [8]. The GA portions of this research were run at M.D. Anderson on Condor, a specialized workload full-featured batch management software system of roughly 80 machines used for handling multiple computationally intensive jobs in parallel (on Windows-based platforms). In this program, the fitness function used to evaluate the candidate d-gene solutions is mahalanobis distance (refer to Equation 2.10). For more information on the *genalg* software, the reader should refer to [8].

The GA software requires a couple of parameterization files in addition to the datafiles – a meta file and a driver file. The meta file, which only needs to be created once for each dataset the GA is to be run on, provides basic information about the dataset as described below:

- <numPeaks> = number of rows p in the dataset (i.e., the number of genes; varies by dataset)
- <numSamples> = number of columns N in the dataset (i.e., the number of samples; varies by dataset)
- <classVector> = a binary vector of length N denoting which samples are of one class and which are of the other (e.g., cancer (1) and normal (0); varies by dataset)
- <dataFileName> = dataset that this meta file describes (tab-delimited dataset with no header files; p rows and N columns),

Parameters to be specified for the implementation of the genetic algorithm are specified in the driver file as follows (if possible, values used in this research are given in parentheses following each parameter):

- <numRuns> = number of iterations to run the genetic algorithm (10)
- <numIndividuals> = population size; i.e., number of d-length solutions generated per generation (1000 for first stage of GA and 500 for second stage of GA)
- <numFeatures> = number of genes d to consider for each solution of the population during the GA runs ($d = 1 : 5$,10, 15, 20, 25), where d is specified by this parameter

- <numGenerations> = number of generations to run for each iteration of the GA (250; large enough to have convergence of solutions)
- <probMutation> = probability in each generation of a feature being randomly changed (0.001; standard value assigned to this parameter in GA applications)
- <outputFilePrefix> = prefix that every output file from a given run will begin with (varies by dataset)
- <outputFrequency> = output will be written to file every g generations, where g is specified by this parameter (the first and last generations are always written to file) (10)
- <metaFileName> = name of meta file (varies by dataset)

5.1.4 Plot Breakdowns

All results shown in this section are based on error rates obtained from 10-fold external and internal cross-validation, with the single-stage and two-stage GA processes serving as the feature selection mechanisms. The same variety of plots generated for the error rates based on univariate feature selection are also presented here. These results include resubstitution, external CV, and internal CV curves plotted against gene subset size, as well as plots investigating the effect of both optimism and selection bias as a function of subset size. The choice for gene subset sizes to use for all datasets was made taking into account the feasibility of the computation that would

be involved in running the GA as well as the desire to maintain relatively small gene subset sizes. In this work, no subsets larger than 25 genes were considered, as preliminary investigation showed no significant increase nor decrease in error rates for these sizes, for all datasets and learning algorithms. This choice of subset size was further supported by the univariate-based results of Chapter 4, in which the advantages gained by having gene subset sizes larger than about 25 were not significantly better, at the expense of having a more complicated model in terms of higher number of genes).

5.2 Resubstitution, External & Internal CV, & Selection & Optimism Bias

Similar to what was done in Chapter 4, in this section, a direct comparison of the 10-fold external and internal cross-validation results for both the single-stage and two-stage GA-based feature selection approaches is provided in this section. Also included are results in which the differences among the resubstitution errors, external CV errors, and internal CV errors are computed as a function of subset size, as a way of assessing the optimism, selection, and total bias incurred.

5.2.1 Resubstitution, Internal CV, and External CV MER's

The results of this section compare the 10-fold external and internal CV error rates, as well as the resubstitution errors, across gene subset sizes for both the single-stage and two-stage GA-based feature selection approaches.

Figure 5.1: 10-Fold CV; Internal CV vs. External CV; 1- and 2-Stage GA FSS; Alon Data

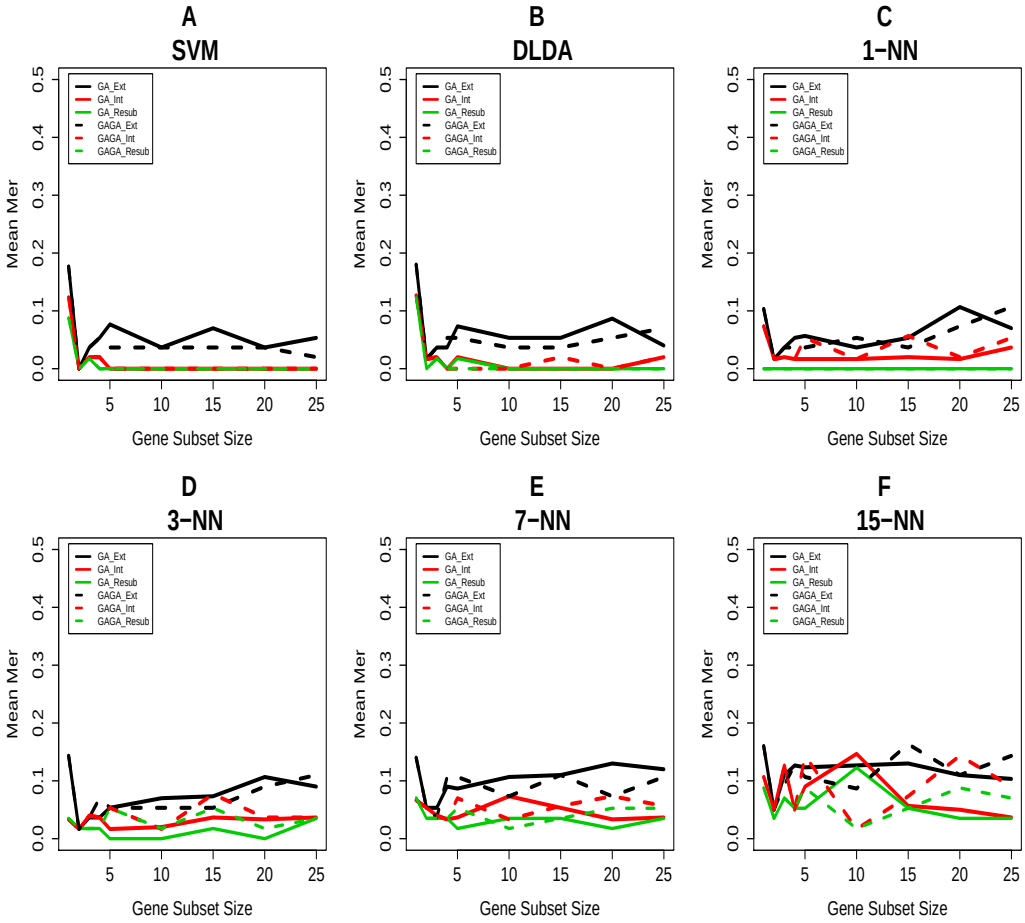


Figure 5.2: 10-Fold CV; Internal CV vs. External CV; 1- and 2-Stage GA FSS; Golub Data

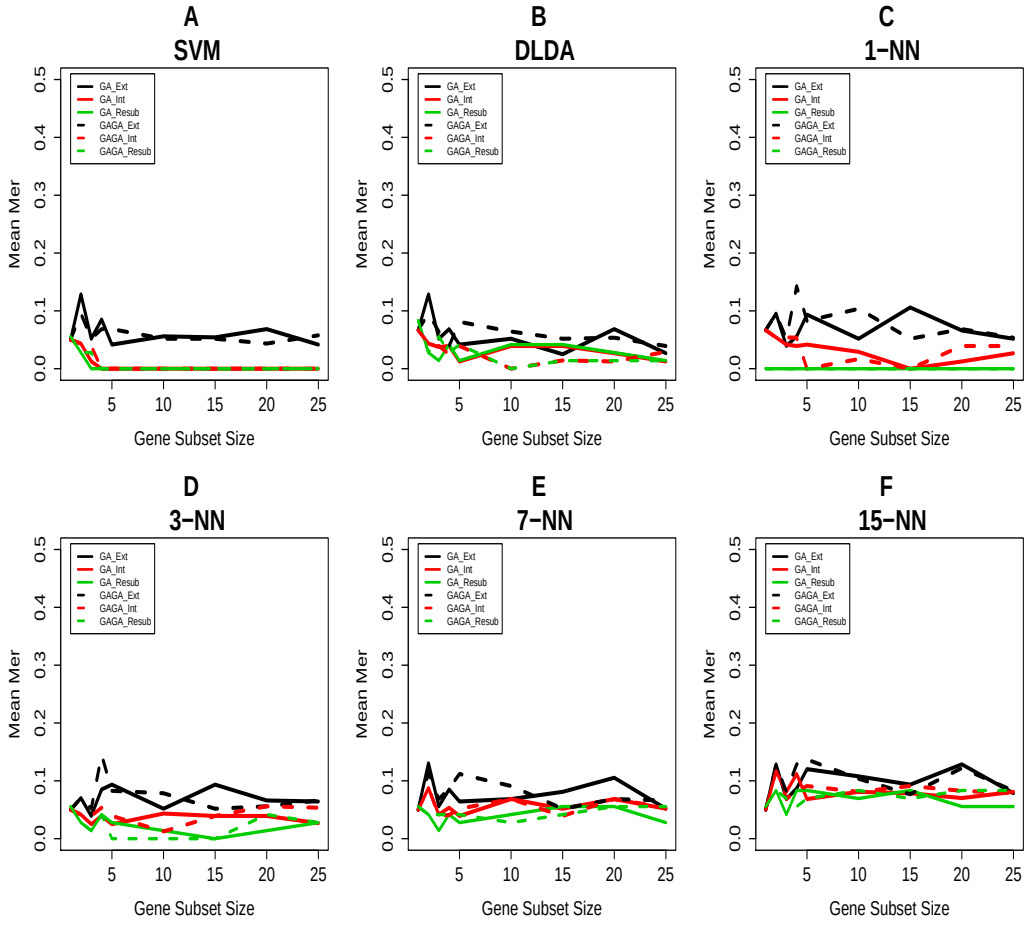


Figure 5.3: 10-Fold CV; Internal CV vs. External CV; 1- and 2-Stage GA FSS; Nutt Data

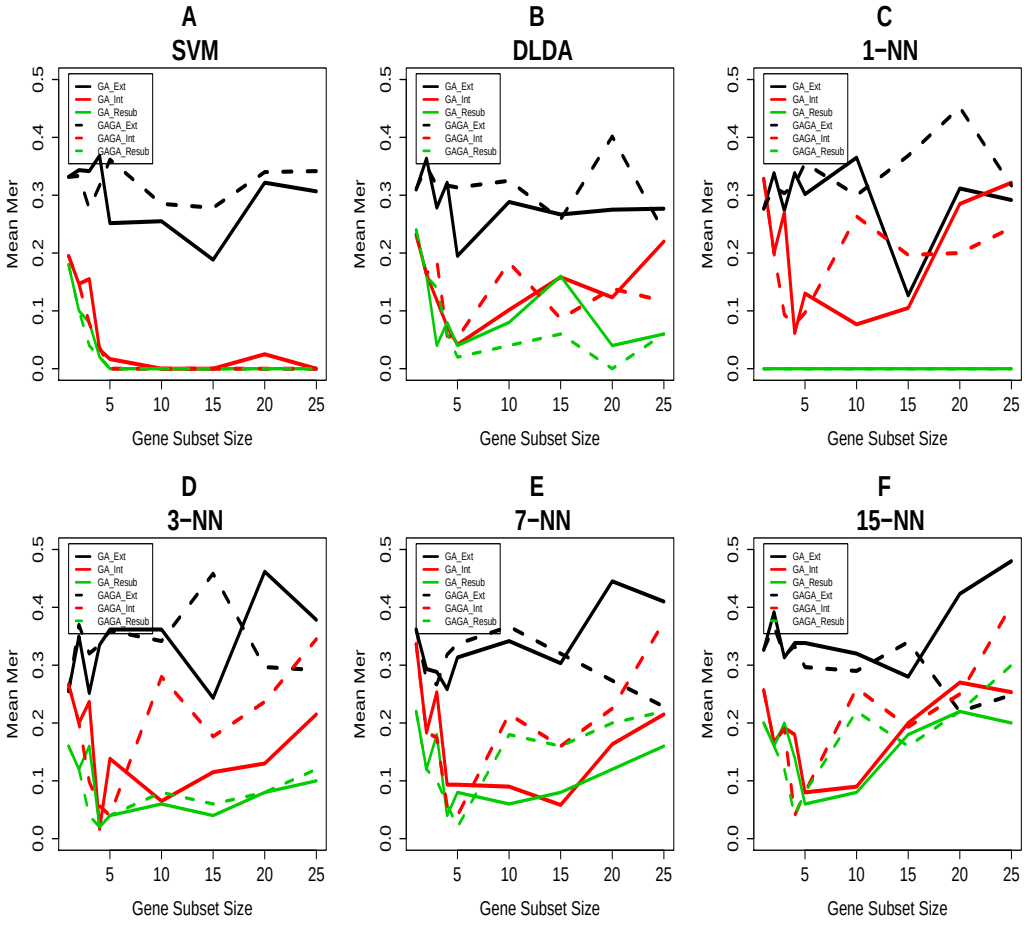


Figure 5.4: 10-Fold CV; Internal CV vs. External CV; 1- and 2-Stage GA FSS; Pomeroy Data

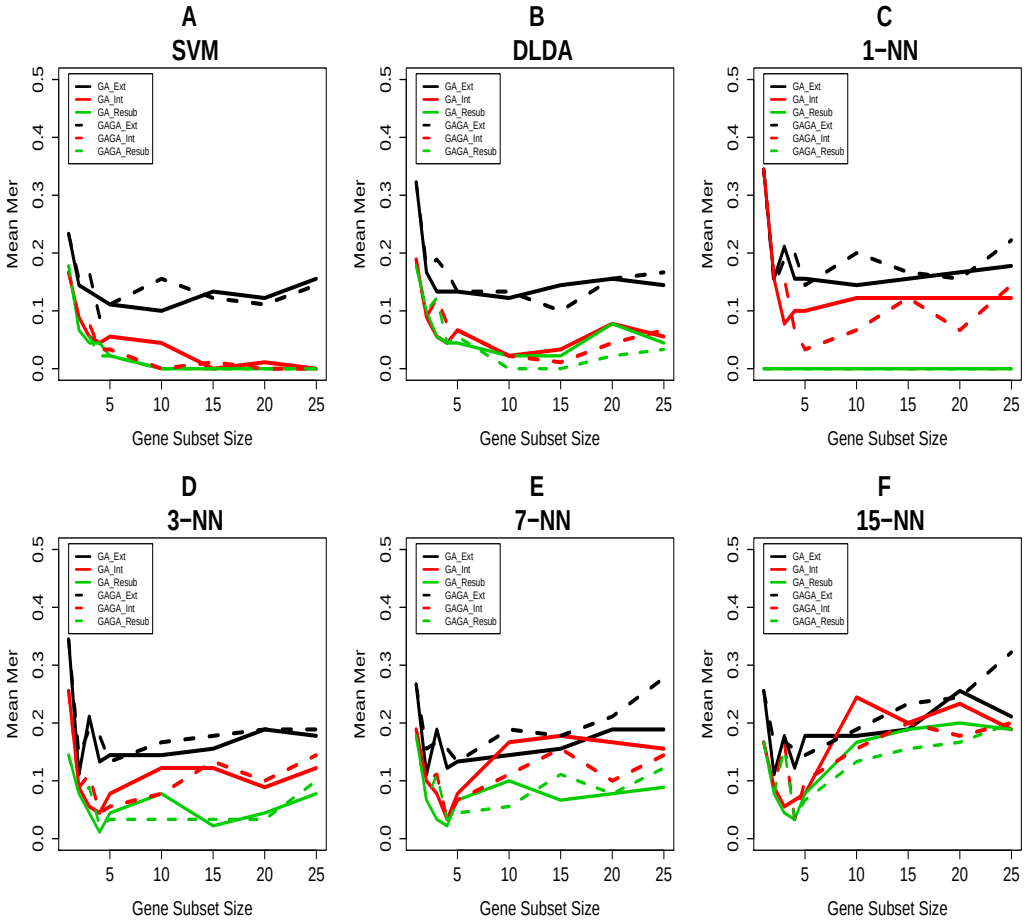


Figure 5.5: 10-Fold CV; Internal CV vs. External CV; 1- and 2-Stage GA FSS; Shipp Data

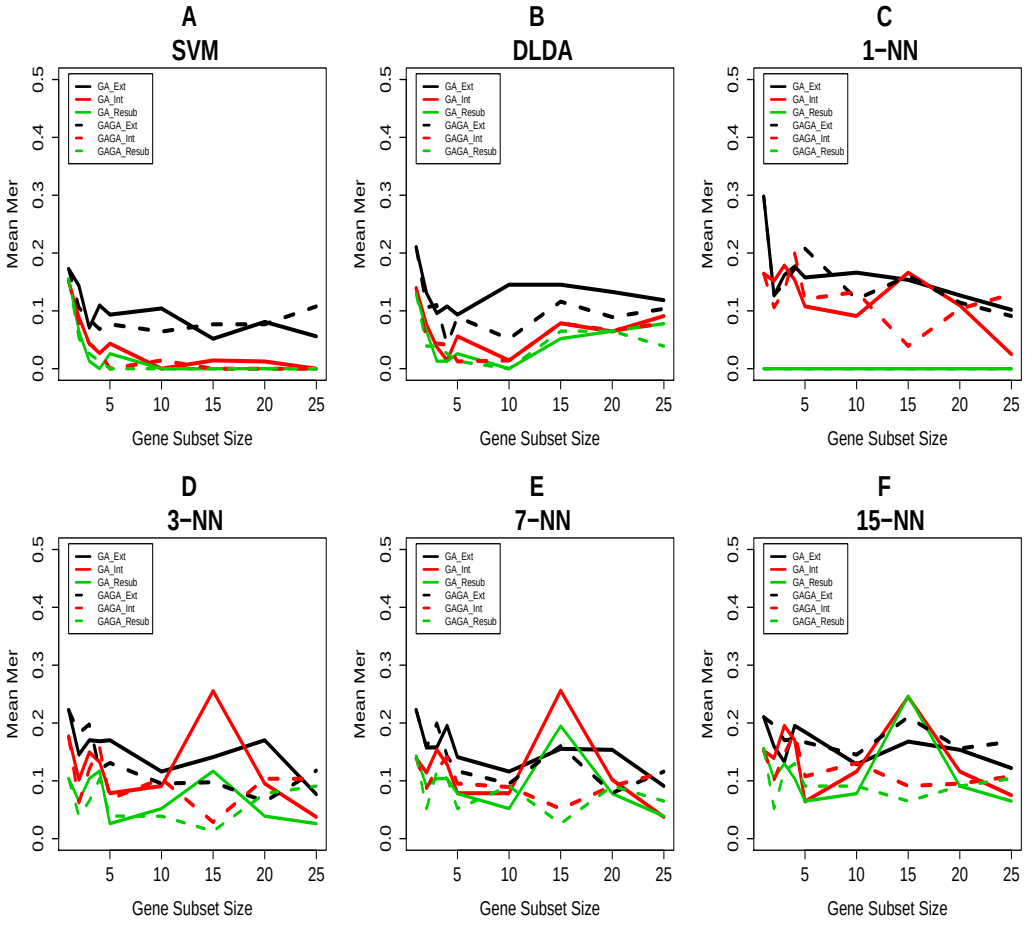
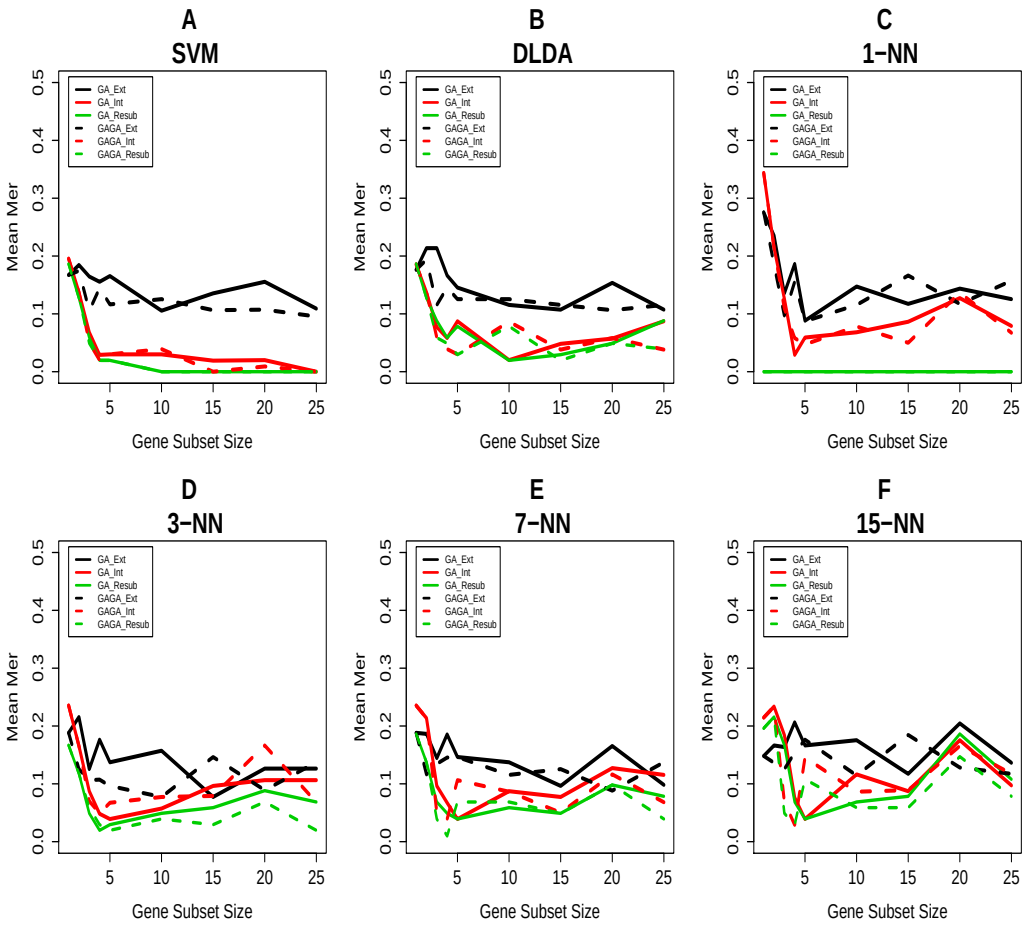


Figure 5.6: 10-Fold CV; Internal CV vs. External CV; 1- and 2-Stage GA FSS; Singh Data



From the plots, one can see that the resubsitution error curves for the 1-NN classifier were always 0, as one would expect. Focusing on the datasets, one should note that the Alon and Golub results possessed the lowest MER's in general across subset sizes (mostly in the range of only 0.05 and 0.10 for the external results; around 0.05 for the internal ones), while the Nutt dataset had the highest MER's (mostly between 0.30 and 0.50 for the external CV results; between 0.10 and 0.35 for the internal results). Also, there seemed to be slightly more variation among the CV curves and even the resubstitution curves across subset sizes for the Nutt data than in the other five datasets. The other three datasets were quite comparable in terms of their error rates across subset sizes. With respect to the single-stage vs. two-stage GA procedure, the more complicated two-stage one surprisingly did not offer a noticeable advantage in terms of lower error rates over the simpler single-stage one. In terms of learning algorithms, it was not totally clear from these plots that one particular algorithm was significantly better than any of the others across all the datasets. As far as what size gene subset was best, this was not immediately clear either, as it varied from classifier to classifier across subset sizes, with no dominant upward or downward trend with increasing size, for each dataset.

Similar to what was found from the univariate-based feature selection results of Chapter 4, there appeared to be a dataset-learning algorithm interaction (i.e., for any given number of genes, the rank ordering of the learning algorithms varied from dataset to dataset), an interaction between learning algorithm and gene subset size

(i.e., for any dataset, the rank ordering of learning algorithms varied from gene subset size to gene subset size), and perhaps minimal interaction between dataset and gene subset size may have been present (i.e., for any learning algorithm, the rank ordering of the datasets varied only slightly from gene subset size to gene subset size). Looking at all three of these experimental parameters together, it would seem that a three-way interaction among them exists, since the effect of the number of genes used for classification seemed to be dependent on both the classfier and the dataset. Interactions aside, it appears that dataset had the biggest main effect on the error rates, followed by the size of the gene subset size, with classifier apparently having the least effect.

Finally, focusing on the internal and external CV curves for all the datasets, one can see that the external CV error rates were greater than the internal ones, as expected. However, for all datasets but the Nutt one, the discrepancy between the external and internal CV results for all classifiers was not too substantial. Furthermore, in using external CV, not only were the error rates comparable, selection bias was taken into account, providing for more honest estimates of the misclassification error rates. The effects of optimism and selection bias are investigated further in the following section.

5.2.2 Optimism Bias, Selection Bias, and Total Bias

For the single-stage GA results, Figures 5.7, 5.9, and 5.11 illustrate how the optimism bias, selection bias, and "total bias" (optimism bias + selection bias) estimates, respectively, vary by classifier within dataset across gene subset sizes (see Equations 2.34, 2.35, and 2.37). The analogous optimism, selection, and total bias plots for the double-stage GA results are shown in Figures 5.8, 5.10, and 5.12, respectively.

Figure 5.7: 10-Fold CV; 1-Stage GA FSS; Optimism Bias vs. Gene Subset Size

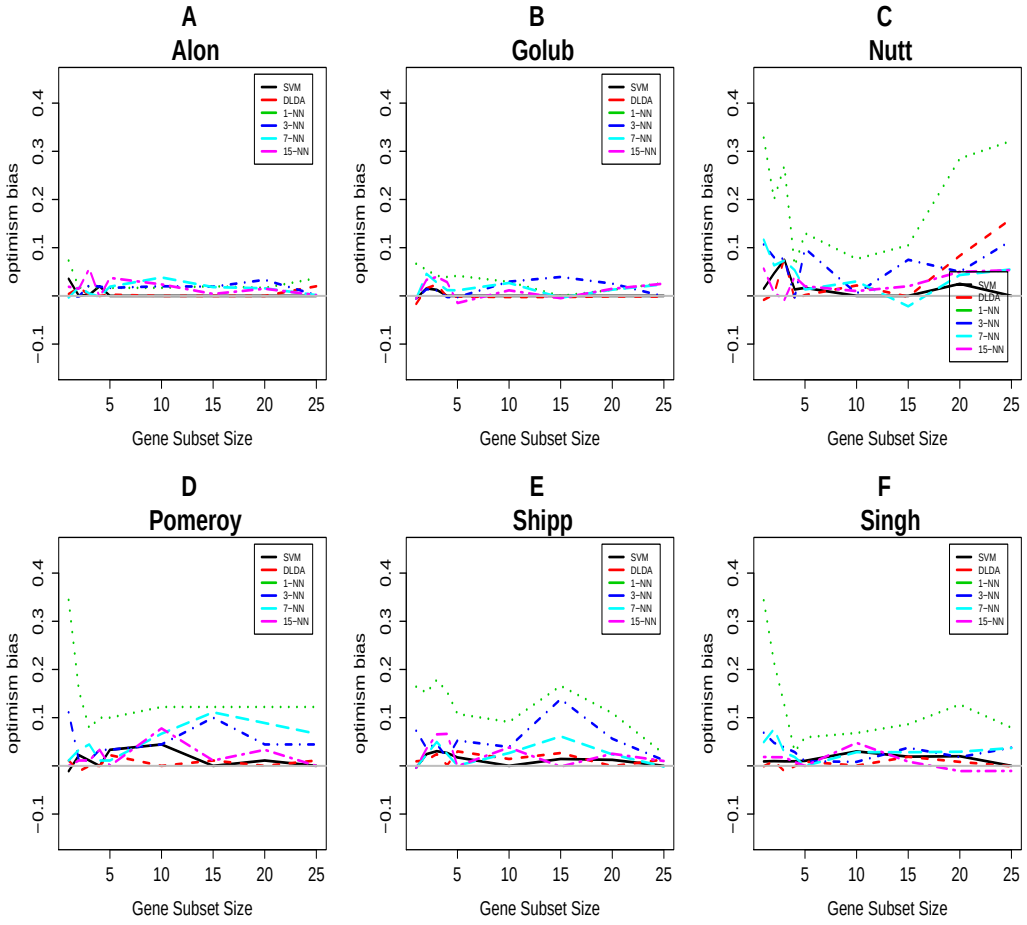


Figure 5.8: 10-Fold CV; 2-Stage GA FSS; Optimism Bias vs. Gene Subset Size

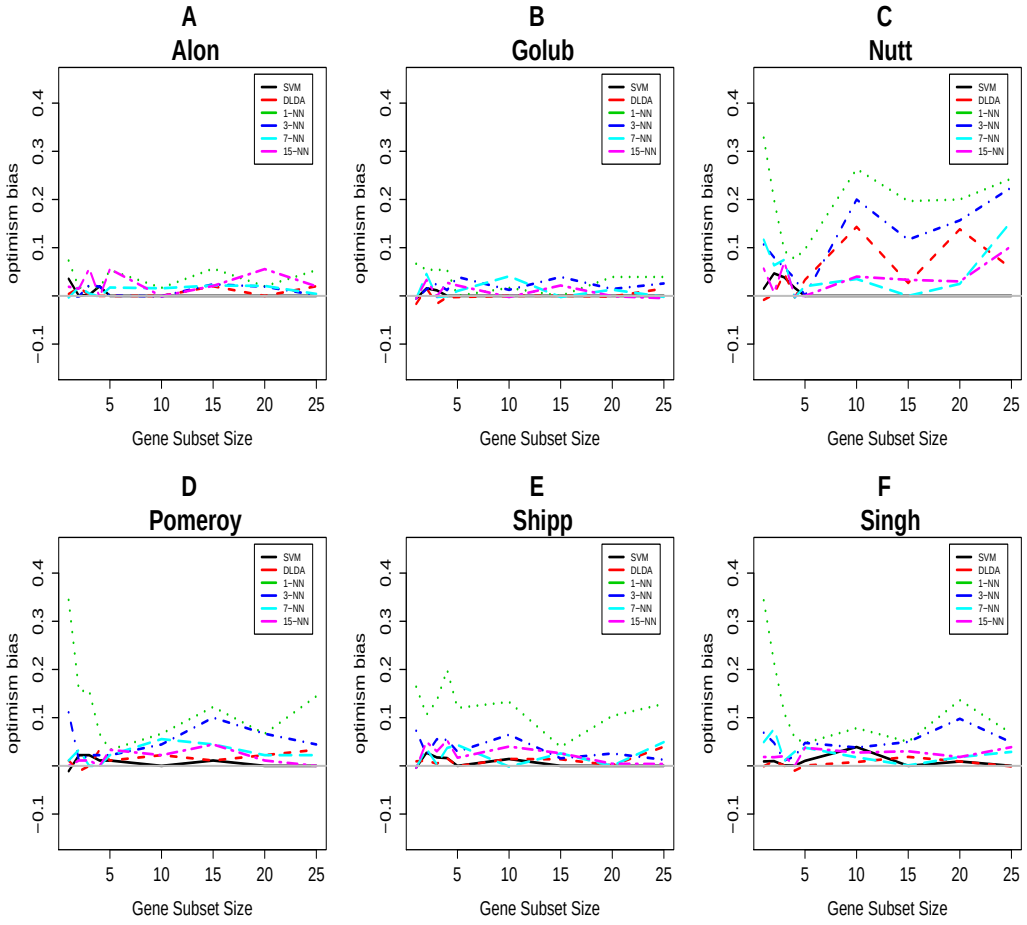


Figure 5.9: 10-Fold CV; 1-Stage GA FSS; Selection Bias vs. Gene Subset Size

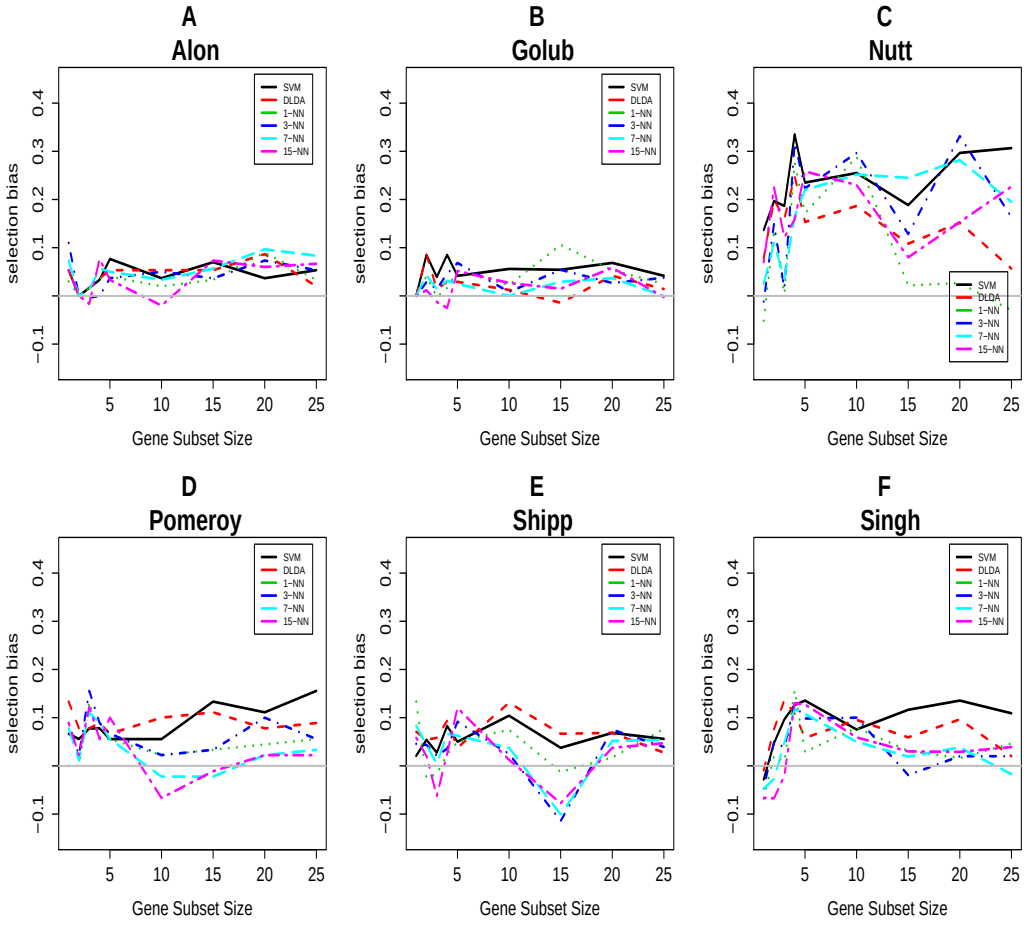


Figure 5.10: 10-Fold CV; 2-Stage GA FSS; Selection Bias vs. Gene Subset Size

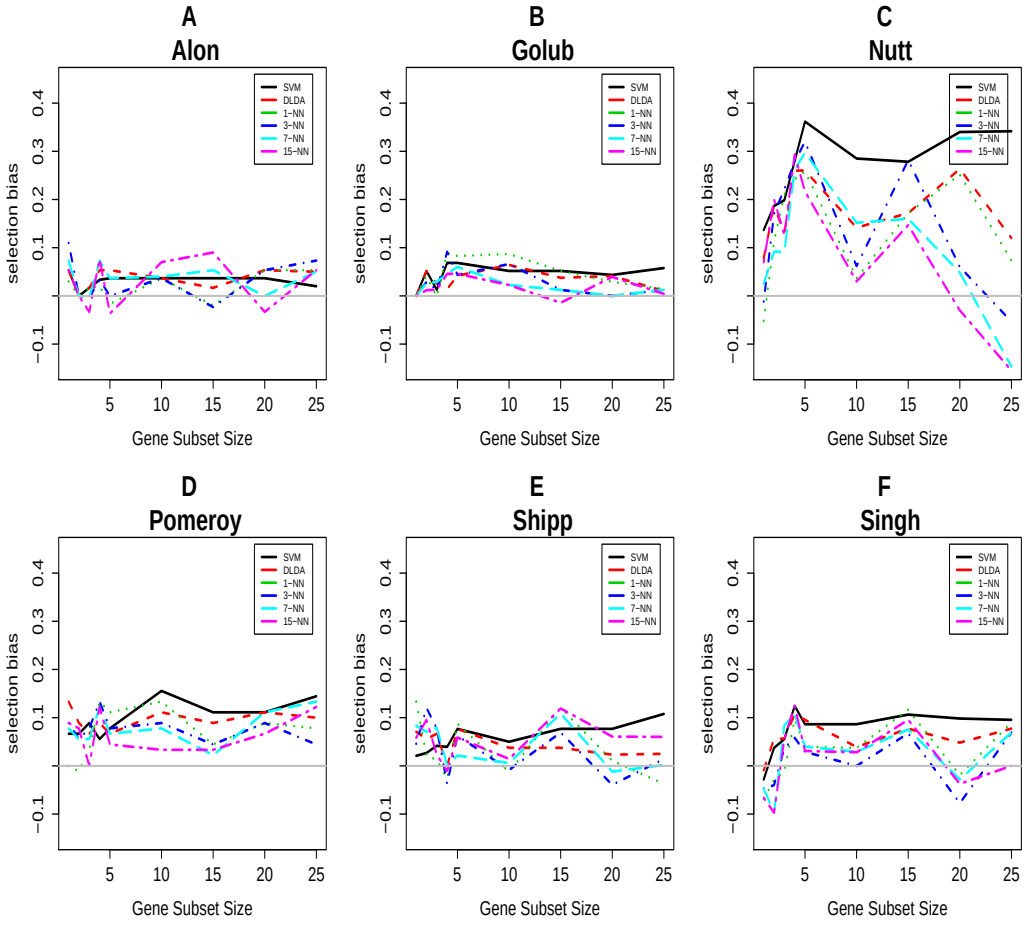


Figure 5.11: 10-Fold CV; 1-Stage GA FSS; Total (Sel + Opt) Bias vs. Gene Subset Size

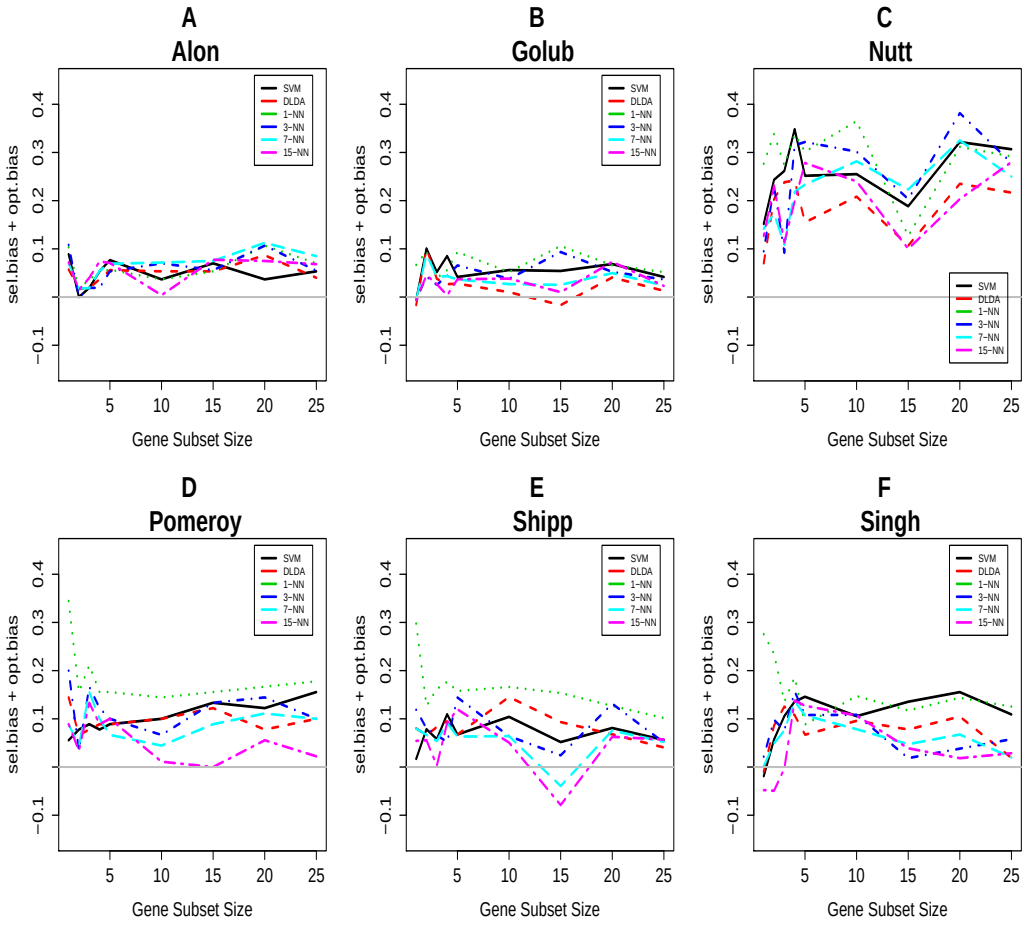
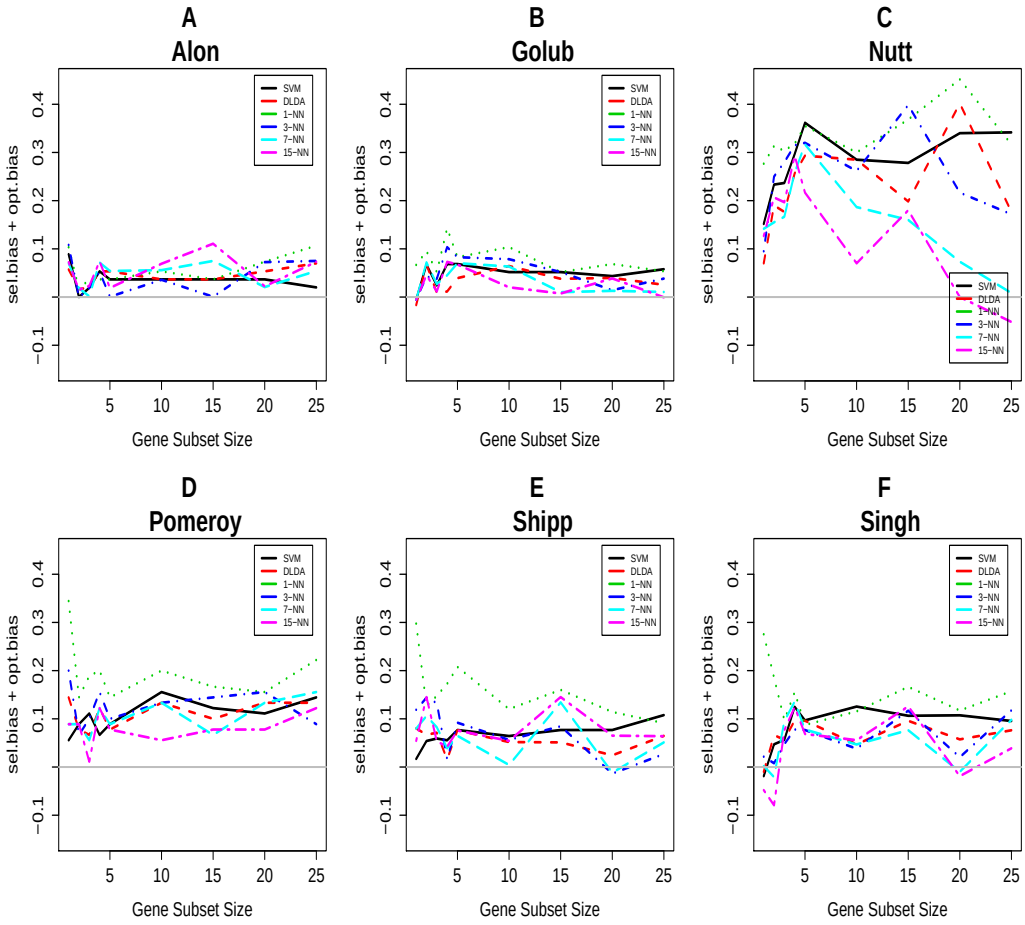


Figure 5.12: 10-Fold CV; 2-Stage GA FSS; Total (Sel + Opt) Bias vs. Gene Subset Size



In general, regardless of whether single- or double-stage GA was used as the feature selection process, it should be noted from Figures 5.7 – 5.12 that all datasets except the Nutt one possessed very little optimism, selection, and total bias across all gene subset sizes. Further, the bias curves for all classifiers of the Nutt dataset were also much more variable than those of the other datasets. The Alon and Golub datasets generally had smaller values of all three bias values for all the classifiers, as was the case with the univariate-based feature selection approaches. No particular subset size emerged with significantly better (or worse) bias values across learning algorithms and datasets. Focusing on the optimism bias plots of Figures 5.7 and 5.8, among the learning algorithms, DLDA was among the lowest bias curves for all six datasets, while 1-NN generally led to the highest bias values. As seen with the univariate-based results of Chapter 4, the optimism bias values were predominantly positive across subset sizes and datasets, indicating there was at least some penalty in terms of higher MER when not using all the samples to both build the classifiers and evaluate the classifiers. For the selection bias plots of Figures 5.8 and 5.9, it should be noted that no particular learning algorithm emerged with significantly better (or worse) bias values across the subset sizes and datasets. The fact that all the curves were predominantly positive across subset sizes and datasets indicated there was at least some penalty in terms of higher MER when performing external 10-fold CV instead of the internal CV approach. These findings were also evident in the univariate-based feature selection results of Chapter 4. Finally, looking at the "total

bias" plots in Figures 5.11 and 5.12, the Nutt dataset again had the largest total bias values. Overall, as was the case with the univariate results, there was clearly some optimism and selection bias present for all the classification rules and across all the datasets. To the extent that "total bias" measures overfit, the results indicate that overfitting was not a consistent function of the number of genes included.

Concluding this section are a couple of tables summarizing the means and standard deviations of the optimism, selection, and total bias estimates across gene subset sizes, for each combination of dataset and learning algorithm. Table 5.1 contains the single-stage GA-based feature selection results, while Table 5.2 contains the results from the two-stage GA feature selection approach. The empirical grand means across all subset sizes, datasets, and learning algorithms for each of the three bias estimates are provided in the last row of each of the tables. Overall, considering all datasets, classifiers, and gene subset sizes together, the average optimism, selection, and total bias estimates for the GA-based results were only 3.6%, 6.5%, and 10.1%, respectively. For the two-stage GA-based results, these averages were virtually identical at 3.7%, 6.3%, and 10.0%, respectively. It should be noted that if the Nutt data were excluded, these averages became 3.1%, 4.5%, and 7.5% for the single-stage GA results and 2.9%, 4.4%, and 7.4% for the two-stage GA results.

5.3 Final Thoughts

For both implementations of GA-based feature selection, for all datasets except the Nutt one, the discrepancy between the external CV results and internal CV results for all classifiers was not very substantial at all. External CV, however, allowed for selection bias to be taken into account and hence provided for more honest estimates of the misclassification error rates. Among the datasets, the Alon and Golub results possessed the lowest CV MER's in general across subset sizes, while the Nutt dataset had the highest MER's. This reinforces what was shown in the MDS plots in Section 4.2.1. Also, with the Nutt dataset, there seemed to be more variation among the CV and resubstitution curves across subset sizes than in the other five datasets. The other three datasets were quite comparable in terms of their error rates across subset sizes. Considering the two GA-based feature selection approaches, the more complicated two-stage one did not offer a noticeable advantage over the single-stage one. In terms of learning algorithms, no particular algorithm emerged as significantly better than any of the others across all the datasets. As far as what subset size was best, this was not immediately clear, as it varied from classifier to classifier across subset sizes for each of the datasets.

Based on the internal and external CV results, there appeared to be a dataset-classifier interaction, as well as some interaction between classifier and gene subset size and between dataset and subset size. Taking all three parameters together, the presence of a three-way interaction among them was evident. Considering each

parameter individually, it appeared that dataset had the biggest main effect on the error rates, followed by subset size and classifier. Further investigating internal CV, external CV, and resubstitution errors, the optimism and selection bias estimates across the majority of the gene subset sizes were positive, indicating there was at least some penalty in terms of a) higher MER when not using all samples to both train and evaluate a given classifier, and b) performing external 10-fold CV instead the internal CV approach. A direct comparison of the repeated-run internal and external 10-fold CV error rates based on the univariate feature selection and the internal and external 10-fold CV results based on the two GA feature selection approaches are provided in the next chapter.

Table 5.1: Optimism, Selection, & Total Bias Across All Subset Sizes; 1-Stage GA

Dataset	Learning Algorithm	Opt.Bias Mean (SD)	Sel.Bias Mean (SD)	Tot.Bias Mean (SD)
Alon	**SVM**	0.006 (0.013)	0.042 (0.024)	0.048 (0.028)
	DLDA	0.005 (0.008)	0.041 (0.026)	0.047 (0.021)
	1-NN	0.026 (0.019)	0.033 (0.025)	0.059 (0.030)
	3-NN	0.014 (0.012)	0.040 (0.038)	0.054 (0.038)
	7-NN	0.012 (0.013)	0.051 (0.032)	0.064 (0.030)
	15-NN	0.019 (0.019)	0.036 (0.039)	0.055 (0.028)
Golub	**SVM**	0.003 (0.007)	0.052 (0.026)	0.055 (0.030)
	DLDA	0.001 (0.012)	0.024 (0.028)	0.025 (0.035)
	1-NN	0.035 (0.020)	0.035 (0.033)	0.070 (0.023)
	3-NN	0.012 (0.016)	0.032 (0.022)	0.043 (0.027)
	7-NN	0.017 (0.016)	0.020 (0.017)	0.037 (0.025)
	15-NN	0.014 (0.019)	0.014 (0.028)	0.028 (0.024)
Nutt	**SVM**	0.021 (0.025)	0.237 (0.066)	0.259 (0.062)
	DLDA	0.037 (0.058)	0.149 (0.060)	0.186 (0.062)
	1-NN	0.198 (0.107)	0.094 (0.129)	0.291 (0.069)
	3-NN	0.067 (0.042)	0.179 (0.125)	0.246 (0.101)
	7-NN	0.047 (0.039)	0.170 (0.094)	0.217 (0.068)
	15-NN	0.028 (0.023)	0.169 (0.069)	0.197 (0.069)
Pomeroy	**SVM**	0.012 (0.018)	0.088 (0.037)	0.100 (0.031)
	DLDA	0.005 (0.010)	0.091 (0.021)	0.096 (0.024)
	1-NN	0.141 (0.079)	0.044 (0.040)	0.185 (0.063)
	3-NN	0.048 (0.035)	0.070 (0.043)	0.119 (0.051)
	7-NN	0.049 (0.036)	0.040 (0.048)	0.089 (0.035)
	15-NN	0.020 (0.025)	0.040 (0.059)	0.059 (0.046)
Shipp	**SVM**	0.014 (0.013)	0.056 (0.027)	0.069 (0.028)
	DLDA	0.014 (0.011)	0.068 (0.031)	0.082 (0.029)
	1-NN	0.128 (0.049)	0.036 (0.053)	0.163 (0.056)
	3-NN	0.052 (0.038)	0.029 (0.058)	0.081 (0.041)
	7-NN	0.023 (0.022)	0.034 (0.055)	0.056 (0.038)
	15-NN	0.026 (0.027)	0.020 (0.060)	0.046 (0.056)
Singh	**SVM**	0.013 (0.009)	0.090 (0.053)	0.103 (0.054)
	DLDA	0.004 (0.009)	0.071 (0.045)	0.075 (0.044)
	1-NN	0.126 (0.098)	0.036 (0.060)	0.162 (0.060)
	3-NN	0.033 (0.019)	0.043 (0.058)	0.076 (0.046)
	7-NN	0.033 (0.021)	0.032 (0.057)	0.065 (0.041)
	15-NN	0.011 (0.018)	0.029 (0.072)	0.040 (0.071)
Grand Avg		0.036 (0.028)	0.065 (0.049)	0.101 (0.044)

Table 5.2: Optimism, Selection, & Total Bias Across All Subset Sizes; 2-Stage GA

Dataset	Learning Algorithm	Opt.Bias Mean (SD)	Sel.Bias Mean (SD)	Tot.Bias Mean (SD)
Alon	**SVM**	0.006 (0.013)	0.030 (0.015)	0.036 (0.025)
	DLDA	0.007 (0.009)	0.037 (0.021)	0.044 (0.018)
	1-NN	0.036 (0.023)	0.019 (0.027)	0.056 (0.032)
	3-NN	0.009 (0.011)	0.031 (0.043)	0.040 (0.039)
	7-NN	0.011 (0.010)	0.036 (0.030)	0.047 (0.027)
	15-NN	0.027 (0.024)	0.027 (0.052)	0.053 (0.034)
Golub	**SVM**	0.002 (0.007)	0.045 (0.024)	0.048 (0.024)
	DLDA	-0.001 (0.011)	0.032 (0.021)	0.031 (0.027)
	1-NN	0.036 (0.025)	0.044 (0.036)	0.080 (0.030)
	3-NN	0.020 (0.014)	0.030 (0.032)	0.049 (0.034)
	7-NN	0.011 (0.019)	0.023 (0.020)	0.034 (0.029)
	15-NN	0.010 (0.016)	0.019 (0.021)	0.029 (0.029)
Nutt	**SVM**	0.013 (0.018)	0.268 (0.078)	0.281 (0.066)
	DLDA	0.049 (0.057)	0.179 (0.069)	0.228 (0.094)
	1-NN	0.189 (0.086)	0.145 (0.109)	0.334 (0.052)
	3-NN	0.109 (0.075)	0.148 (0.138)	0.256 (0.089)
	7-NN	0.054 (0.054)	0.109 (0.132)	0.163 (0.091)
	15-NN	0.038 (0.035)	0.100 (0.138)	0.138 (0.112)
Pomeroy	**SVM**	0.007 (0.011)	0.098 (0.035)	0.105 (0.033)
	DLDA	0.015 (0.015)	0.095 (0.022)	0.110 (0.030)
	1-NN	0.128 (0.093)	0.068 (0.054)	0.196 (0.062)
	3-NN	0.049 (0.036)	0.078 (0.028)	0.127 (0.042)
	7-NN	0.023 (0.019)	0.080 (0.036)	0.104 (0.034)
	15-NN	0.015 (0.016)	0.065 (0.042)	0.080 (0.034)
Shipp	**SVM**	0.008 (0.011)	0.057 (0.029)	0.065 (0.025)
	DLDA	0.012 (0.012)	0.044 (0.026)	0.056 (0.023)
	1-NN	0.125 (0.044)	0.034 (0.064)	0.160 (0.062)
	3-NN	0.039 (0.022)	0.034 (0.055)	0.073 (0.055)
	7-NN	0.021 (0.021)	0.040 (0.045)	0.061 (0.046)
	15-NN	0.024 (0.021)	0.053 (0.040)	0.078 (0.039)
Singh	**SVM**	0.009 (0.012)	0.074 (0.046)	0.082 (0.047)
	DLDA	0.004 (0.008)	0.061 (0.034)	0.065 (0.034)
	1-NN	0.123 (0.099)	0.028 (0.064)	0.151 (0.057)
	3-NN	0.048 (0.025)	0.011 (0.054)	0.058 (0.041)
	7-NN	0.030 (0.023)	0.026 (0.069)	0.056 (0.055)
	15-NN	0.023 (0.012)	0.015 (0.073)	0.038 (0.073)
Grand Avg		0.037 (0.028)	0.063 (0.051)	0.100 (0.046)

Chapter 6

Univariate or Multivariate: Comparing the Results

6.1 Introduction

This chapter reflects on the results of Chapters 4 and 5. Some time will first be given to looking at how effective the single- and two-stage GA-based feature selection processes were at selecting jointly discriminatory subsets of genes that would otherwise not be detected by combining individually predictive genes from the univariate screen. In addition, a direct comparison will be made among the univariate- and multivariate-based approaches, in terms of the 10-fold internal and external CV error curves across a selected group of gene subset sizes, for all learning algorithms and datasets. All analyses of this chapter were performed using the *R* statistical software

package [35] on a Red Hat Linux machine (dual Intel(R) 3.06 GHz processors and 4 GB memory).

6.2 Gene Selection: Univariate vs. Multivariate

This section investigates whether or not the GA feature selection approaches were capable of detecting jointly discriminative groups of genes that would not be easily selected by combining individually predictive genes from the simpler T-test feature selection approach. Results for this section were generated in a resubstitution setting. That is, results were based on performing the feature selection processes on *all* samples of each dataset. This way, specific genes could be easily monitored, as opposed to having the gene subsets vary at each stage of a 10-fold CV procedure. Although for modeling purposes these particular genes would lead to biased results when using cross-validation to assess the predictive accuracy since the same samples would be used for both training and prediction purposes, the purpose of this study is to allow for some quick insights into whether or not the multivariate-based feature selection approaches were able to choose subsets of genes that would have gone otherwise undetected from the univariate approach.

Tables A.7 – A.18 of Appendix A show the genes selected after both the first and second stages of GA, for each of the subset sizes considered. It should be noted that the columns entitled "Gene Index" provide the row (or column, depending on how one imports the dataset) number of each particular gene, so the reader is referred to

the original dataset for cross-referencing any gene of interest (see Section 3.2 for more information pertaining to each dataset's website where the data was downloaded). Regarding the subset sizes, it should be noted that they are the same ones used in the multivariate-based results of Chapter 5, since they represent a common subset from those also considered from the univariate results of Chapter 4, for all six datasets. In addition, these subset sizes yielded desirable CV error rates in both analyses. The low error rates and small subset sizes are both favorable aspects to consider when designing models for use in classifying gene expression data. The tables show where on the univariate ranked gene list each of the selected genes was positioned, as well as each gene's p-value. From these tables, it is interesting to note where many of the genes selected by the GA-based feature selection schemes were located on the univariate-based feature selection list. For all situations except the 1st stage GA results from the Shipp dataset, the gene subsets of size 1 and 2 were made up of the same genes from the univariate and both GA feature selection schemes. Beyond these subset sizes, though, the matches were not exact. In fact, considering as many as the top 100 genes of the univariate list, the GA-based processes selected genes that were not even considered within these top 100 univariately selected genes. Table 6.1 provides a breakdown of the percentage of genes (relative to each gene subset size) among the GA-based feature selection processes that were not even among the top 100 univariately significant genes, for all six datasets.

Another interesting aspect with respect to the effect of performing the second

stage of GA is to consider how many genes selected for the final subsets after the two-stage GA process were not among the final subsets after the first stage of GA. Table 6.2 shows a breakdown of what percentage of genes (relative to each subset size of genes) fell into this category.

Table 6.1: Percentage of Genes Per Subset Size Not Within Top 100 Univ List (Feature Selection Based on All Samples)

	Alon		Golub		Nutt		Pomeroy		Shipp		Singh	
Size	**GA1**	**GA2**	**GA1**	**GA2**	**GA1**	**GA2**	**GA1**	**GA2**	**GA1**	**GA2**	**GA1**	**GA2**
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	0.0	0.0	0.0
3	66.7	66.7	33.3	33.3	33.3	66.7	33.3	100.0	100.0	66.7	0.0	33.3
4	50.0	50.0	50.0	25.0	75.0	75.0	50.0	50.0	75.0	100.0	50.0	50.0
5	20.0	60.0	40.0	20.0	100.0	80.0	80.0	60.0	60.0	80.0	60.0	60.0
10	70.0	70.0	60.0	60.0	70.0	90.0	70.0	80.0	70.0	80.0	60.0	60.0
15	73.3	66.7	60.0	80.0	93.3	86.7	66.7	60.0	93.3	80.0	73.3	73.3
20	75.0	75.0	70.0	80.0	95.0	95.0	85.0	80.0	85.0	85.0	85.0	95.0
25	84.0	84.0	76.0	72.0	96.0	92.0	76.0	84.0	96.0	96.0	84.0	76.0

Table 6.2: Percentage of Genes Per Subset Size Selected from 2-Stage, but not Single-Stage, GA Process
(Feature Selection Based on All Samples)

Size	**Alon**	**Golub**	**Nutt**	**Pomeroy**	**Shipp**	**Singh**
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	50.0	0.0
3	0.0	66.7	100.0	0.0	66.7	33.3
4	0.0	50.0	100.0	50.0	25.0	50.0
5	80.0	40.0	40.0	100.0	40.0	20.0
10	70.0	70.0	100.0	60.0	40.0	90.0
15	73.3	46.7	80.0	73.3	73.3	80.0
20	65.0	60.0	90.0	90.0	90.0	100.0
25	92.0	80.0	100.0	92.0	84.0	88.0

From Table 6.2, just looking at the gene indices selected after the initial stage of GA and after the second stage of GA, for all datasets except the Shipp one, the gene subsets for both GA-based processes were the same for subset sizes 1 and 2. Beyond that, one can see that for all datasets, there were a number of genes selected after the two-stage GA process that were not in the final subsets after the 1st stage of GA, across all six datasets. This was especially true for gene subset sizes 10, 15, 20, and 25. It is feasible to consider that these genes may have appeared in only a small proportion of the 1000 solutions for any run in the first stage of GA, but still appeared in multiple runs of the GA. With the two-stage GA process, these genes had a better chance of being selected for use in building classifiers, since the second implementation of the GA considered as its gene intial gene pool all genes that appeared at least once in the final generation's solutions for each of the 10 runs from the intial GA stage. Hence, the second implementation of the GA attempted to select the 'best of the best' genes, some of which may not necessarily have been selected as part of the final subset of genes for a given subset size after the initial GA stage.

Ultimately though, perhaps the more important thing to compare is the genes selected by the more sophisticated GA-based processes and those selected by the univariate feature selection process. From Table 6.1, one should note that for all datasets except the Golub one in the case of the two-stage GA feature selection process, for subset sizes of four or more, both the single- and double-stage GA processes

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 – all of which yielded very favorable 10-fold internal and external CV misclassification error rates, aside from the unusally high error rates of the Nutt dataset. Thus, these GA approaches provided the opportunity to take into account the presence of discriminatory gene subsets composed of genes that otherwise would have gone undetected from a ranked list of 100 genes generated by the univariate testing method.

6.3 Head-to-Head CV MER Results

The 10-fold external and internal CV results based on a) the univariate- (T-test-) based feature selection process, b) the single-stage GA process, and c) the 2-stage GA process are compared directly in a series of plots (Figures 6.1 – 6.6). For each dataset, a series of six plots are presented, corresponding to each of the six classifiers. Each plot includes the CV MER curves for the univariate- and two multivariate-based feature selection results. It should be noted that for the univariate-based results, the repeated- (10-) run CV results are used. Also note that these subset sizes are those used in the multivariate-based results of Chapter 5, for reasons given in Section 6.2.

Figure 6.1: 10-Fold Ext & Int CV; Univ FSS vs. 1- & 2-Stage GA FSS; Alon Data

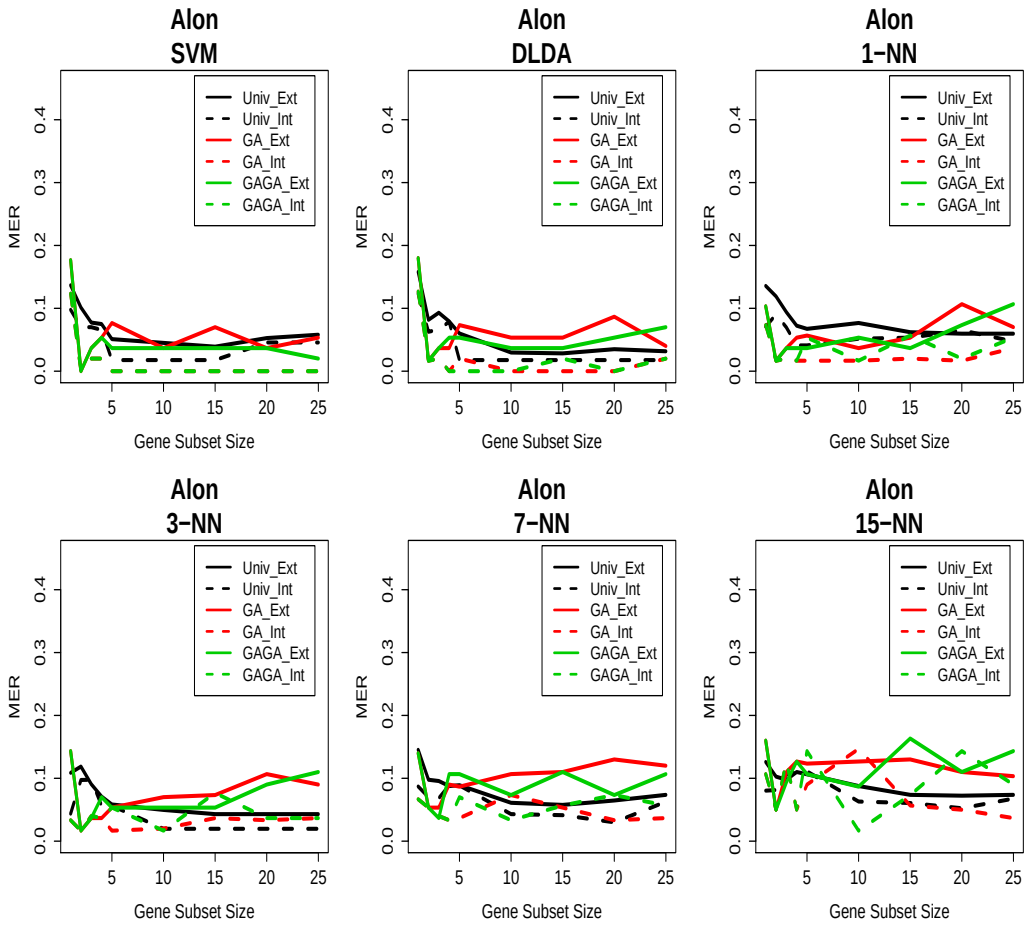


Figure 6.2: 10-Fold Ext & Int CV; Univ FSS vs. 1- & 2-Stage GA FSS; Golub Data

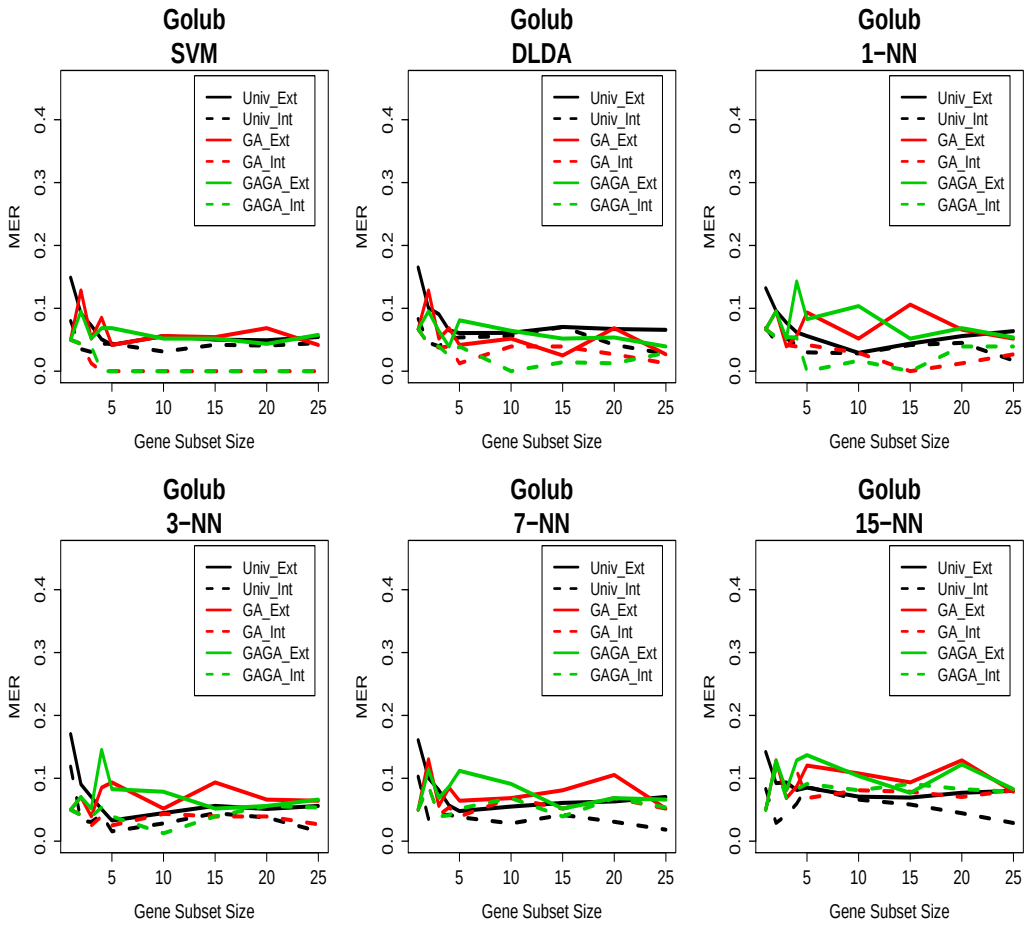


Figure 6.3: 10-Fold Ext & Int CV; Univ FSS vs. 1- & 2-Stage GA FSS; Nutt Data

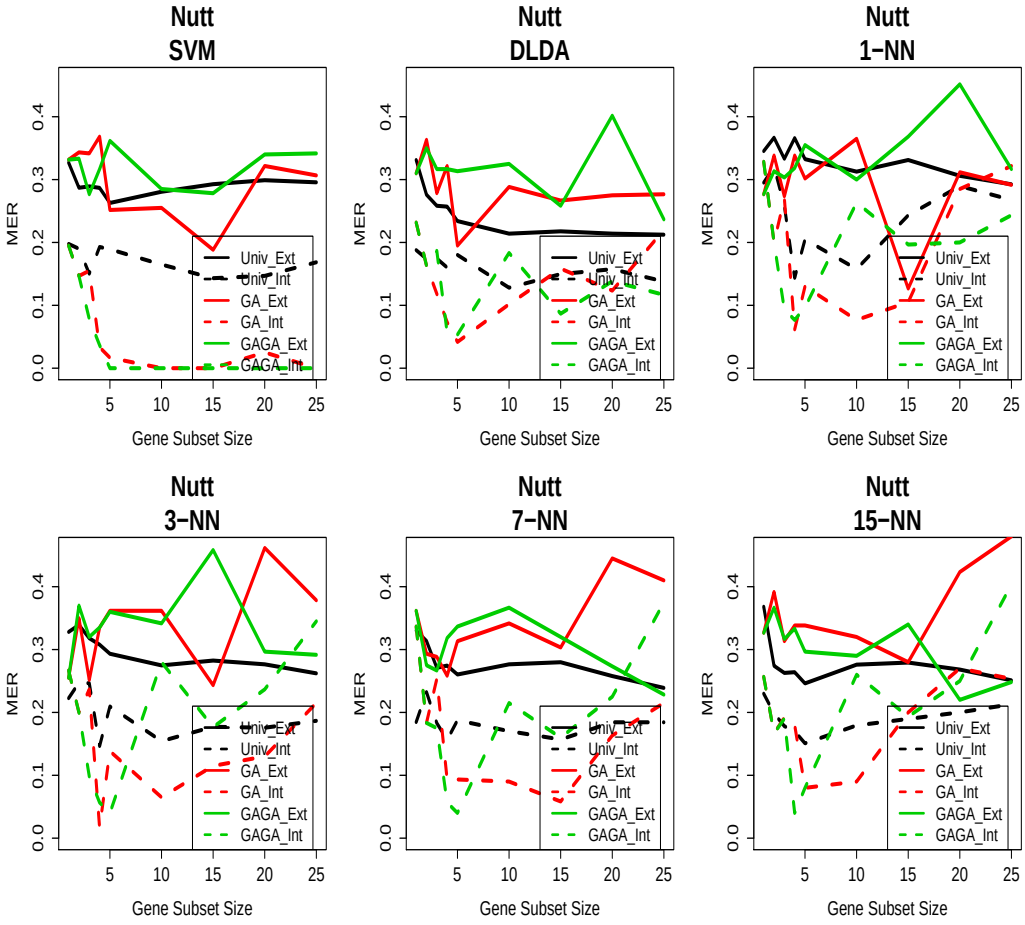


Figure 6.4: 10-Fold Ext & Int CV; Univ FSS vs. 1- & 2-Stage GA FSS; Pomeroy Data

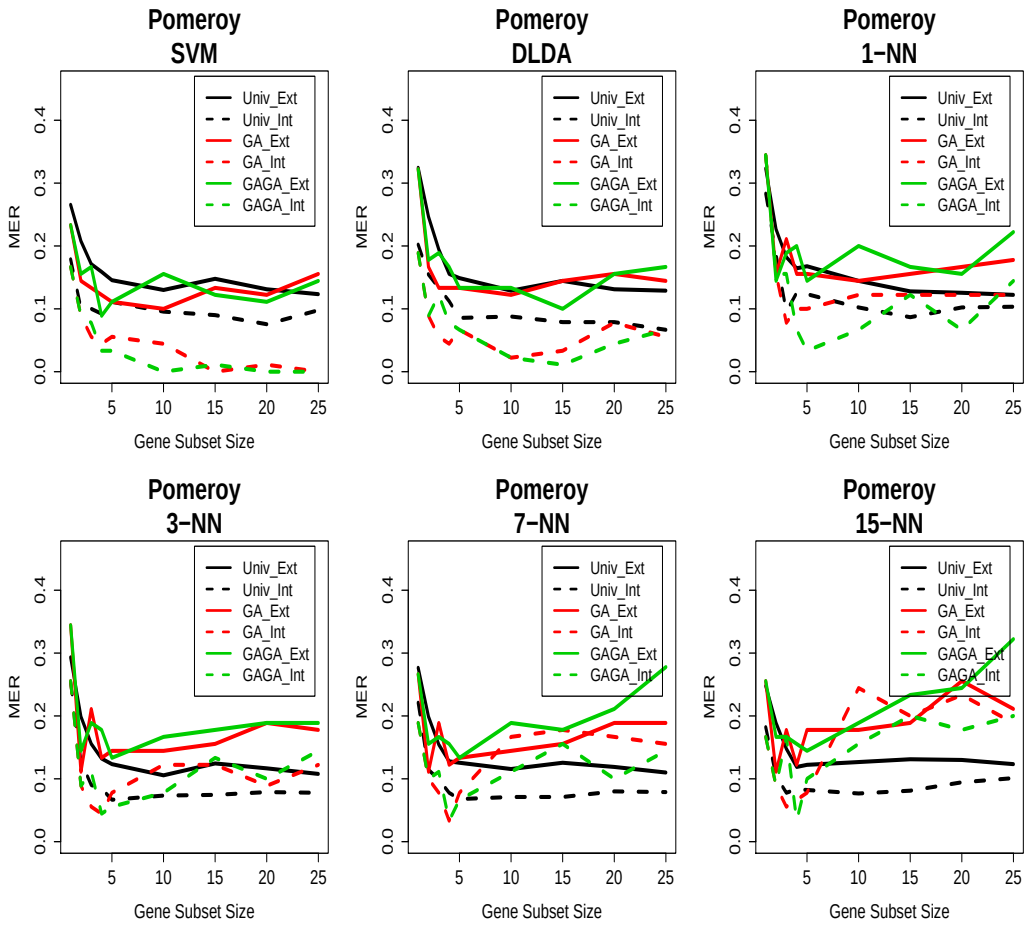


Figure 6.5: 10-Fold Ext & Int CV; Univ FSS vs. 1- & 2-Stage GA FSS; Shipp Data

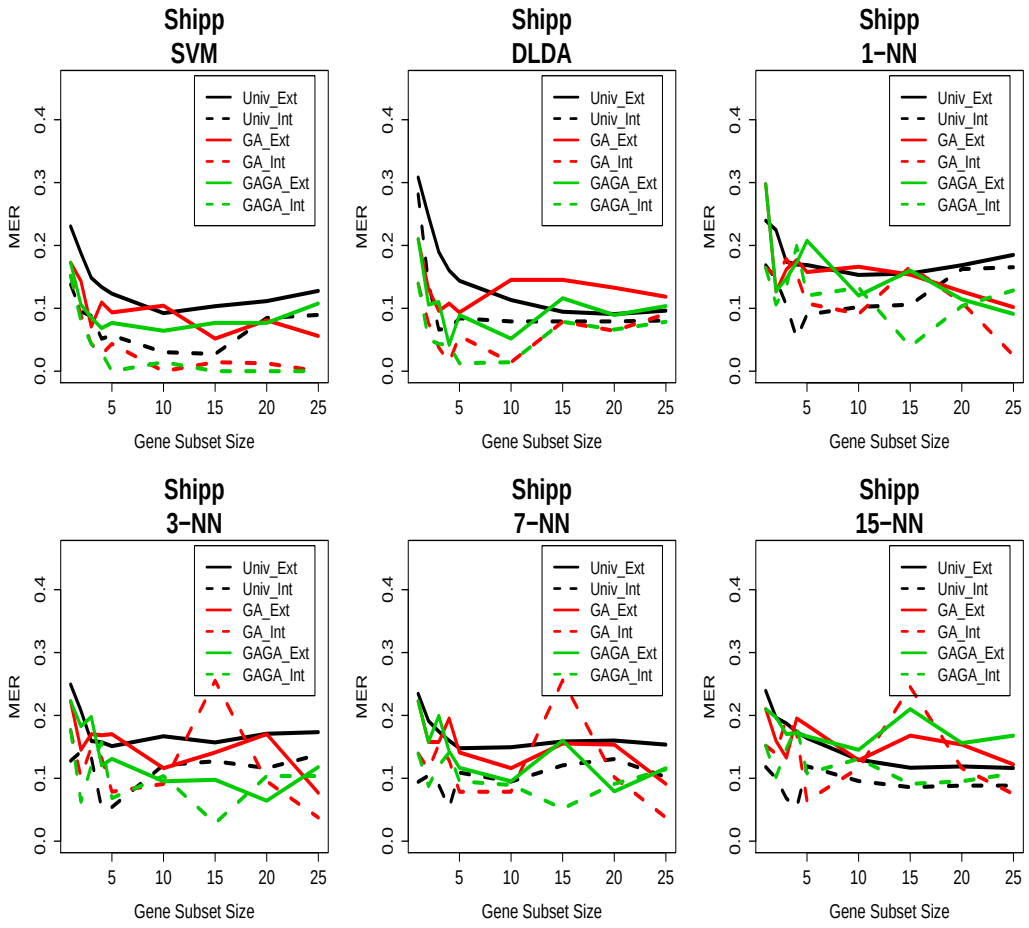
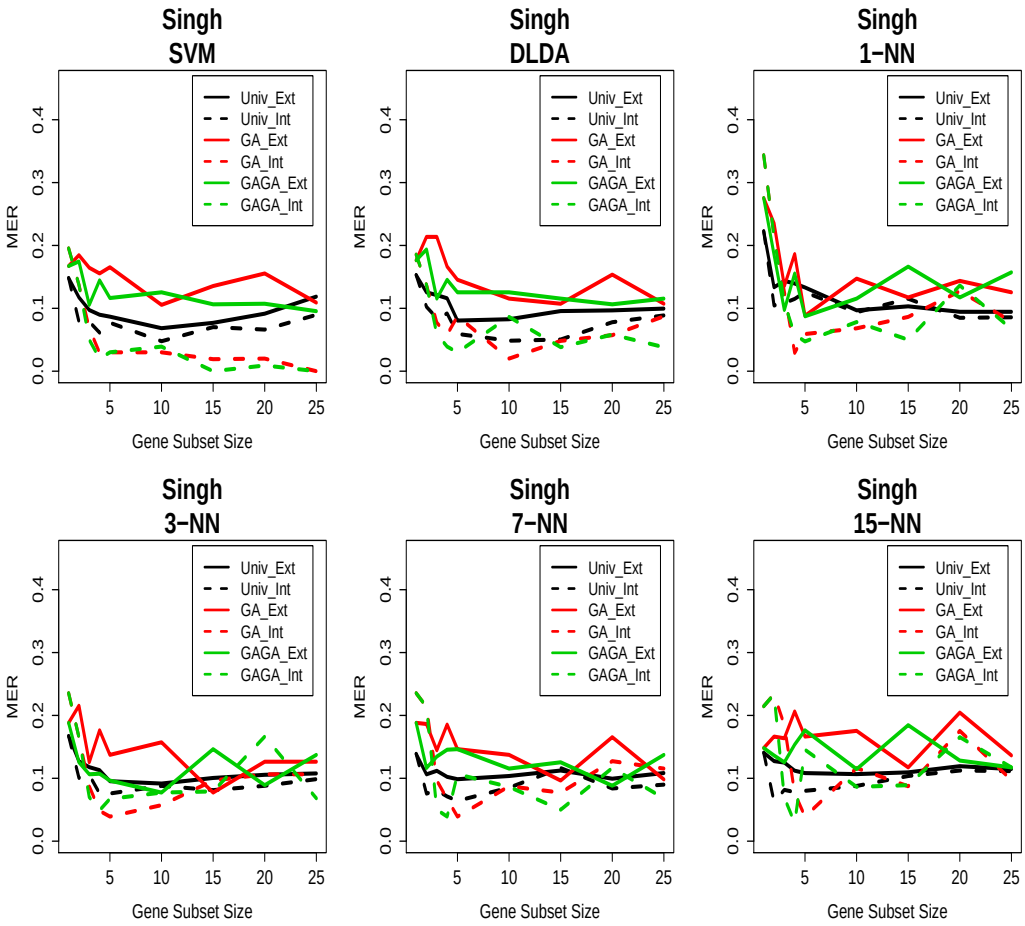


Figure 6.6: 10-Fold Ext & Int CV; Univ FSS vs. 1- & 2-Stage GA FSS; Singh Data



First off, with respect to the external and internal CV results in Figures 6.1 – 6.6, several things should be noted. Overall, a finding that was quite evident throughout all the classifiers and datasets was the fact that the external CV error rates were in general higher than the corresponding internal CV error rates. This trend was expected since the feature selection procedure was incorporated into the CV process, thus providing for classifiers that yielded more realistic error rates than would be the case if the feature selection were performed on all N samples of a given dataset prior to performing the cross-validation. For all datasets, whether internal or external CV was used, it was clear that the single- and double-stage GA-based processes led to very comparable misclassification errors across learning algorithms and subset sizes. Furthermore, it was interesting to note that if one were to compare the univariate-based results with either of the GA-based results, for either type of CV and any choice of learning algorithm, one can see that the GA-based analyses did not offer any significant advantage over the univariate results in terms of lower error rates across subset sizes. This finding was evident across all six datasets.

It should be also be noted that as mentioned in Chapters 4 and 5, the Alon and Golub datasets (Figures 6.1 and 6.2, respectively), had internal and external CV error rates that were in general lower across subset sizes, learning algorithms, and feature selection methods than those from the other four datasets. On the other hand, the Nutt dataset had error rates that were markedly higher than those of the other four datasets. These findings reinforce what was suggested in the MDS plots in Section

4.2.1. In terms of learning algorithms and subset sizes, no particular algorithm or size emerged as clearly the "best" in terms of lowest error rates (internal or external) across all the datasets.

Side Note: Confirming the Presence of a Three-Way Interaction

It should be noted that to further investigate the presence of a three-way interaction among dataset, learning algorith, and subset size (a notion suggested in the results from Chapters 4 and 5.1), a linear model approach was taken. Treating the external CV MER as the response variable and dataset, learning algorithm, and subset size as the three predictor variables, a linear model was constructed. Dataset and learning algorithm were treated as factor variables, while subset size was treated as a continuous variable. This model included the two- and three-way interactions as well. Although the normality of residuals was verified from this model, there was a piece-wise linear trend evident in a plot of the residuals vs. gene subset size – decreasing residuals from subset size 1 to 5 and then increasing residuals from size 5 to 25. This trend was successfully removed by fitting a piece-wise linear function for gene subset size. In order to test the significance of the three-way interaction, this first model was compared to a second model which contained all the nested main effects and two-way interactions, but not the three-way interaction. Since the second model was nested within the first and since the first model included 25 additional parameters, an F test was used to assess the significance of the interaction term. It was found that the

F-statistic was indeed significant at the 0.001 significance level. This study confirmed the suggestion that a three-way interaction exists – that is, that the subset size effect was dependent on both the choice of learning algorithm and which dataset was being investigated.

Further Insights: Collapsing the Results Across Learning Algorithms, Subset Sizes, and Datasets

First off, combining the results across learning algorithms, it was interesting to observe which subset size led to the minimum average error rate for each of the three feature selection approaches and for each of the six datasets. These results are shown in Tables 6.3 and 6.4.

Table 6.3: Minimum 10-Fold Internal CV Average MER's Across All Classifiers Univ FSS vs. 1- & 2-Stage GA FSS

	Univ FSS		**1-Stage GA FSS**		**2-Stage GA FSS**	
Dataset	**MER (SD)**	**Subset Size**	**MER (SD)**	**Subset Size**	**MER (SD)**	**Subset Size**
Alon	0.035 (0.020)	10 (also 15)	0.026 (0.018)	4	0.014 (0.013)	10
Golub	0.025 (0.011)	25	0.031 (0.024)	5	0.030 (0.036)	10
Nutt	0.159 (0.018)	10	0.071 (0.037)	10	0.052 (0.034)	5
Pomeroy	0.080 (0.007)	15	0.056 (0.024)	4	0.048 (0.019)	4
Shipp	0.055 (0.006)	4	0.044 (0.033)	25	0.048 (0.033)	15
Singh	0.075 (0.021)	10	0.049 (0.021)	5	0.039 (0.014)	4

Table 6.4: Minimum 10-Fold External CV Average MER's Across All Classifiers Univ FSS vs. 1- & 2-Stage GA FSS

	Univ FSS		**1-Stage GA FSS**		**2-Stage GA FSS**	
Dataset	**MER (SD)**	**Subset Size**	**MER (SD)**	**Subset Size**	**MER (SD)**	**Subset Size**
Alon	0.051 (0.017)	15	0.026 (0.021)	2	0.026 (0.021)	2
Golub	0.053 (0.014)	10	0.051 (0.011)	3	0.056 (0.009)	1 (also 15)
Nutt	0.259 (0.032)	25	0.235 (0.066)	15	0.277 (0.046)	25
Pomeroy	0.119 (0.008)	25	0.131 (0.013)	4	0.133 (0.012)	5
Shipp	0.131 (0.029)	15	0.094 (0.025)	25	0.095 (0.035)	10
Singh	0.092 (0.014)	10	0.108 (0.020)	15	0.106 (0.016)	20

For all datasets except the Singh data, Table 6.3 shows that the minima of the average 10-fold internal CV MER's across classifiers using either of the GA-based analyses were based on subset sizes that were smaller than or equal to those that led to the minimum MER's using the univariate-based analyses. Also, the minimum average MER's from performing univariate feature selection were each slightly higher than the two GA-based approaches for all datasets except the Golub one, although the Golub error rates for the 1- and 2-stage GA-based results were based on subset sizes of 20 and 15 genes less, respectively. Furthermore, in looking at the 1-stage and 2-stage GA results, one can see that the minimum average 2-stage GA error rates were slightly lower than those of the 1-stage process for all datasets except the Shipp one, although these MER's were still very comparable. From the external CV results in Table 6.4, the 2-stage GA results were very comparable to those of the 1-stage process for all datasets. The univariate-based results were all higher than those of the 1-stage GA results for all datasets except the Pomeroy and Singh datasets. However, they were lower than those of the 2-stage GA results for all datasets except the Alon and Shipp data. In terms of subset sizes at which the minimum average MER's occurred, for all datasets except the Shipp and Singh ones, the sizes from the univariate-based results were all at least as big as those from the two GA-based feature selection approaches. Considering both tables together, one can see that all minimum MER's based on the external CV analyses were larger than their counterparts from the internal CV analyses, as expected since the feature selection was built into the

CV process. With respect to subset sizes, in the majority of the dataset and feature selection method combinations, the sizes for the external CV results were larger than those of the internal CV results.

Finally, the empirical grand means for both internal and external CV results, as well as the resubstitution results, are presented in Table 6.5. Results for each of the three feature selection approaches are averaged across all datasets, learning algorithms, and subset sizes. The average resubstitution and CV error rates from the univariate-based feature selection results were quite comparable to each of the GA-based analyses. In comparing the average internal CV error rates and the average resubstitution error rates, the two GA-based processes led to internal CV errors that were slightly lower than the resubstitution ones, but the opposite relationship held for the univariate-based feature selection results. Also, it should be noted that the discrepancy between the average external and internal CV error rates, as well as between the external CV and resubstitution error rates, was quite evident for all three feature selection approaches. These relationships among resubstitution, internal CV, and external CV, are investigated further in Section 6.4.

Table 6.5: IntCV, ExtCV, Resub MER: Empir Grand Means Across All Datasets, Classifiers, & Subset Sizes; Univ FSS vs. 1 & 2-Stage GA FSS

	Univ FSS	**1-Stage GA FSS**	**2-Stage GA FSS**
Resub MER	6.2 %	5.5 %	5.3 %
IntCV MER	5.7 %	9.1 %	9.0 %
ExtCV MER	14.7 %	15.6 %	15.4 %

6.4 Head-to-Head Optimism and Selection Bias Results

Again referring to Figures 6.1 – 6.6, it is of particular interest to note the effect of performing internal CV instead of external CV. Whether univariate or multivariate feature selection was implemented, one can see that the external CV curves for all learning algorithms were in general slightly higher across subset sizes than those of the internal CV. This was evident for all datasets, although it should be noted that the discrepancies between error rates between the internal and external CV curves from the Nutt dataset were larger than those from the other five datasets. The slightly higher external CV error rates did not come as a surprise, though. With external CV, the feature selection process was built into the CV procedure. Because of this, the error rates were more reflective of true generalization error, since the feature selection was performed on 90% of the samples; that is, externally to each CV stage's test set samples. Also, although resubstitution curves were not explicitly shown in these plots, it should be noted that there was also a penalty in terms of higher error rates across all subset sizes, learning algorithms, and datasets from performing internal CV over resubstitution error estimation. This was expected since resubstitution estimation was based on performing the feature selection, classifier training, and classifier evaluation on all the samples, which of course resulted in extremely optimistic error rates. As first discussed in Section 2.7, the higher error rates incurred from perform-

ing a) resubstitution instead of internal CV and b) external CV instead of internal CV bring up the issues of optimism and selection bias, respectively. Both of these bias measures are plotted against subset size in Figures 6.7 – 6.12, for each dataset and learning algorithm combination. These tables provide for a direct comparison between the optimism and selection bias values incurred from using univariate-based feature selection and those incurred from using either of the two GA-based feature selection approaches. For simplicity purposes, it should be noted that the "total bias" curves are not shown on each of these plots. One can get an idea of these results knowing that the "total bias" is merely the sum of the optimism bias and selection bias values at each subset size.

Figure 6.7: 10-Fold CV; Univ, 1-, & 2-Stage GA FSS; Opt & Sel Bias vs. Subset Size; Alon Data

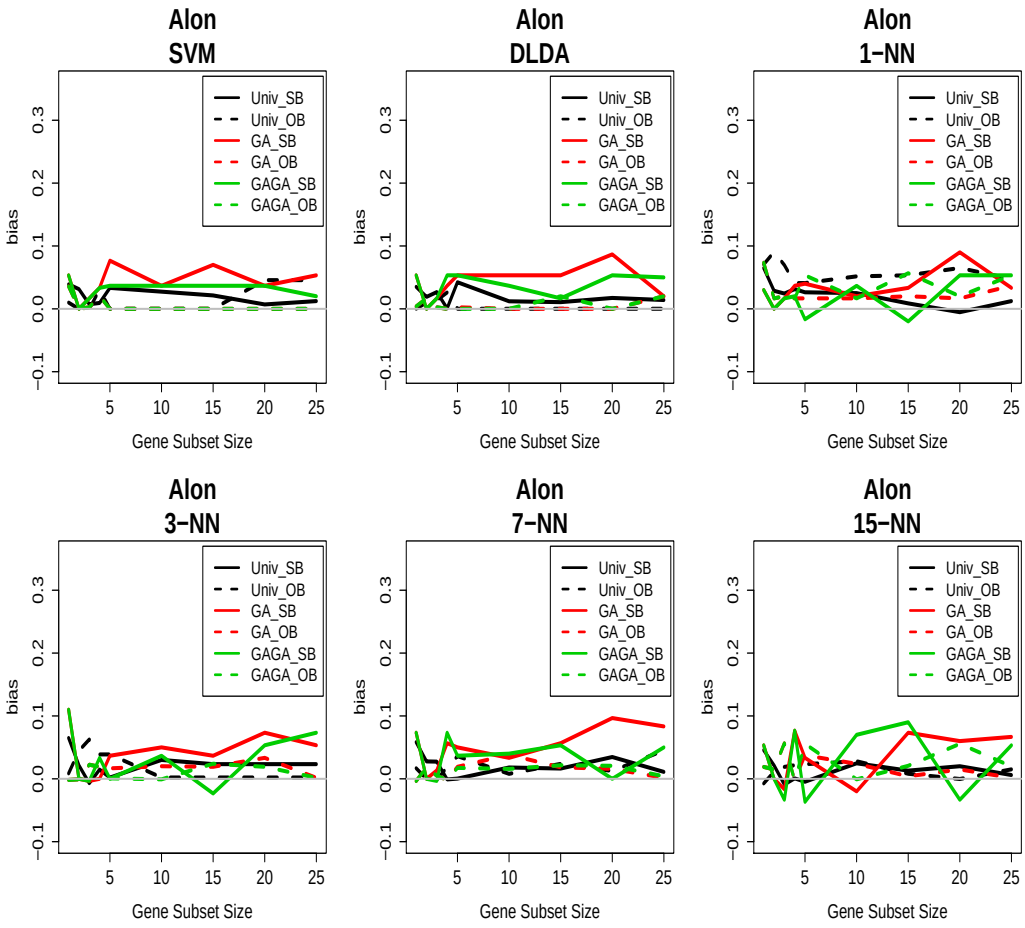


Figure 6.8: 10-Fold CV; Univ, 1-, & 2-Stage GA FSS; Opt & Sel Bias vs. Subset Size; Golub Data

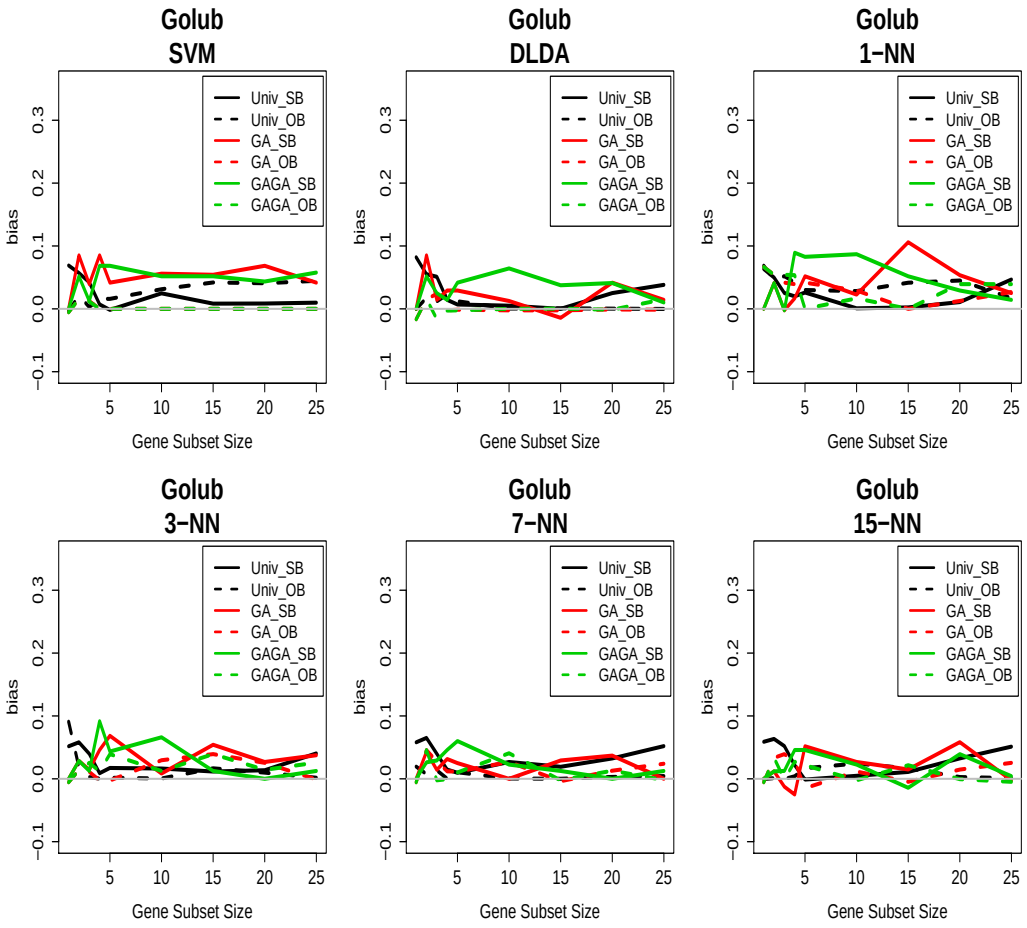


Figure 6.9: 10-Fold CV; Univ, 1-, & 2-Stage GA FSS; Opt & Sel Bias vs. Subset Size; Nutt Data

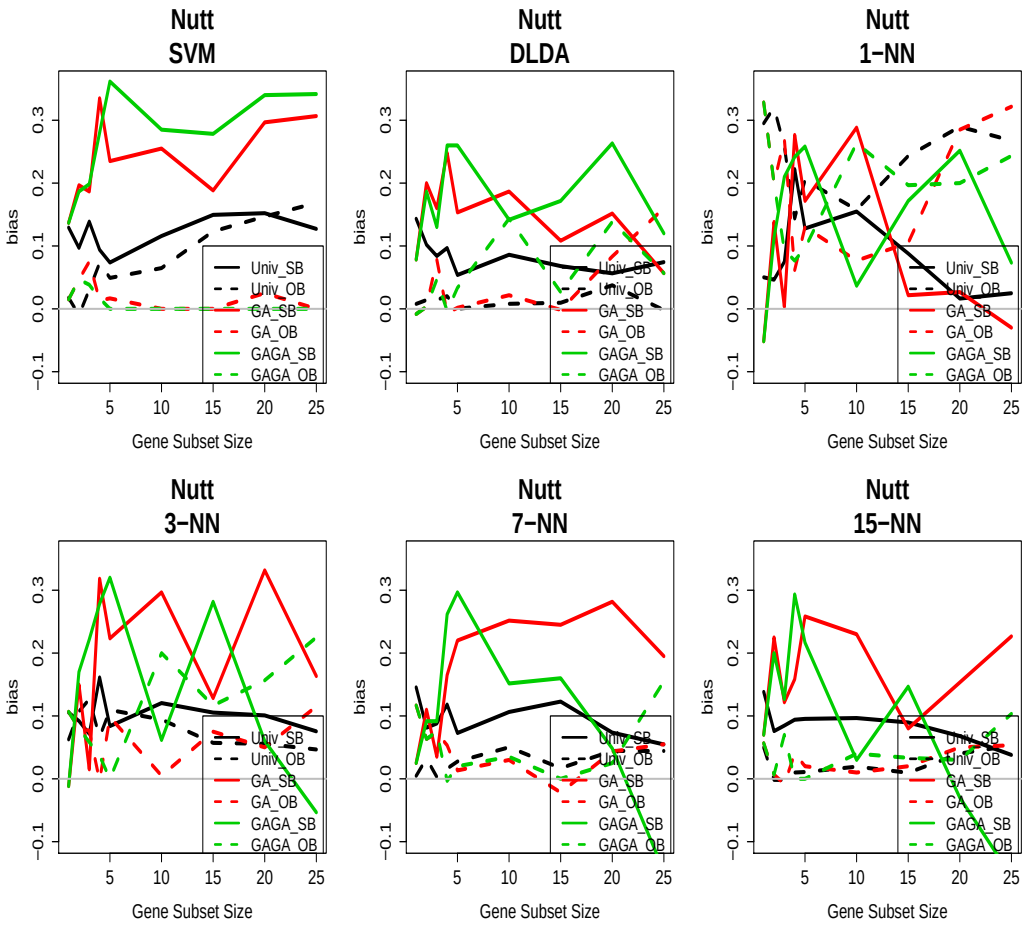


Figure 6.10: 10-Fold CV; Univ, 1-, & 2-Stage GA FSS; Opt & Sel Bias vs. Subset Size; Pomeroy Data

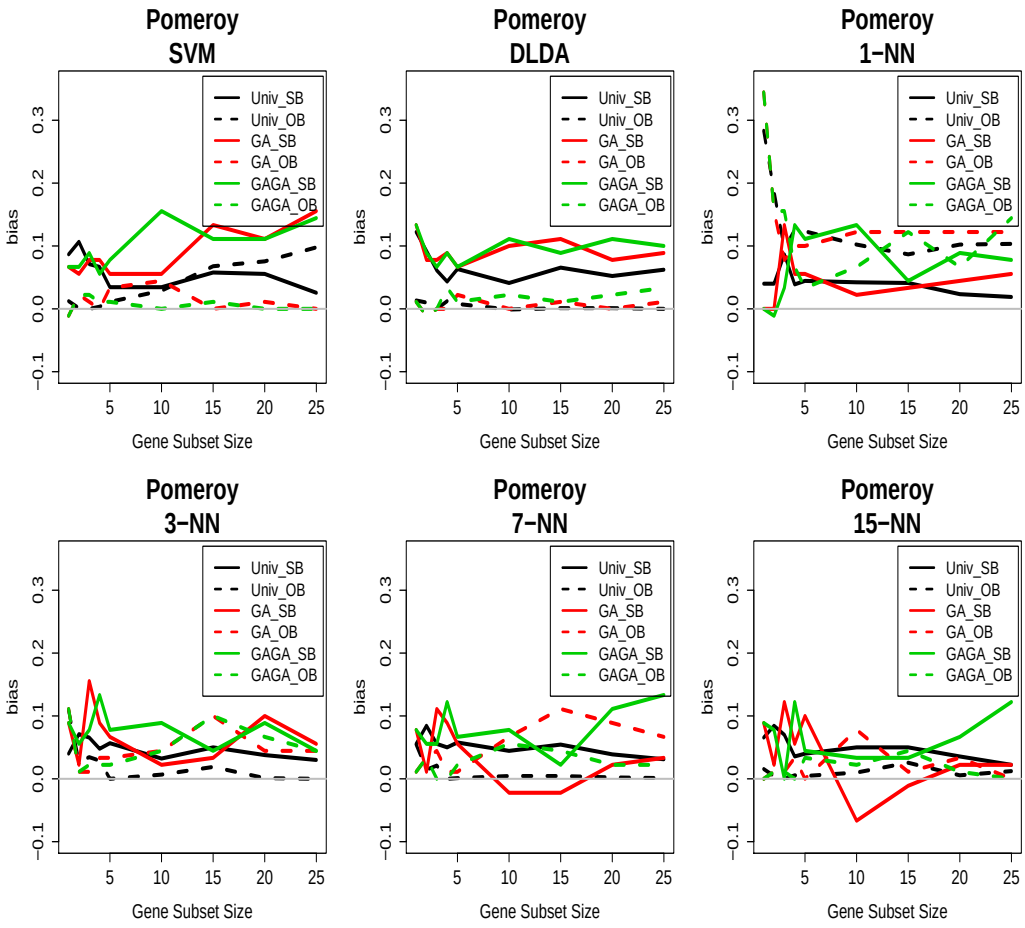


Figure 6.11: 10-Fold CV; Univ, 1-, & 2-Stage GA FSS; Opt & Sel Bias vs. Subset Size; Shipp Data

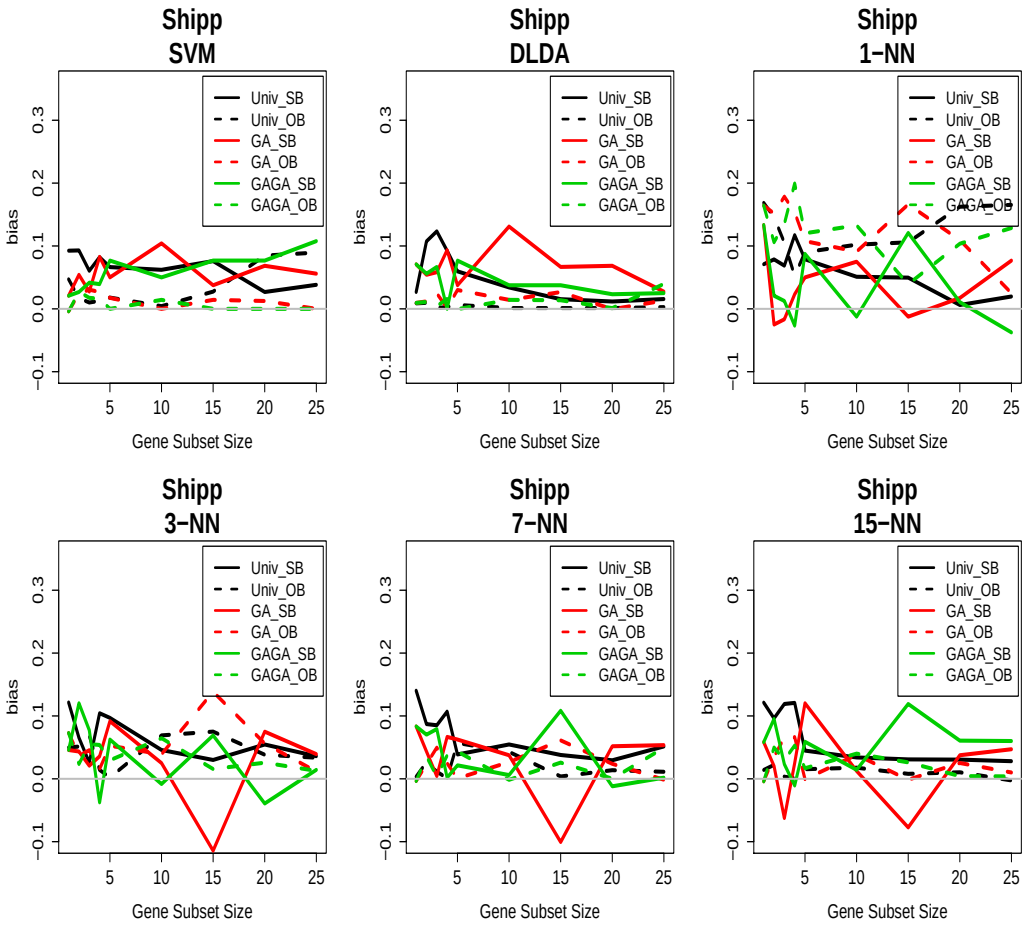
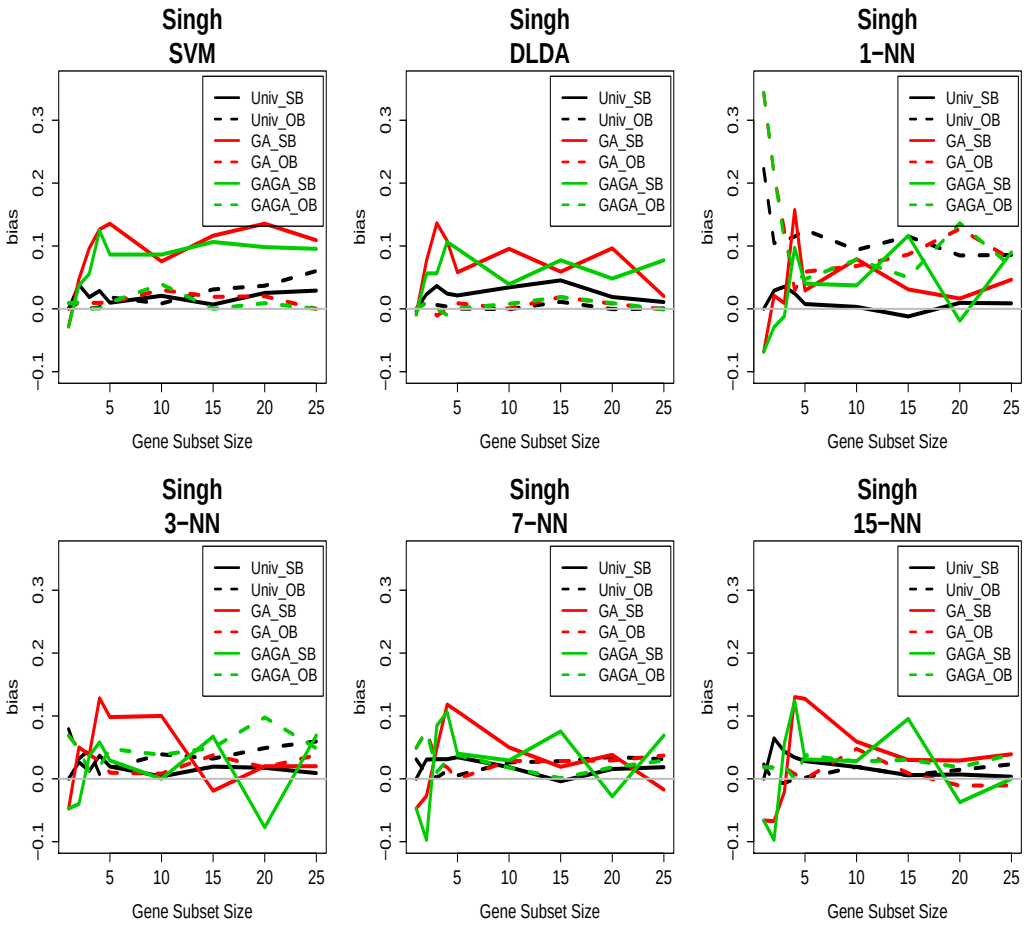


Figure 6.12: 10-Fold CV; Univ, 1-, & 2-Stage GA FSS; Opt & Sel Bias vs. Subset Size; Singh Data



From Figures 6.7 – 6.12, one can see the results from the univariate- and multivariate-based analyses from Chapters 4 and 5, respectively, together. That is, with respect to datasets, the Alon and Golub ones generally had smaller optimism and selection bias values across the six learning algorithms and all the gene subset sizes. On the other hand, the Nutt dataset generally had the largest bias values among the classifiers and subset sizes. In terms of learning algorithms, both DLDA and 1-NN generally led to smaller optimism bias values than did the other four algorithms, while there was no clear best (or worst) learning algorithm that emerged for the selection bias results. Similarly, no particular subset size emerged with significantly better (or worse) optimism and selection bias values across learning algorithms and datasets.

Furthermore, directly comparing the optimism and selection bias curves from the univariate, 1-stage, and 2-stage feature selection processes, several things should be noted. Focusing on the two GA-based feature selection curves in each of the datasets' plots, the two-stage GA process did not offer noticeably lower optimism nor selection bias values across the learning algorithms and subset sizes than those from the single-stage GA process. With respect to the optimism bias curves only, no immediately obvious conclusions were evident, in terms of the univariate-based results versus either of the two GA-based results. At certain subset sizes the univariate-based bias values were higher than the multivariate ones, but for other sizes the opposite was true, and even these trends varied by learning algorithm and dataset. With respect to the selection bias curves, though, it was very interesting to note that in general the

univariate-based bias values were often lower than those of each of the two GA-based selection bias values across subset sizes and learning algorithms. Regarding the latter, these findings seemed to be even more evident for the SVM and DLDA learning algorithms than for the four k-NN algorithms. To further investigate these bias findings, let us revisit the empirical grand means for the optimism, selection, and total bias results, averaged across all datasets, learning algorithms, and subset sizes. These results are presented for all three feature selection approaches together in Table 6.6. The average optimism bias estimates from each of the GA-based analyses were only slightly less than that from the univarate-based analyses. However, the average selection and total bias estimates from each of the GA-based analyses were roughly 2.5 and 1.4 times, respectively, the average selection and total bias estimates incurred from performing univariate-based feature selection.

Table 6.6: Opt & Sel Bias: Empir Grand Means Across All Datasets, Classifiers, & Subset Sizes; Univ FSS vs. 1 & 2-Stage GA FSS

	Univ FSS	**1-Stage GA FSS**	**2-Stage GA FSS**
Opt. Bias	4.7 %	3.6 %	3.7 %
Sel. Bias	2.6 %	6.5 %	6.3 %
Tot. Bias	7.3 %	10.1 %	10.0 %

6.5 Final Thoughts

Overall, the results of this chapter have addressed several important issues regarding feature subset selection and its role with respect to the use of cross-validation

when assessing classifiers. In particular, questions involving a) the use of univariate versus multivariate (GA-based) feature subset selection methods and b) how to most honestly evaluate candidate classifiers such that the effects of optimism and selection bias are properly taken into account, were addressed across six published two-class microarray datasets, six learning algorithms, and a number of gene subset sizes. General conclusions that can be inferred from the results presented in Chapters 4, 5, and 6 are the subject of the following chapter.

Chapter 7

Conclusions and Further Thoughts

Chapter 4 presented 10-fold internal and external cross-validation misclassification error rate results, as well as optimism, selection, and total bias results, across six learning algorithms and six published microarray datasets. All CV results were based on a feature selection approach that was univariate in nature (rank-based, unequal variance T-tests). Chapter 5 presented the same types of results based on two feature selection approaches that were multivariate in nature – a single-stage and two-stage genetic algorithm. In Chapter 6, the external and internal cross-validation results based on all three feature selection methods were directly compared. This chapter provides an overall summary of all these results by focusing on four important areas from this large-scale empirical comparison study of feature selection in binary classification with microarray data: choice of learning algorithm and subset size, choice of feature subset selection technique, choice of using internal CV versus external CV,

and finally the presence of optimism bias and selection bias.

7.1 Learning Algorithms and Subset Sizes

Based on these studies, the findings across all six datasets have suggested that the choice of learning algorithm and gene subset size were not clear-cut choices when constructing a prediction model for performing binary classification on a given microarray dataset, based on the 10-fold internal and external CV prediction errors of the classification process implemented. As far as subset sizes go, however, it was clear that desirable classification results could be obtained using gene subset sizes of 25 or less, based on the six published microarray datasets of this study at least.

7.2 Feature Subset Selection Approaches

7.2.1 Gene Selection

It is important for the reader to recall that for the purpose of actually determining particular genes to comprise each subset size, a resubstitution setting was used to obtain the results discussed in this section. For all datasets, there were a number of genes selected by the two-stage GA process for the final subsets of genes for a given subset size, but not for the final subsets of genes selected by the initial stage of GA. For gene subset sizes 10, 15, 20, and 25, this was especially true (see Table 6.2). These genes may have appeared in only a small proportion of the 1000 solutions

found within any single GA run, but they still may have appeared in multiple runs of the GA. With the two-stage GA process, these genes had a better chance of being selected for use in building classifiers. Recall that the second implementation of the GA attempted to take this issue into account by considering for its initial gene pool all genes that appeared at least once among the final generation's 1000 solutions for each of the 10 runs from the initial implementation of the GA procedure (i.e., at least once in the "superpopulation" of 10000 solutions from the initial GA procedure). If there were intuition for a particular microarray dataset that as many genes as possible should be taken into account when performing feature subset selection, then the GA-based multivariate approaches to selecting genes from the initial gene pool would be warranted. The motivation for implementing the two-stage GA process is as just described – to take into account *all* genes chosen among the final generation's solutions from all runs of an initial (single-stage) implementation of GA.

Also in a resubstitution setting, for the same reason given above, this gene selection issue was investigated in comparing the univariate feature selection results versus both the single- and the double-stage GA-based feature selection results (refer to Table 6.1). For subset sizes of four or greater, both GA-based feature selection processes generated final gene subsets in which the majority of the genes relative to each subset size were *not* among the top 100 univariately-ranked gene list. For subset sizes of 10, 15, 20, and 25, all of which yielded favorable misclassification error rates (aside from the unusually high error rates in general incurred with the Nutt data), this finding

was especially true. It should be noted that this finding was true for all datasets except the Golub one, in the case of the two-stage GA feature selection process.

7.2.2 CV Error Rates

The ability of the two-stage GA feature selection process to select genes that the single-stage procedure would not have included in its final subsets has been shown. However, of particular interest is that in terms of the actual 10-fold internal and external CV error misclassification error rates, the two procedures were very comparable across all six learning algorithms, all subset sizes, and all six datasets, with the more sophisticated two-stage procedure offering only minimal advantages with respect to lower error rates. Perhaps even more surprising, however, was how well all the CV misclassification error rates based on the simple rank-based T-tests did in comparison to the more complicated and computationally intensive GA-based feature selection procedures. Despite the ability of the GA procedures to select subsets of genes that would most likely go undetected via the combination of individually predictive genes from say, the top 100 genes from a ranked genes list, the results in Chapter 6 clearly showed that neither of the two GA-based feature selection methods led to significantly better 10-fold CV error rates. This finding was true across all learning algorithms, all subset sizes, and all six datasets, whether internal CV or external CV was implemented. More discussion on the topics of internal and external CV in particular is provided in the following section.

7.3 Internal CV vs. External CV

One of the most important aspects of this research was the investigation of the role of feature selection with respect to cross-validation. The standard role of feature selection with respect to cross-validation has been to perform the feature selection on all N samples of a given dataset, such that the test samples at each stage of a CV process were also used during the feature selection process ("internal CV"). Hence, overly optimistic error rates would be incurred. As a major part of this large-scale comparative study of gene expression-based clinical outcome classifiers, this research considered another approach to doing the cross-validation – external CV, in which the feature selection was performed at each stage of a CV process only on that stage's training set samples, external to each stage's test set samples. In doing so, the idea was to provide for more realistic and honest misclassification error rates than would normally be the case with internal CV and of course resubstitution error estimation. The results presented in Chapters 4, 5, and 6 confirmed this idea across all six datasets, all six learning algorithms, and all subset sizes, using both univariate- and GA-based feature selection approaches; the latter FSS technique of which had never been implemented in a CV setting of this nature before. Whether the feature selection method was univariate or multivariate in nature, the results of this research showed that a 10-fold external CV procedure did not suffer much at all in terms of higher error rates than those of a 10-fold internal CV procedure to assess the predictive accuracy of a given binary classification procedure, while at the

same time incorporating the feature selection process into the cross-validation. In general, the plots shown in Figures 6.1 – 6.6 illustrated these discrepancies between the external CV results and internal CV results. Only in the case of the Nutt dataset were these discrepancies noticeably larger. This dataset, for reasons unknown at this point, generally had unusually high error rates for all classifiers, relative to the other five datasets' corresponding results (a finding that what was suggested in the exploratory MDS plots in Section 4.2.1). It is extremely important, however, to not ignore the discrepancies that exist in terms of the external CV error rates being higher in general than those obtained from internal CV. This situation arose as a result of the feature selection procedure being built into the CV process, thus providing for classifiers that yielded more realistic error rates than would be the case if the feature selection were performed on all N samples of a given dataset prior to performing the cross-validation. Taking all the internal CV error rates together and averaging them across datasets, learning algorithms, feature subset sizes, the empirical grand means for the error rates, based on univariate, single-stage, and double-stage GA feature selection, were 5.7%, 9.1%, and 9.0%, respectively. For external CV, these empirical grand means were quite higher, as one would expect – 14.7%, 15.6%, and 15.4%, respectively. More insights into internal and external CV were provided during the investigation of optimism and selection bias, which is summarized in the following section.

7.4 Optimism and Selection Bias

Directly related to the resubstitution, internal CV, and external CV methods of assessing the predictive accuracy of a classification process are the issues of optimism and selection bias, incurred from performing a) resubstitution instead of internal CV and b) external CV over internal CV, respectively. Figures 6.7 – 6.12 allowed for a direct comparison to be made between the optimism bias curves from the univariate-based and each of the two GA-based feature selection methods, as well as between the selection bias curves from the univariate and GA-based feature selection approaches. Although no clear conclusions were evident with respect to the optimism bias curves for each of the three feature selection methods, in general the single- and two-stage GA-based selection bias estimates were higher than those of the univaritate-based analyses across subset sizes, learning algorithms, and datasets. In fact, taking all the selection bias estimates together across all subset sizes, learning algorithms, and datasets, the average selection bias estimates from each of the GA-based methods were roughly 2.5 times that of the univariate-based method. Thus, although the more sophisticated GA feature selection techniques were able to select subsets of genes that would likely go undetected via combining univariately discriminatory genes from a ranked list of genes, it is important to realize that they could nevertheless also have greater potential to select spurious genes than would be the case with a univariate-based feature selection 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 viewed as one type of data overfit. 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 test data, it would be safe to say that the GA-based methods tend to overfit the data. This notion of data overfit can also be recognized by considering the total bias measure defined in Equation 2.37. That is, taking total bias to be the difference between external CV MER and training error, the average total bias estimates across all subset sizes, learning algorithms, and datasets from each of the GA-based methods were nearly 1.5 times that of the univariate-based approach. Ultimately, whether a univariate or a GA-based feature selection approach is implemented, the presence of both optimism and selection bias should be taken into account through the use of external CV.

7.5 Impact

This research provided for a large-scale empirical comparative study on feature subset selection in binary classification with DNA microarray data. This research builds on the findings of the studies by Ambroise and McLachlan [7], Dudoit and Fridlyand [14],

Dudoit et al. [15], and Xiong et al. [41], 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. [23] and [24] in that the GA is actually incorporated into each stage of a 10-fold (external) CV procedure, rather than have the data split into reduced training and test sets. It also builds on these results in that once subsets of genes are selected by the initial stage of GA (single-stage approach), all unique genes selected are *not* then pooled together again such that the final subsets used for modeling would actually be selected based on their frequency of selection from the first stage 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, Mahalanobis distance, a simpler and less computationally intensive objective function than k-NN, is employed in the GA implemented in this research.

Ultimately, this study has provided a more extensive comparative analysis than any type of microarray classification study to date in terms of the number of datasets considered, the range of classifiers (including learning algorithms, feature subset selection methods, and subset sizes), and the investigation of both internal *and* external

K-fold CV as a means of assessing the predictive accuracy of the classification rules. This research has also put to test the more traditional implementations of the statistical learning aspects of cross-validation and feature selection. With respect to the former, there has not been a strong enough emphasis in the microarray classification literature placed on the importance of building the feature selection process into a traditional CV procedure, as a means of taking into account selection bias and hence providing more realistic and honest estimates of generalization accuracy when performing limited-sample classification studies using gene expression data. With respect to feature selection, this research has shown across multiple datasets and classifiers how a simple univariate feature selection approach can perform quite comparably to a more sophisticated type of high-dimensional optimization search algorithm such as a genetic algorithm. At the same time though, this research showed the effectiveness of the genetic algorithm as a feature subset selection technique that can select combinations of genes that would most likely go otherwise undetected by combining highly ranked discriminatory genes from among a ranked list of univariately significant genes.

7.6 Future Directions

I believe this research has provided a solid foundation in terms of the magnitude of the empirical study that has been undertaken – a study involving the comparison of feature selection and binary classification techniques, as well as two general approaches to implementing cross-validation to assess the predictive accuracy of candidate clas-

sifiers. At the same time, I also believe this research has provided fertile ground from which a number of other possible interesting investigations could be pursued. This study has put to test some of the more traditional assertions of a) cross-validation (i.e., how to implement it to achieve more honest error rate estimates), and b) feature subset selection (i.e., determining whether a more sophisticated high-dimensional search technique such as the GA really offers a significant advantage when performing binary classification using gene expression data).

Although no particular learning algorithm emerged as a clear-cut best choice to use across all the datasets, I still believe more in-depth investigation to determine if one can somehow infer from each of the datasets any clues that suggest what methods of feature selection and learning algorithm would be best suited to a particular dataset is warranted. The unsupervised learning technique of multidimensional scaling was used in this research as one form of exploratory analysis to investigate the structure of the datasets – in particular, to see obtain an idea of how difficult the classification task may be in each case. Unfortunately, there does not seem to be a clear-cut way of inferring from the features of a given dataset what type of feature selection and supervised learning algorithm should definitely be used for performing binary classification with gene expression data. However, I envision this particular topic should become a prime area of very interesting and challenging research that could have a very large impact on the microarray classification realms.

In the area of performance assessment, it is important to keep in mind that in

performing external CV, the results are designed to assess the process of assessing predictive accuracy for binary classification of microarray data. That is, external CV is not inherently geared towards determining which specific group of d genes are the best predictors (for a given subset size d), since the feature selection process is built into each stage of the CV procedure and hence the particular genes selected at each stage will inevitably vary. Perhaps some form of gene monitoring operation could be developed such that specific subsets of genes can be identified.

An extension of the genetic algorithm aspect of this research is to implement a wider variety of objective functions within the genetic algorithm to see how the results may be affected by them. In particular, it would be interesting to see if any particular type of GA implementation yields subsets of genes that a) would not easily be discovered via univariate feature selection and b) would offer a significant advantage over univariate feature selection-based analyses in terms of minimal subset size and low misclassification error rates. Also, one could conduct further research investigating the probability of certain genes known to be strongly discriminatory *not* being selected by a GA in the final generation's solution sets. One parameter that warrants further study is the mutation rate and in particular how it relates to the solution (chromosome) size being pursued in a GA. For example, if the GA were designed to seek the best (i.e., most discriminatory) single-gene solution, one would think that a larger mutation rate should be used to allow for a wider population of the original gene pool to be given consideration during the evolution of the algorithm,

and vice versa as the solution size increases. Overall, I believe more in-depth research on how to optimally select GA parameters such as population size, crossover rate, and mutation rate would be a much appreciated and needed contribution to the GA community. Another extension of this research that would be interesting would be to consider the application of some type of aggregation of different GA-based classifiers to form a "meta-classifier." For example, using different objective functions within a GA, one could consider combining the results of GA/Mahalanobis distance, GA/SVM, and GA/k-NN analyses in some fashion.

Further investigation into the nature of the two-class problem could also prove to be useful. That is, one could more closely investigate the results obtained from problems in which the two classes are tumor and normal, to compare with a situation in which the two classes are subtypes of a particular type of cancer. Finally, beyond the realms of microarrays, the results and insights from this research can hopefully be applied to other binary classification problems marked by high-dimensional datasets with relatively small samples, and perhaps motivate a similar type of large-scale empirical study focused on working with more complex multi-class classification problems.

Appendix A

Other Results from Current Research

Table A.1: Raw and Adjusted P-values for Top 25 Genes; Alon Data

Gene Index	Raw	BY	Holm	Bonferroni	WY
493	2.38E-16	3.90E-12	4.77E-13	4.77E-13	2.00E-04
267	1.88E-15	1.54E-11	3.77E-12	3.77E-12	2.00E-04
245	3.51E-15	1.91E-11	7.01E-12	7.02E-12	2.00E-04
1423	8.71E-15	3.56E-11	1.74E-11	1.74E-11	2.00E-04
1635	3.74E-14	1.17E-10	7.47E-11	7.48E-11	2.00E-04
377	4.28E-14	1.17E-10	8.53E-11	8.55E-11	2.00E-04
1042	1.58E-13	3.42E-10	3.14E-10	3.15E-10	2.00E-04
780	1.67E-13	3.42E-10	3.34E-10	3.35E-10	2.00E-04
897	5.07E-13	9.22E-10	1.01E-09	1.02E-09	2.00E-04
765	9.13E-13	1.48E-09	1.82E-09	1.83E-09	2.00E-04
964	9.93E-13	1.48E-09	1.98E-09	1.99E-09	2.00E-04
1494	1.70E-12	2.31E-09	3.37E-09	3.39E-09	2.00E-04
1730	2.08E-12	2.61E-09	4.13E-09	4.15E-09	2.00E-04
1843	8.25E-12	9.64E-09	1.64E-08	1.65E-08	2.00E-04
1771	1.56E-11	1.70E-08	3.10E-08	3.12E-08	2.00E-04
365	1.93E-11	1.95E-08	3.82E-08	3.85E-08	2.00E-04
513	2.02E-11	1.95E-08	4.02E-08	4.05E-08	2.00E-04
1263	3.90E-11	3.54E-08	7.73E-08	7.80E-08	2.00E-04
138	1.05E-10	9.01E-08	2.07E-07	2.09E-07	2.00E-04
824	1.14E-10	9.30E-08	2.25E-07	2.28E-07	2.00E-04
249	1.26E-10	9.78E-08	2.49E-07	2.51E-07	2.00E-04
625	2.12E-10	1.58E-07	4.20E-07	4.24E-07	2.00E-04
1421	2.42E-10	1.72E-07	4.79E-07	4.84E-07	2.00E-04
1060	6.70E-10	4.57E-07	1.33E-06	1.34E-06	4.00E-04
1892	1.19E-09	7.78E-07	2.35E-06	2.38E-06	4.00E-04

Table A.2: Raw and Adjusted P-values for Top 25 Genes; Golub Data

Gene Index	Raw	BY	Holm	Bonferroni	WY
1834	5.782E-20	2.072E-15	4.122E-16	4.122E-16	2.000E-04
6855	6.151E-20	2.072E-15	4.385E-16	4.385E-16	2.000E-04
4847	1.771E-19	3.978E-15	1.263E-15	1.263E-15	2.000E-04
6041	5.906E-19	9.947E-15	4.209E-15	4.211E-15	2.000E-04
1882	9.517E-19	1.282E-14	6.781E-15	6.785E-15	2.000E-04
2354	9.459E-18	9.812E-14	6.738E-14	6.743E-14	2.000E-04
3252	1.020E-17	9.812E-14	7.262E-14	7.268E-14	2.000E-04
4377	2.618E-17	2.205E-13	1.865E-13	1.867E-13	2.000E-04
1685	1.521E-16	1.138E-12	1.083E-12	1.084E-12	2.000E-04
1144	5.289E-16	3.425E-12	3.766E-12	3.771E-12	2.000E-04
760	5.594E-16	3.425E-12	3.982E-12	3.988E-12	2.000E-04
1745	1.316E-15	7.387E-12	9.367E-12	9.381E-12	2.000E-04
2121	1.636E-15	8.480E-12	1.165E-11	1.167E-11	2.000E-04
4328	6.828E-15	3.285E-11	4.859E-11	4.867E-11	2.000E-04
4366	1.702E-14	7.643E-11	1.211E-10	1.213E-10	2.000E-04
5501	2.526E-14	1.063E-10	1.797E-10	1.801E-10	2.000E-04
4973	5.710E-14	2.263E-10	4.062E-10	4.071E-10	2.000E-04
1909	7.337E-14	2.746E-10	5.218E-10	5.231E-10	2.000E-04
6281	2.426E-13	8.603E-10	1.725E-09	1.730E-09	2.000E-04
2642	2.903E-13	9.778E-10	2.064E-09	2.070E-09	2.000E-04
4107	3.424E-13	1.098E-09	2.434E-09	2.441E-09	2.000E-04
1630	5.141E-13	1.574E-09	3.654E-09	3.665E-09	2.000E-04
804	8.550E-13	2.504E-09	6.076E-09	6.095E-09	2.000E-04
7119	9.136E-13	2.564E-09	6.492E-09	6.513E-09	2.000E-04
2020	1.012E-12	2.726E-09	7.189E-09	7.213E-09	2.000E-04

Table A.3: Raw and Adjusted P-values for Top 25 Genes; Nutt Data

Gene Index	Raw	BY	Holm	Bonferroni	WY
12194	5.836E-09	7.383E-04	7.368E-05	7.368E-05	1.200E-03
10017	2.415E-08	1.527E-03	3.048E-04	3.049E-04	2.000E-03
1614	5.178E-08	1.941E-03	6.536E-04	6.537E-04	4.000E-03
8501	6.138E-08	1.941E-03	7.748E-04	7.750E-04	5.000E-03
5908	7.734E-08	1.957E-03	9.761E-04	9.764E-04	5.600E-03
10943	9.877E-08	2.083E-03	1.247E-03	1.247E-03	6.200E-03
216	1.546E-07	2.795E-03	1.951E-03	1.952E-03	7.200E-03
1272	3.281E-07	5.188E-03	4.140E-03	4.142E-03	1.720E-02
221	4.167E-07	5.236E-03	5.258E-03	5.261E-03	2.200E-02
6223	4.343E-07	5.236E-03	5.480E-03	5.484E-03	1.920E-02
7367	4.561E-07	5.236E-03	5.754E-03	5.758E-03	2.200E-02
6682	4.967E-07	5.236E-03	6.265E-03	6.270E-03	2.420E-02
12195	6.850E-07	6.067E-03	8.640E-03	8.648E-03	2.760E-02
8495	7.013E-07	6.067E-03	8.845E-03	8.854E-03	2.920E-02
4574	7.194E-07	6.067E-03	9.072E-03	9.082E-03	2.660E-02
7645	7.785E-07	6.155E-03	9.816E-03	9.828E-03	3.460E-02
5212	1.033E-06	7.262E-03	1.302E-02	1.304E-02	4.280E-02
5682	1.079E-06	7.262E-03	1.360E-02	1.362E-02	3.480E-02
10681	1.091E-06	7.262E-03	1.375E-02	1.377E-02	4.460E-02
6681	1.329E-06	8.404E-03	1.675E-02	1.677E-02	4.820E-02
12012	1.493E-06	8.747E-03	1.882E-02	1.885E-02	5.040E-02
4723	1.521E-06	8.747E-03	1.917E-02	1.921E-02	5.460E-02
8397	1.590E-06	8.747E-03	2.004E-02	2.008E-02	5.660E-02
8455	1.719E-06	9.064E-03	2.167E-02	2.171E-02	5.800E-02
11837	2.222E-06	1.124E-02	2.800E-02	2.805E-02	6.720E-02

Table A.4: Raw and Adjusted P-values for Top 25 Genes; Pomeroy Data

Gene Index	Raw	BY	Holm	Bonferroni	WY
3127	2.326E-14	8.843E-10	1.658E-10	1.658E-10	2.000E-04
2365	2.625E-14	8.843E-10	1.871E-10	1.872E-10	2.000E-04
1422	4.498E-14	1.010E-09	3.206E-10	3.207E-10	2.000E-04
2322	2.979E-13	5.017E-09	2.123E-09	2.124E-09	2.000E-04
2967	1.738E-12	2.341E-08	1.238E-08	1.239E-08	2.000E-04
4457	3.422E-12	3.652E-08	2.438E-08	2.440E-08	2.000E-04
6512	3.795E-12	3.652E-08	2.703E-08	2.705E-08	2.000E-04
2511	4.950E-12	4.168E-08	3.526E-08	3.529E-08	2.000E-04
4484	7.940E-12	5.808E-08	5.654E-08	5.661E-08	2.000E-04
6718	8.622E-12	5.808E-08	6.139E-08	6.146E-08	2.000E-04
5585	1.028E-11	6.294E-08	7.316E-08	7.327E-08	2.000E-04
6435	1.177E-11	6.355E-08	8.379E-08	8.392E-08	2.000E-04
5669	1.262E-11	6.355E-08	8.980E-08	8.995E-08	2.000E-04
3136	1.383E-11	6.355E-08	9.839E-08	9.857E-08	2.000E-04
3525	1.415E-11	6.355E-08	1.007E-07	1.009E-07	2.000E-04
2545	1.626E-11	6.474E-08	1.156E-07	1.159E-07	2.000E-04
6625	1.634E-11	6.474E-08	1.162E-07	1.165E-07	2.000E-04
6181	2.643E-11	9.890E-08	1.879E-07	1.884E-07	2.000E-04
2344	4.980E-11	1.766E-07	3.541E-07	3.550E-07	2.000E-04
4632	6.317E-11	2.105E-07	4.491E-07	4.503E-07	2.000E-04
5581	6.562E-11	2.105E-07	4.665E-07	4.678E-07	2.000E-04
3032	7.643E-11	2.340E-07	5.432E-07	5.449E-07	2.000E-04
3001	9.046E-11	2.650E-07	6.429E-07	6.449E-07	2.000E-04
588	1.058E-10	2.970E-07	7.518E-07	7.542E-07	2.000E-04
3043	1.304E-10	3.513E-07	9.263E-07	9.294E-07	2.000E-04

Table A.5: Raw and Adjusted P-values for Top 25 Genes; Shipp Data

Gene Index	Raw	BY	Holm	Bonferroni	WY
972	2.235E-16	1.506E-11	1.594E-12	1.594E-12	2.000E-04
4569	7.299E-16	2.055E-11	5.203E-12	5.204E-12	2.000E-04
4194	1.199E-15	2.055E-11	8.545E-12	8.548E-12	2.000E-04
506	1.231E-15	2.055E-11	8.772E-12	8.776E-12	2.000E-04
4028	1.525E-15	2.055E-11	1.087E-11	1.087E-11	2.000E-04
2988	2.786E-15	3.128E-11	1.985E-11	1.986E-11	2.000E-04
5386	5.256E-15	5.058E-11	3.744E-11	3.747E-11	2.000E-04
699	7.482E-15	6.300E-11	5.328E-11	5.334E-11	2.000E-04
4292	1.744E-14	1.305E-10	1.242E-10	1.243E-10	2.000E-04
1092	2.950E-14	1.987E-10	2.101E-10	2.103E-10	2.000E-04
6815	4.955E-14	3.034E-10	3.527E-10	3.532E-10	2.000E-04
605	7.921E-14	4.447E-10	5.638E-10	5.647E-10	2.000E-04
6179	9.972E-14	5.168E-10	7.097E-10	7.109E-10	2.000E-04
2043	3.096E-13	1.490E-09	2.203E-09	2.207E-09	2.000E-04
2137	3.726E-13	1.585E-09	2.651E-09	2.656E-09	2.000E-04
4372	3.764E-13	1.585E-09	2.678E-09	2.684E-09	2.000E-04
1080	5.327E-13	2.111E-09	3.789E-09	3.798E-09	2.000E-04
4183	1.530E-12	5.726E-09	1.088E-08	1.091E-08	2.000E-04
5994	2.142E-12	7.594E-09	1.523E-08	1.527E-08	2.000E-04
2121	2.323E-12	7.824E-09	1.652E-08	1.656E-08	2.000E-04
1790	2.508E-12	8.047E-09	1.783E-08	1.788E-08	2.000E-04
1352	2.789E-12	8.432E-09	1.983E-08	1.988E-08	2.000E-04
1612	2.879E-12	8.432E-09	2.046E-08	2.052E-08	2.000E-04
6476	3.044E-12	8.544E-09	2.163E-08	2.170E-08	2.000E-04
1188	3.724E-12	9.951E-09	2.646E-08	2.655E-08	2.000E-04

Table A.6: Raw and Adjusted P-values for Top 25 Genes; Singh Data

Gene Index	Raw	BY	Holm	Bonferroni	WY
6185	3.669E-24	4.632E-19	4.623E-20	4.623E-20	2.000E-04
10494	6.929E-17	4.373E-12	8.729E-13	8.730E-13	2.000E-04
9850	1.740E-16	5.719E-12	2.193E-12	2.193E-12	2.000E-04
4365	1.812E-16	5.719E-12	2.283E-12	2.283E-12	2.000E-04
10138	1.141E-15	2.880E-11	1.437E-11	1.437E-11	2.000E-04
9172	1.996E-15	4.200E-11	2.514E-11	2.515E-11	2.000E-04
5944	6.554E-15	1.117E-10	8.254E-11	8.258E-11	2.000E-04
9034	7.080E-15	1.117E-10	8.915E-11	8.920E-11	2.000E-04
3649	4.314E-14	6.051E-10	5.432E-10	5.436E-10	2.000E-04
2839	6.507E-14	7.687E-10	8.192E-10	8.198E-10	2.000E-04
8554	6.698E-14	7.687E-10	8.433E-10	8.440E-10	2.000E-04
7557	9.176E-14	9.653E-10	1.155E-09	1.156E-09	2.000E-04
205	2.395E-13	2.326E-09	3.015E-09	3.018E-09	2.000E-04
3794	3.493E-13	3.121E-09	4.397E-09	4.402E-09	2.000E-04
10956	3.708E-13	3.121E-09	4.667E-09	4.673E-09	2.000E-04
8850	5.905E-13	4.659E-09	7.432E-09	7.441E-09	2.000E-04
7520	8.218E-13	6.102E-09	1.034E-08	1.035E-08	2.000E-04
9050	1.012E-12	7.095E-09	1.273E-08	1.275E-08	2.000E-04
10537	1.617E-12	1.074E-08	2.034E-08	2.037E-08	2.000E-04
5757	1.999E-12	1.262E-08	2.515E-08	2.519E-08	2.000E-04
8123	2.951E-12	1.774E-08	3.713E-08	3.719E-08	2.000E-04
8768	5.099E-12	2.926E-08	6.414E-08	6.424E-08	2.000E-04
6462	7.598E-12	4.170E-08	9.557E-08	9.574E-08	2.000E-04
7768	1.562E-11	8.218E-08	1.965E-07	1.969E-07	2.000E-04
7247	1.691E-11	8.539E-08	2.127E-07	2.131E-07	2.000E-04

Table A.7: Gene Selection Based on All Samples: Alon Data (a)

	GA			GA-GA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pval**	**Univ Rank**
1	493	2.38E-16	1	493	2.38E-16	1
2	66	3.70E-09	31	66	3.70E-09	31
	1423	8.71E-15	4	1423	8.71E-15	4
3	1058	5.69E-04	274	1058	5.69E-04	274
	1423	8.71E-15	4	1423	8.71E-15	4
	1484	1.69E-03	339	1484	1.69E-03	339
4	66	3.70E-09	31	66	3.70E-09	31
	1058	5.69E-04	274	1058	5.69E-04	274
	1423	8.71E-15	4	1423	8.71E-15	4
	1484	1.69E-03	339	1484	1.69E-03	339
5	267	1.88E-15	2	26	1.80E-07	54
	377	4.28E-14	6	377	4.28E-14	6
	441	3.14E-01	1557	624	7.26E-04	288
	493	2.38E-16	1	1286	2.33E-05	153
	1836	1.15E-07	49	1423	8.71E-15	4
10	245	3.51E-15	3	66	3.70E-09	31
	339	8.28E-03	478	93	2.06E-02	615
	493	2.38E-16	1	245	3.51E-15	3
	792	1.26E-04	201	258	1.22E-02	529
	895	1.20E-02	525	622	6.24E-04	279
	1135	3.32E-02	695	882	6.61E-05	183
	1445	4.32E-01	1841	895	1.20E-02	525
	1567	3.94E-03	415	1565	2.70E-01	1447
	1585	4.73E-01	1934	1836	1.15E-07	49
	1836	1.15E-07	49	1873	1.12E-06	78
15	13	4.16E-04	256	13	4.16E-04	256
	43	1.00E-06	77	79	2.39E-01	1373
	128	4.38E-01	1859	267	1.88E-15	2
	164	1.44E-02	559	377	4.28E-14	6
	190	1.67E-04	210	409	2.29E-01	1350
	267	1.88E-15	2	624	7.26E-04	288
	493	2.38E-16	1	769	3.01E-01	1522
	792	1.26E-04	201	897	5.07E-13	9
	897	5.07E-13	9	1052	2.94E-01	1511
	1024	7.98E-02	922	1058	5.69E-04	274
	1082	6.42E-02	851	1400	1.19E-03	316
	1150	3.52E-01	1653	1423	8.71E-15	4
	1676	7.29E-02	891	1522	3.34E-01	1604
	1819	5.76E-02	823	1653	3.22E-01	1576
	1969	7.35E-02	893	1873	1.12E-06	78

Table A.8: Gene Selection Based on All Samples: Alon Data (b)

	GA			GAGA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pvbal**	**Univ Rank**
20	66	3.70E-09	31	104	2.90E-02	668
	141	5.09E-06	107	137	1.36E-08	37
	213	2.06E-02	614	213	2.06E-02	614
	234	4.68E-01	1927	533	2.48E-01	1393
	431	4.94E-02	783	542	3.86E-01	1730
	572	2.96E-02	674	624	7.26E-04	288
	581	3.79E-08	41	895	1.20E-02	525
	624	7.26E-04	288	897	5.07E-13	9
	1058	5.69E-04	274	1058	5.69E-04	274
	1103	3.79E-01	1716	1225	4.10E-01	1786
	1186	7.00E-04	284	1295	4.74E-01	1939
	1295	4.74E-01	1939	1306	1.65E-03	337
	1423	8.71E-15	4	1423	8.71E-15	4
	1597	1.43E-03	329	1571	4.00E-01	1769
	1647	1.63E-01	1163	1644	3.94E-02	729
	1713	5.90E-02	832	1668	4.71E-02	770
	1790	7.48E-03	472	1808	2.06E-06	95
	1808	2.06E-06	95	1829	3.67E-01	1692
	1858	1.87E-01	1231	1851	1.31E-01	1082
	1873	1.12E-06	78	1873	1.12E-06	78
25	13	4.16E-04	256	13	4.16E-04	256
	31	3.31E-07	63	189	2.92E-03	390
	66	3.70E-09	31	190	1.67E-04	210
	188	1.84E-03	345	270	1.45E-01	1109
	634	2.85E-02	665	326	2.50E-02	645
	792	1.26E-04	201	495	1.30E-06	80
	970	1.42E-01	1103	537	1.08E-01	1006
	982	3.18E-04	241	623	5.00E-01	1999
	983	4.73E-01	1933	624	7.26E-04	288
	1018	1.73E-01	1193	659	2.78E-01	1474
	1049	5.68E-02	817	734	2.84E-03	387
	1050	4.65E-02	769	879	1.68E-01	1174
	1102	7.22E-02	887	897	5.07E-13	9
	1158	2.71E-01	1451	931	2.81E-05	157
	1239	5.21E-02	799	1058	5.69E-04	274
	1243	2.28E-01	1348	1085	2.93E-01	1506
	1342	4.95E-02	786	1136	6.07E-04	277
	1423	8.71E-15	4	1152	6.83E-02	872
	1440	2.65E-02	653	1346	1.91E-03	348
	1563	6.25E-03	456	1423	8.71E-15	4
	1629	4.32E-01	1838	1478	4.45E-02	756
	1634	1.54E-08	38	1593	8.53E-02	941
	1654	4.25E-01	1821	1724	4.70E-05	177
	1672	9.54E-05	196	1859	4.59E-03	426
	1927	2.93E-01	1508	1873	1.12E-06	78

Table A.9: Gene Selection Based on All Samples: Golub Data (a)

	GA			GA-GA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pval**	**Univ Rank**
1	4847	1.77E-19	3	4847	1.77E-19	3
2	1834	5.78E-20	1	1834	5.78E-20	1
	6539	3.55E-10	58	6539	3.55E-10	58
3	1775	1.57E-02	1634	1779	4.73E-12	32
	4847	1.77E-19	3	1941	4.37E-03	1169
	4951	4.85E-09	85	4847	1.77E-19	3
4	1779	4.73E-12	32	1745	1.32E-15	12
	2121	1.64E-15	13	1779	4.73E-12	32
	5949	3.98E-01	6115	1796	3.75E-01	5895
	6277	1.26E-02	1525	2121	1.64E-15	13
5	1456	1.58E-04	551	1745	1.32E-15	12
	1779	4.73E-12	32	1779	4.73E-12	32
	1796	3.75E-01	5895	1796	3.75E-01	5895
	1834	5.78E-20	1	1834	5.78E-20	1
	4847	1.77E-19	3	1882	9.52E-19	5
10	1745	1.32E-15	12	1779	4.73E-12	32
	1779	4.73E-12	32	1796	3.75E-01	5895
	1796	3.75E-01	5895	1829	1.12E-11	39
	1829	1.12E-11	39	1975	8.12E-03	1350
	2221	3.46E-01	5629	2426	1.23E-05	343
	2288	1.91E-12	27	3847	5.27E-10	64
	2681	4.70E-01	6832	4200	3.94E-04	674
	3795	5.99E-02	2479	4847	1.77E-19	3
	4642	1.16E-01	3202	5778	2.89E-02	1950
	4664	7.49E-04	763	6184	1.21E-07	155
15	325	3.20E-02	2019	877	4.75E-01	6897
	1779	4.73E-12	32	905	2.19E-01	4325
	1796	3.75E-01	5895	1779	4.73E-12	32
	1829	1.12E-11	39	1796	3.75E-01	5895
	1834	5.78E-20	1	1834	5.78E-20	1
	1882	9.52E-19	5	1882	9.52E-19	5
	1963	8.79E-02	2875	2242	2.79E-05	398
	2242	2.79E-05	398	2317	2.65E-01	4778
	2643	2.13E-01	4262	2337	4.49E-01	6618
	3551	4.20E-02	2192	2725	1.61E-01	3701
	3892	1.68E-01	3784	3191	2.51E-02	1869
	4642	1.16E-01	3202	4746	2.02E-01	4157
	4947	1.29E-02	1533	4863	4.84E-01	6979
	6041	5.91E-19	4	5307	6.66E-02	2572
	6401	2.89E-01	5039	6012	1.34E-03	864

Table A.10: Gene Selection Based on All Samples: Golub Data (b)

	GA			GAGA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pvbal**	**Univ Rank**
20	94	1.78E-01	3894	49	9.82E-03	1424
	348	4.84E-01	6966	368	1.52E-03	886
	1745	1.32E-15	12	1109	1.65E-02	1652
	1779	4.73E-12	32	1779	4.73E-12	32
	1796	3.75E-01	5895	1796	3.75E-01	5895
	1834	5.78E-20	1	1834	5.78E-20	1
	1891	2.92E-04	624	1891	2.92E-04	624
	2121	1.64E-15	13	1941	4.37E-03	1169
	2288	1.91E-12	27	2121	1.64E-15	13
	2635	1.95E-01	4080	2240	1.93E-01	4049
	3320	1.45E-06	244	2288	1.91E-12	27
	3813	2.04E-01	4183	2389	1.57E-04	549
	3886	7.58E-02	2695	2635	1.95E-01	4080
	4341	6.16E-02	2503	3921	3.56E-01	5713
	4840	3.22E-01	5376	4246	3.47E-01	5638
	4847	1.77E-19	3	4847	1.77E-19	3
	5182	6.64E-02	2567	5599	2.24E-03	994
	5252	4.55E-01	6677	5742	8.43E-03	1365
	5395	2.09E-01	4222	5934	8.33E-02	2808
	5481	3.47E-01	5639.5	6184	1.21E-07	155
25	381	2.88E-01	5025	727	4.69E-02	2281
	424	2.40E-01	4538	1307	1.49E-01	3571
	487	1.17E-01	3215	1477	2.08E-01	4217
	605	6.28E-02	2524	1674	3.02E-07	185
	774	3.14E-05	401	1771	5.54E-02	2417
	1107	4.43E-01	6555	1779	4.73E-12	32
	1779	4.73E-12	32	1796	3.75E-01	5895
	1796	3.75E-01	5895	1834	5.78E-20	1
	1829	1.12E-11	39	1882	9.52E-19	5
	1977	4.49E-02	2249	2121	1.64E-15	13
	2497	6.37E-09	94	2426	1.23E-05	343
	3576	3.56E-03	1121	2518	4.51E-01	6635
	3714	4.34E-07	199	2752	3.93E-01	6071
	3737	9.96E-02	3013	3714	4.34E-07	199
	4166	3.27E-02	2033	3847	5.27E-10	64
	4197	5.08E-05	435	4847	1.77E-19	3
	4200	3.94E-04	674	5048	8.37E-02	2811
	4217	4.71E-01	6860	5065	3.44E-02	2071
	4601	3.00E-01	5148	5102	1.94E-01	4067
	4847	1.77E-19	3	5364	1.35E-01	3402
	5206	4.27E-01	6403	5432	1.61E-02	1640
	5357	4.24E-01	6376	6043	3.21E-01	5359
	5552	2.97E-09	77	6184	1.21E-07	155
	6283	2.13E-11	43	6283	2.13E-11	43
	6848	4.88E-01	7010	6845	1.67E-01	3776

Table A.11: Gene Selection Based on All Samples: Nutt Data (a)

	GA			GA-GA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pval**	**Univ Rank**
1	1614	5.18E-08	3	1614	5.18E-08	3
2	2202	3.05E-06	33	2202	3.05E-06	33
	10017	2.41E-08	2	10017	2.41E-08	2
3	1272	3.28E-07	8	1614	5.18E-08	3
	9780	2.89E-04	239	4066	7.75E-05	125
	11432	2.86E-06	30	12359	1.41E-01	5242
4	7855	6.98E-06	46	1614	5.18E-08	3
	9299	3.76E-03	836	8691	2.19E-04	206
	10474	2.95E-02	2261	9299	3.76E-03	836
	12154	4.15E-03	872	12154	4.15E-03	872
5	4257	4.53E-05	107	1614	5.18E-08	3
	6171	1.76E-02	1775	4257	4.53E-05	107
	8691	2.19E-04	206	8691	2.19E-04	206
	9299	3.76E-03	836	9299	3.76E-03	836
	10112	5.63E-03	1001	10118	6.17E-03	1066
10	398	2.22E-02	1979	690	1.57E-02	1662
	666	8.57E-02	3977	723	4.39E-05	103
	1614	5.18E-08	3	1058	8.94E-02	4075
	2099	4.52E-01	11654	2148	7.32E-02	3660
	2202	3.05E-06	33	5244	1.24E-04	160
	3805	1.95E-04	190	5579	3.37E-01	9380
	8526	1.43E-01	5294	8334	2.35E-01	7311
	10926	2.68E-03	722	11280	1.96E-03	628
	12194	5.84E-09	1	11618	3.69E-02	2555
	12569	4.09E-01	10784	11821	1.14E-05	57
15	93	3.60E-01	9837	1614	5.18E-08	3
	618	1.98E-01	6508	1727	1.83E-01	6204
	3086	1.00E-02	1348	3418	2.02E-02	1879
	3096	1.38E-01	5176	3748	1.32E-02	1537
	3748	1.32E-02	1537	4257	4.53E-05	107
	5398	1.98E-04	193	4575	4.82E-02	2915
	6785	2.00E-04	196	5244	1.24E-04	160
	7121	3.95E-01	10535	6081	2.11E-05	77
	8846	1.86E-02	1810	6708	1.25E-02	1492
	9392	3.11E-01	8856	7077	1.84E-01	6229
	9783	7.68E-06	48	10118	6.17E-03	1066
	10261	7.23E-02	3618	10494	2.69E-01	7972
	11196	2.27E-01	7133	11196	2.27E-01	7133
	11376	6.82E-05	119	11376	6.82E-05	119
	11404	8.48E-02	3959	11515	4.12E-02	2699

Table A.12: Gene Selection Based on All Samples: Nutt Data (b)

	GA			GAGA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pvbal**	**Univ Rank**
20	334	6.43E-04	373	1614	5.18E-08	3
	723	4.39E-05	103	2068	3.80E-01	10235
	1068	8.14E-04	414	3250	8.26E-04	419
	2088	3.03E-01	8682	4066	7.75E-05	125
	3555	2.54E-01	7675	6049	4.24E-01	11099
	3794	2.61E-01	7823	6144	1.38E-01	5167
	4575	4.82E-02	2915	6162	9.50E-03	1305
	4689	1.75E-03	586	6876	1.04E-01	4391
	5000	7.58E-02	3728	7215	1.95E-02	1852
	5478	1.17E-01	4660	7741	1.50E-02	1627.5
	6162	9.50E-03	1305	8412	1.21E-01	4753
	7064	2.37E-01	7371	8432	2.52E-02	2086
	7477	2.43E-05	84	8589	7.73E-02	3771
	8054	5.89E-02	3240	8696	2.52E-04	226
	8101	2.32E-02	2020	8948	3.40E-01	9432.5
	8696	2.52E-04	226	9830	2.26E-01	7105
	10118	6.17E-03	1066	11280	1.96E-03	628
	12517	6.79E-02	3487	11376	6.82E-05	119
	12542	2.90E-01	8422	12359	1.41E-01	5242
	12567	4.62E-01	11834	12581	3.62E-01	9863
25	142	3.17E-01	8970.5	1038	4.69E-01	11989
	723	4.39E-05	103	1377	6.98E-02	3537
	2551	2.42E-01	7454	1545	3.74E-02	2574
	3588	4.03E-01	10706	1687	3.45E-01	9523
	4390	4.84E-01	12270	2570	3.92E-02	2630
	5510	3.04E-01	8693	4005	1.93E-01	6413
	5830	2.63E-01	7861	4066	7.75E-05	125
	6715	8.02E-02	3843	4257	4.53E-05	107
	7477	2.43E-05	84	4384	1.97E-01	6481
	7667	2.13E-01	6826	5475	4.69E-01	11995
	8149	4.92E-05	108	5660	3.29E-01	9220
	8223	4.70E-01	12098	6510	4.84E-02	2923
	8274	7.74E-02	3774	7058	7.11E-02	3587
	8393	1.05E-01	4400	7785	1.47E-01	5392
	8943	1.66E-02	1719	8352	1.53E-01	5531
	9368	1.36E-03	533	8455	1.72E-06	24
	9420	2.16E-02	1956	8803	4.25E-01	11127
	9522	7.84E-03	1200	9367	2.75E-01	8093
	10078	1.96E-03	630	9510	1.53E-01	5543
	10596	2.77E-01	8134	9637	1.22E-05	59
	11022	2.38E-01	7388	10395	3.56E-01	9741
	11406	4.39E-01	11366	10501	5.00E-01	12587
	11919	3.85E-01	10335	11482	3.48E-03	814
	12130	1.04E-02	1364	11502	4.92E-01	12428
	12272	3.36E-01	9353	11711	1.17E-01	4653

Table A.13: Gene Selection Based on All Samples: Pomeroy Data (a)

	GA			GA-GA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pval**	**Univ Rank**
1	379	7.83E-09	66	379	7.83E-09	66
2	2967	1.74E-12	5	2967	1.74E-12	5
	3127	2.33E-14	1	3127	2.33E-14	1
3	1783	2.33E-08	83	1040	2.47E-02	3417
	3127	2.33E-14	1	1746	1.69E-04	971
	6377	6.64E-05	721	3681	8.49E-04	1560
4	14	8.20E-07	193	3127	2.33E-14	1
	2545	1.63E-11	16	4031	2.06E-06	249
	3127	2.33E-14	1	6377	6.64E-05	721
	6377	6.64E-05	721	6432	3.20E-08	86
5	576	4.38E-07	164	2830	3.55E-05	592
	2511	4.95E-12	8	3080	2.61E-04	1095
	3258	1.68E-07	125	3127	2.33E-14	1
	6401	1.99E-07	133	6377	6.64E-05	721
	6967	8.65E-02	4484	6432	3.20E-08	86
10	527	1.20E-05	408	1407	5.53E-04	1363
	692	2.95E-01	6087	1788	2.46E-01	5750
	1389	9.08E-03	2762	2203	3.78E-02	3754
	2365	2.63E-14	2	3127	2.33E-14	1
	3127	2.33E-14	1	4031	2.06E-06	249
	4031	2.06E-06	249	4134	2.46E-02	3413
	5664	1.71E-07	126	4701	2.01E-06	247
	6432	3.20E-08	86	5664	1.71E-07	126
	6971	5.07E-03	2419	6377	6.64E-05	721
	7097	1.62E-01	5175	6432	3.20E-08	86
15	423	2.41E-03	2000	702	6.96E-04	1464
	1286	1.16E-03	1691	1951	1.02E-09	36
	1520	9.28E-02	4563	2203	3.78E-02	3754
	1951	1.02E-09	36	2365	2.63E-14	2
	2135	3.75E-02	3747	3097	2.30E-03	1983
	2203	3.78E-02	3754	3127	2.33E-14	1
	2830	3.55E-05	592	3128	8.25E-02	4437
	2967	1.74E-12	5	3563	2.49E-03	2026
	3043	1.30E-10	25	4454	1.24E-01	4863
	4388	4.97E-01	7109	4475	1.14E-01	4772
	4482	2.00E-01	5463	5585	1.03E-11	11
	4912	5.34E-02	4034	6377	6.64E-05	721
	5585	1.03E-11	11	6432	3.20E-08	86
	6314	4.57E-01	6921	6801	1.87E-04	998
	6432	3.20E-08	86	7121	2.31E-08	81

Table A.14: Gene Selection Based on All Samples: Pomeroy Data (b)

	GA			GAGA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pvbal**	**Univ Rank**
20	27	5.61E-02	4076	249	8.50E-03	2716
	412	1.98E-01	5450	496	2.64E-05	534
	546	4.47E-08	95	529	1.79E-08	77
	576	4.38E-07	164	1267	1.63E-01	5182
	1040	2.47E-02	3417	1286	1.16E-03	1691
	1746	1.69E-04	971	1520	9.28E-02	4563
	1776	4.45E-02	3885	2203	3.78E-02	3754
	2062	3.97E-01	6623	2964	2.28E-01	5638
	2203	3.78E-02	3754	3220	5.38E-02	4037
	2354	1.41E-09	41	3306	1.21E-05	411
	2376	2.33E-05	512	4340	3.83E-01	6559
	2568	7.97E-02	4393	4407	1.33E-01	4944
	4378	2.69E-01	5902	4499	9.28E-07	201
	4380	1.05E-03	1643	4578	5.20E-03	2433
	4623	4.10E-02	3821	5240	3.59E-04	1201
	5585	1.03E-11	11	5585	1.03E-11	11
	5804	1.07E-03	1651	6068	5.06E-02	3989
	5867	3.85E-01	6570	6377	6.64E-05	721
	6444	4.22E-02	3845	6432	3.20E-08	86
	6914	2.44E-05	519	7121	2.31E-08	81
25	549	4.81E-05	644	27	5.61E-02	4076
	716	3.18E-02	3613	141	4.81E-01	7033
	1142	1.87E-01	5361	174	7.27E-02	4314
	1233	7.73E-02	4364	496	2.64E-05	534
	1520	9.28E-02	4563	1110	9.54E-09	71
	1712	2.37E-02	3387	1121	1.29E-02	2994
	2203	3.78E-02	3754	1319	3.01E-02	3564
	2262	3.08E-06	283	2033	2.41E-04	1066
	2511	4.95E-12	8	2076	7.87E-03	2668
	3032	7.64E-11	22	2203	3.78E-02	3754
	3343	2.75E-01	5954	2356	2.74E-01	5949
	3466	8.08E-03	2679	2511	4.95E-12	8
	3804	8.98E-02	4533	2855	1.19E-01	4817
	4378	2.69E-01	5902	3220	5.38E-02	4037
	4831	3.26E-07	147	4172	2.40E-09	47
	4951	4.09E-02	3818	4299	2.57E-02	3449
	5585	1.03E-11	11	4745	1.15E-03	1685
	5928	1.60E-04	955	4931	3.64E-02	3722
	6343	3.00E-03	2122	5211	1.27E-03	1726
	6377	6.64E-05	721	5603	1.26E-03	1723
	6432	3.20E-08	86	5795	2.72E-01	5924
	6435	1.18E-11	12	6064	4.56E-01	6914
	6625	1.63E-11	17	6321	4.09E-01	6694
	6778	1.21E-01	4828	6377	6.64E-05	721
	7121	2.31E-08	81	6748	1.26E-09	38

Table A.15: Gene Selection Based on All Samples: Shipp Data (a)

	GA			GA-GA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pval**	**Univ Rank**
1	699	7.48E-15	8	699	7.48E-15	8
2	699	7.48E-15	8	699	7.48E-15	8
	3053	5.40E-05	639	5077	7.95E-10	67
3	555	1.34E-08	102	555	1.34E-08	102
	6725	9.72E-08	152	2609	1.56E-08	106
	7102	6.33E-08	140	4194	1.20E-15	3
4	555	1.34E-08	102	555	1.34E-08	102
	6040	5.99E-09	84	2609	1.56E-08	106
	6725	9.72E-08	152	6725	9.72E-08	152
	7102	6.33E-08	140	7102	6.33E-08	140
5	555	1.34E-08	102	555	1.34E-08	102
	1818	6.13E-11	42	2609	1.56E-08	106
	2609	1.56E-08	106	4194	1.20E-15	3
	4194	1.20E-15	3	5233	2.37E-04	975
	5518	1.69E-02	3463	6141	7.99E-05	711
10	555	1.34E-08	102	555	1.34E-08	102
	613	3.26E-10	58	613	3.26E-10	58
	1685	2.79E-02	3851	1685	2.79E-02	3851
	2337	9.29E-02	4925	2173	1.39E-01	5347
	2937	6.30E-10	63	2609	1.56E-08	106
	4194	1.20E-15	3	3984	7.65E-02	4756
	4454	7.48E-03	2848	4194	1.20E-15	3
	4497	2.17E-01	5887	5233	2.37E-04	975
	6079	3.34E-04	1072	6079	3.34E-04	1072
	7102	6.33E-08	140	7102	6.33E-08	140
15	555	1.34E-08	102	555	1.34E-08	102
	699	7.48E-15	8	699	7.48E-15	8
	991	8.38E-05	720	801	4.50E-05	601
	1077	5.99E-04	1294	1585	1.65E-08	108
	1585	1.65E-08	108	2006	2.35E-08	115
	3259	1.96E-01	5750	2079	9.53E-02	4952
	3285	3.45E-06	318	2226	1.35E-02	3305
	3429	8.60E-04	1442	2524	1.69E-02	3462
	4934	1.75E-03	1841	2937	6.30E-10	63
	5198	6.76E-02	4625	3033	2.69E-01	6155
	5932	9.20E-02	4914	3787	2.28E-01	5945
	6323	7.47E-02	4728	4580	9.75E-09	98
	6337	9.31E-06	412	5198	6.76E-02	4625
	6493	3.20E-08	122	5940	2.50E-04	995
	6814	8.29E-03	2930	6432	1.14E-01	5142

Table A.16: Gene Selection Based on All Samples: Shipp Data (b)

	GA			GAGA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pvbal**	**Univ Rank**
20	426	4.66E-03	2487	219	9.73E-02	4982
	486	4.60E-06	341	555	1.34E-08	102
	555	1.34E-08	102	699	7.48E-15	8
	613	3.26E-10	58	726	1.53E-01	5453
	657	1.39E-03	1702	899	2.84E-05	534
	1309	4.39E-01	6855	2006	2.35E-08	115
	2267	3.02E-01	6297	3212	8.24E-05	715
	2396	2.61E-06	299	3395	1.33E-01	5294
	2477	1.02E-01	5039	3647	3.09E-01	6333
	2893	7.77E-02	4768	3787	2.28E-01	5945
	2937	6.30E-10	63	4114	2.26E-02	3685
	3212	8.24E-05	715	4218	1.11E-02	3161
	4194	1.20E-15	3	4236	2.81E-04	1021
	4307	2.46E-06	295	5198	6.76E-02	4625
	4903	9.93E-06	416	5877	2.56E-01	6092
	4943	6.81E-03	2780	5976	4.21E-01	6786
	6731	6.69E-02	4614	6141	7.99E-05	711
	6967	8.33E-03	2938	6337	9.31E-06	412
	7024	1.45E-01	5390	6493	3.20E-08	122
	7102	6.33E-08	140	6927	5.94E-03	2664
25	86	1.56E-04	857	22	5.96E-03	2665
	699	7.48E-15	8	122	6.54E-03	2756
	801	4.50E-05	601	373	2.37E-08	116
	1389	3.41E-02	4026	439	4.51E-01	6909
	1437	4.38E-03	2445	555	1.34E-08	102
	1910	1.53E-01	5455	613	3.26E-10	58
	2006	2.35E-08	115	749	1.36E-02	3314
	2482	4.90E-06	348	1437	4.38E-03	2445
	2714	4.58E-01	6942	1909	2.78E-02	3848
	3023	3.94E-01	6664	2142	1.01E-02	3089
	3131	3.31E-01	6442	2990	2.96E-02	3906
	3215	3.07E-04	1056	2995	2.50E-02	3758
	3285	3.45E-06	318	3086	8.44E-06	401
	4010	4.56E-06	339	3515	3.27E-02	3989
	4118	1.44E-03	1720	4137	8.43E-02	4830
	4397	1.27E-06	246	4469	1.52E-01	5448
	5158	5.99E-03	2671	5086	3.85E-02	4137
	5198	6.76E-02	4625	5198	6.76E-02	4625
	5497	7.87E-02	4773	5233	2.37E-04	975
	6011	2.80E-03	2147	5601	2.49E-01	6053
	6136	9.14E-04	1473	5846	6.54E-04	1328
	6573	8.65E-06	403	6243	2.92E-02	3899
	6611	1.85E-03	1872	6725	9.72E-08	152
	6746	2.38E-02	3716	6746	2.38E-02	3716
	7102	6.33E-08	140	7102	6.33E-08	140

Table A.17: Gene Selection Based on All Samples: Singh Data (a)

	GA			GA-GA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pval**	**Univ Rank**
1	9050	1.01E-12	18	9050	1.01E-12	18
2	205	2.40E-13	13	205	2.40E-13	13
	7520	8.22E-13	17	7520	8.22E-13	17
3	6185	3.67E-24	1	6185	3.67E-24	1
	7768	1.56E-11	24	7768	1.56E-11	24
	11942	2.12E-10	45	10234	1.30E-07	165
4	6185	3.67E-24	1	5045	2.46E-09	85
	8092	5.50E-03	2861	6185	3.67E-24	1
	10234	1.30E-07	165	10234	1.30E-07	165
	11871	4.18E-09	98	12067	3.63E-04	1205
5	1247	2.10E-02	4393	205	2.40E-13	13
	6185	3.67E-24	1	1247	2.10E-02	4393
	6323	9.91E-04	1651	6185	3.67E-24	1
	8965	7.82E-11	35	6323	9.91E-04	1651
	10234	1.30E-07	165	10234	1.30E-07	165
10	2607	1.10E-02	3598	1561	3.46E-03	2461
	3617	1.68E-03	1976	2694	3.75E-01	11108
	8200	4.99E-11	32	4767	5.44E-05	687
	8965	7.82E-11	35	4899	8.49E-07	233
	9860	3.98E-01	11351	5045	2.46E-09	85
	10234	1.30E-07	165	6185	3.67E-24	1
	11190	1.82E-02	4218	8965	7.82E-11	35
	11871	4.18E-09	98	9002	5.78E-03	2914
	11942	2.12E-10	45	9093	2.25E-11	26
	11947	4.61E-01	12180	11588	2.22E-01	9138
15	205	2.40E-13	13	497	1.45E-01	7918
	1107	9.78E-02	6981	1308	1.98E-01	8775
	2714	1.96E-01	8738	2434	2.70E-01	9786
	3113	1.89E-01	8638	3062	6.64E-03	3048
	4613	4.74E-01	12308	4525	1.32E-08	112
	6185	3.67E-24	1	5045	2.46E-09	85
	6539	2.25E-01	9195	5862	4.13E-08	132
	7362	1.68E-01	8289	6185	3.67E-24	1
	8603	2.78E-01	9884	6838	2.11E-04	1015
	10234	1.30E-07	165	8443	7.54E-07	231
	10426	6.19E-03	2977	8603	2.78E-01	9884
	10494	6.93E-17	2	9034	7.08E-15	8
	11245	3.70E-03	2505	10234	1.30E-07	165
	11871	4.18E-09	98	11730	4.27E-02	5453
	12535	5.00E-07	212	11942	2.12E-10	45

Table A.18: Gene Selection Based on All Samples: Singh Data (b)

	GA			GAGA		
Subset Size	**Gene Index**	**Univ Pval**	**Univ Rank**	**Gene Index**	**Univ Pvbal**	**Univ Rank**
20	205	2.40E-13	13	1221	1.10E-06	253
	2597	1.33E-03	1828	2608	9.62E-02	6944
	2861	1.61E-01	8194	3094	6.22E-03	2983
	5255	1.37E-02	3875	3206	4.32E-01	11788
	5578	1.27E-05	471	3530	6.95E-03	3102
	5890	1.26E-10	37	3938	2.17E-03	2124
	5991	2.35E-05	546	5629	6.53E-04	1454
	7217	3.00E-01	10163	6390	1.78E-04	969
	7499	4.97E-01	12567	7297	5.54E-03	2866
	7800	2.96E-01	10110	7465	1.07E-06	251
	7900	4.40E-01	11904	8123	2.95E-12	21
	7903	3.21E-05	594	8297	2.89E-01	10037
	7956	4.39E-01	11877	8729	2.25E-04	1038
	8610	1.51E-02	3985	8814	1.30E-05	477
	9034	7.08E-15	8	9630	2.64E-02	4715
	9315	2.84E-01	9958	10234	1.30E-07	165
	9964	1.10E-01	7233	11331	1.04E-01	7127.5
	10417	4.50E-05	650	11858	1.44E-06	275
	10426	6.19E-03	2977	12378	1.02E-02	3503
	11661	3.73E-01	11068	12495	4.74E-08	136
25	205	2.40E-13	13	49	4.79E-01	12373
	1098	3.17E-01	10421	205	2.40E-13	13
	2182	1.24E-02	3741	1023	8.89E-03	3371
	2872	5.25E-03	2817	1355	3.05E-01	10240
	3794	3.49E-13	14	1534	1.22E-01	7496
	4161	4.24E-01	11705	1912	6.90E-03	3094
	4636	3.01E-05	585	2403	2.99E-02	4900
	4652	4.53E-01	12077	2752	2.66E-02	4723
	4725	1.06E-02	3547	5045	2.46E-09	85
	5838	9.94E-02	7011	5890	1.26E-10	37
	5890	1.26E-10	37	5915	1.82E-02	4216
	6468	9.06E-05	801	6185	3.67E-24	1
	7875	8.69E-07	234	6719	4.00E-02	5344
	8681	2.11E-02	4398	8295	4.87E-01	12455
	9133	6.59E-08	146	9034	7.08E-15	8
	9608	4.92E-01	12513	9263	2.12E-01	8993
	9850	1.74E-16	3	10000	1.45E-05	487
	9882	2.82E-02	4821	10234	1.30E-07	165
	10234	1.30E-07	165	10861	1.07E-01	7180
	10512	3.07E-01	10273	10979	8.98E-03	3379
	11026	2.67E-01	9737	11091	5.47E-02	5860
	11946	4.42E-01	11931	11858	1.44E-06	275
	12194	3.74E-04	1219	11871	4.18E-09	98
	12402	2.21E-03	2134	11918	1.24E-04	873
	12406	3.93E-01	11310	12478	5.40E-04	1362

Bibliography

[1] Affymetrix. Genechip analysis suite. User guide, version 3.3, Affymetrix, 1999.

[2] Affymetrix. Expression analysis technical manual. Technical report, Affymetrix, 2000.

[3] Affymetrix. Genecip expression analysis. Technical manual, Affymetrix, 2000.

[4] Affymetrix. Statistical algorithms description document. Technical report, Affymetrix, 2002.

[5] A Alizadeh, M Eisen, R Davis, C Ma, I Lossos, A Rosenwalkd, J Broldrick, H Sabet, T Tran, X Yu, JI Powell, L Yang, GE Marti, T Moore, J Hudson Jr., L Lu, DB Lewis, R Tibshirani, G Sherlock, WC Chan, TC Greiner, DD Weisenburger, JO Armitage, R Warnke, R Levy, W Wilson, MR Grever, JC Byrd, D Botstein, PO Brown, and LM Staudt. Different types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[6] U Alon, N Barkai, DA Notterman, K Gish, S Ybarra, D Mack, and AJ Levine. 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, 1999.

[7] C Ambroise and G McLachlan. 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, May 2002.

[8] K Baggerly, J Morris, J Wang, D Gold, L Xiao, and K Coombes. 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, 2003.

[9] Y Benjamini and D Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[10] CE Bonferroni. *Il calcolo delle assicurazioni su guppi di teste.* Rome, 1935. In 'Studi in Onore del Professore Salvatore Ortu Carboni'.

[11] CE Bonferroni. Teori statistica delle classi e calcolo delle probabilita. *Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[12] UM Braga-Neto and ER Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, 2004.

[13] N Cristianini and J Shawe-Taylor. *Support Vector Machines and other kernel-based learning methods.* Cambridge, Cambridge, England, 2000.

[14] S Dudoit and J Fridlyand. *Classification in Microarray Experiments*, chapter 3, pages 93–158. Chapman and Hall/CRC, 2003. Appearing in 'Statistical Analysis of Gene Expression Microarray Data' (ed. Terry Speed).

[15] S Dudoit, J Fridlyand, and T Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576, University of California, Berkeley, Dept. of Statistics, June 2000a.

[16] RA Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[17] A Freitas. *A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery*. Springer-Verlag, 2001. Appearing in 'Advances in Evolutionary Computing' (eds. A. Ghosh and S. Tsutsui).

[18] T Golub, D Slonim, P Tamayo, C Huard, M Gaasenbeek, J Mesirov, H Coller, M Loh, J Downing, M Caligiuri, C Bloomfield, and E Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[19] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[20] J Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, MI, 1975.

[21] R Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143, Montreal, Canada, 1995. in 'Proceedings of the 14th International Joint Conference on Artificial Intelligence' (IJCAI-95).

[22] R Kohavi and G John. *The Wrapper Approach*, chapter 1, pages 33–50. Kluwer Academic Publishers, 1998. Appearing in 'Feature Selection for Knowledge Discovery and Data Mining' (eds. H. Liu and H. Motoda).

[23] L Li, T Darden, C Weinberg, A Levine, and L Pedersen. 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, 2001.

[24] L Li, C Weinberg, T Darden, and L Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics*, 17(12):1131–1142, 2001.

[25] KV Mardia, JT Kent, and JM Bibby. *Multivariate Analysis*. Academic Press, San Diego, 1979.

[26] G McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[27] M Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, 1997.

[28] CL Nutt, DR Mani, RA Betensky, P Tamayo, JG Cairncross, C Ladd, U Pohl, C Hartmann, ME McLauhglin, TT Batchelor, PM Black, A von Deimling, S Pomeroy, TR Golub, and DN Louis. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, 63:1602–1607, 2003.

[29] CM Perou, SS Jeffrey, MVD Rijn, CA Rees, MB Eisen, RT Ross, A Pergamenschikov, CV Williams, SX Zhu, JC Lee, D Lashkari, D Shalon, PO Brown, and D Botstein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci.*, 96:9212–9217, 1999.

[30] SL Pomeroy, P Tamayo, M Gaasebeek, LM Sturla, M Angelo, ME McLaughlin, JYH Kim, LC Goumnerova, PM Black, C Lau, JC Allen, D Zagzag, JM Olson, T Curran, C Wetmore, JA Biegel, T Poggio, S Mukherjee, R Rifkin, A Califano, G Stolovitzky, DN Louis, JP Mesirov, ES Lander, and TR Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.

[31] P Pudil, J Novovicova, and J Kittler. Floating search methods in feature selection. *Patt. Recogn. Lett.*, 15:1119–1125, 1994.

[32] BD Ripley. *Pattern Recognition and Neural Networks*. Cambridge, Cambridge, England, 1996.

[33] MA Shipp, KN Ross, P Tamayo, AP Weng, JL Kutor, RCT Aguiar, M Gaasenbeer, M Angelo, M Reich, GS Pinkus, TS Ray, MA Koval, KW Last, A Norton, A Lister, J Mesirov, DS Neuberg, ES Lander, JC Aster, and TR Golub. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[34] D Singh, PG Febbo, K Ross, DG Jackson, J Manola, C Ladd, P Tamayo, AA Renshaw, AV D'Amico, JP Richie, ES Lander, M Loda, PW Kantoff, TR Golub, and WR Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

[35] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3.

[36] S Theodoridis. *Pattern Recognition*. Academic Press, San Diego, 1999.

[37] V Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1996.

[38] X West, C Blanchette, H Dressman, E Hunag, S Ishida, R Spang, H Zuzan, JA Olson Jr, JR Marks, and JR Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.*, 98:11462–11467, 2001.

[39] PH Westfall and SS Young. *Resampling-based multiple testing: Examples and Methods for p-value adjustment*. Wiley, New York, 1993.

[40] E Xing. *Feature Selection in Microarray Analysis*, chapter 6, pages 110–131. Kluwer Academic Publishers, 2002. Appearing in 'A Practical Approach to Microarray Data Analysis' (eds. D. Berrar, W. Dubitzky, and M. Granzow).

[41] M Xiong, L Wuju, J Zhao, L Jin, and E Boerwinkle. Feature (gene) selection in gene expression-based tumor classification. *Mol. Genet. Metab.*, 73:239–247, 2001.

[42] YH Yang, T Speed, S Dudoit, and P Luu. Normalization for cdna microarrya data. In ML Bittner et al., editors, *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proc. SPIE*, pages 141–152, May 2001.