JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Number 6, 2003

Pp. 961–980

Regression Approaches for Microarray Data Analysis

MARK R. SEGAL,[1] KAM D. DAHLQUIST,[2] and BRUCE R. CONKLIN[2]

ABSTRACT

A variety of new procedures have been devised to handle the two-sample comparison (e.g., tumor versus normal tissue) of gene expression values as measured with microarrays. Such new methods are required in part because of some defining characteristics of microarray-based studies: (i) the very large number of genes contributing expression measures which far exceeds the number of samples (observations) available and (ii) the fact that by virtue of pathway/network relationships, the gene expression measures tend to be highly correlated. These concerns are exacerbated in the regression setting, where the objective is to relate gene expression, simultaneously for multiple genes, to some external outcome or phenotype. Correspondingly, several methods have been recently proposed for addressing these issues. We briefly critique some of these methods prior to a detailed evaluation of *gene harvesting*. This reveals that gene harvesting, without additional constraints, can yield artifactual solutions. Results obtained employing such constraints motivate the use of regularized regression procedures such as the lasso, least angle regression, and support vector machines. Model selection and solution multiplicity issues are also discussed. The methods are evaluated using a microarray-based study of cardiomyopathy in transgenic mice.

Key words: cardiomyopathy, gene harvesting, least angle regression, microarray, support vector machine.

1. INTRODUCTION

MUCH HAS BEEN WRITTEN ON THE POTENTIAL USE of DNA microarrays in studying the relationship between phenotype and gene expression profiles on a whole-genome scale. Early attention was focused on categoric phenotypes, for example, differing cancer classes (Golub *et al.*, 1999) for which classification/discrimination methods were employed (Dudoit *et al.*, 2002). More recently, however, there has been investigation of continuous (Li and Hong, 2001) or survival (Hastie *et al.*, 2001a) phenotypes for which a regression framework is appropriate. The need to develop regression approaches for the microarray setting derives principally from the "large p, small n" problem (West *et al.*, 2001) whereby the number (p) of available and potentially interesting predictors (which we will loosely refer to as genes but are actually individual probe sets on the array that target full-length cDNAs or ESTs) vastly exceeds the number (n) of samples. An additional consideration is that, by virtue of pathway and gene network relationships, there will likely be strong and complex correlations between expression levels of various genes across samples.

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143-0560.
[2]Gladstone Institute of Cardiovascular Disease and Cardiovascular Research Institute, University of California, San Francisco, CA 94143.

We start by giving a very brief overview of some recent proposals for tackling these issues, highlighting shortcomings. Subsequently, we describe the dataset that will be used throughout to illustrate methods. This features microarray-based measures of gene expression and an attendant outcome used in a study of dilated cardiomyopathy in transgenic mice (Redfern *et al.*, 2000). We then proceed, in Section 2, to a detailed evaluation of a promising new technique, gene harvesting (Hastie *et al.*, 2001a). Again, some deficiencies are identified and improvements examined. These serve to motivate the use of the lasso (Tibshirani, 1996), least angle regression (Efron *et al.*, 2002), and support vector machines (Vapnik, 1998; Brown *et al.*, 2000), described in Section 3, as alternate regression tools for microarray studies. All these methods have "tuning parameters," the determination of which is crucial for model fit and interpretation. Such model selection issues are addressed in Section 4. Section 5 provides concluding discussion.

1.1. Some microarray regression approaches

As mentioned, the challenges of pursuing regression analyses with microarray data have spawned several new methodologic approaches. Here we provide a brief overview of a selection of these.

We note that separate consideration of continuous phenotypes and associated regression procedures is warranted even though some of the classification methods already employed for categoric (especially binary) phenotypes are generalizations of these procedures. There are important differences in how the bias-variance tradeoff operates for classification problems using 0-1 loss as compared with regression problems using squared error loss; see Hastie *et al.* (2001b). Nonetheless, some concerns (e.g., the cost of selection/adaptive procedures) will be common irrespective of loss as noted below.

West *et al.* (2001) develop a Bayesian regression framework customized to phenotype-gene expression association studies in the microarray context. They argue for allowing all genes to contribute to regression models, as opposed to applying prefiltering methods that yield small gene subsets and thereby mitigate the "large p, small n" problem. The cited difficulty with such strategies, as based on univariate (individual gene) association summaries, is that genes whose expression patterns jointly relate to phenotype may be eliminated. Accordingly, West *et al.* (2001) effect analyses by employing a singular value decomposition (SVD) of the full matrix of expression measurements and pursuing regression on the resultant latent factor variables. These latent variables (supergenes) provide for dimension reduction and summarize patterns of covariation among the original genes. Via the SVD, it is possible to map the standard linear regression formulation on the original genes to an equivalent regression on the latent factors. While the approach emphasizes careful, informative prior specification with attendant development of new classes of structured priors, there are some drawbacks to SVD-based regression that cannot be overcome by the Bayesian framework. These result from the fact that the latent factors are derived independently of the outcome or phenotype. So, in settings such as the study described below, where there are several different phenotypes associated with the disease, the same latent factors would be employed for each. Further, as with principal component regression, variation explained by the leading latent factors may not correlate with phenotypic variation (Hastie *et al.*, 2001b).

Li and Hong (2001) take a different approach to dimension reduction. In pursuing microarray regression they employ a Rasch model but with preliminary gene clustering. Since the clustering is performed independent of phenotype, the same concerns as above pertain: the same clusters will be used irrespective of phenotype, and within cluster variation may not correlate with phenotypic variation. Furthermore, results will be sensitive to the clustering algorithm and distance measure used, as we subsequently illustrate for the related gene harvesting procedure.

Zhang *et al.* (2001) use tree-structured (or recursive partitioning) techniques with gene expression data from a colon cancer study. While the application is classification (tumor versus normal tissue) rather than regression, important issues regarding degrees of freedom or effective numbers of parameters emerge that deserve further attention. Tree methods are highly adaptive and greedy: for each node of the tree, the best cut-point (expression level) of the best covariate (gene) is determined so as to optimize homogeneity of the resultant daughter nodes. In order to allay attendant concerns with overfitting, Breiman *et al.* (1984) employ cross-validation (CV) to pick appropriate tree size. Indeed, CV is used for this purpose in a multitude of settings. However, on account of the large p (6,500 reduced by filtering out low expression genes to 2,000), small n (62) setting, Zhang *et al.* (2001) use a very limited form of cross validation. Here, the tree topology is fixed—the number and identity of genes and the sequence in which they are

used is locked—and all that is subject to cross-validation is the expression level cut-point for a given gene. A point of reference is provided by the "generalized degrees of freedom" (gdf) construct of Ye (1998) which accounts for adaptivity. For regression tree procedures, Ye finds that the cost of a single split in $p = 10$ dimensional noise is ≈ 15 degrees of freedom. With the covariate dimensions encountered in microarray studies, it becomes evident that very few, if any, splits will withstand cross-validation. Indeed, this is the case for the colon data where, even for the reduced gene set ($p = 2000$), CV supports only one split. We expand on these concerns below, utilizing a straightforward way to compute effective numbers of parameters analogous to gdf (Section 2), as well as commenting on properties of cross-validation in the microarray setting (Section 4).

In view of these limitations with the above methods, we chose to further explore the new and promising gene harvesting technique, described in Section 2.

1.2. Cardiomyopathy data

The microarray data are from a transgenic mouse model of dilated cardiomyopathy (Redfern *et al.*, 2000). The mice overexpress a G protein-coupled receptor, designated Ro1, that is a mutated form of the human kappa opioid receptor, and that signals through the Gi pathway. Expression of Ro1 is controlled temporally and spatially through the use of an inducible expression system (Redfern *et al.*, 1999). When the receptor is overexpressed in the hearts of adult mice, the mice develop a lethal dilated cardiomyopathy that has many hallmarks of the human disease such as chamber dilation, left ventricular conduction delay, systolic dysfunction, and fibrosis. When expression of the receptor is turned off, the mice recover. The cardiomyopathy is due to hyperactive signaling of the receptor because treatment of the mice with a receptor antagonist or with pertussis toxin (which blocks Gi signaling) reverses certain phenotypes associated with the disease. To determine which changes in gene expression were due to the hyperactive signaling of Ro1 and led to cardiomyopathy in these mice, Affymetrix Mu6500 arrays were used. Labeled cRNA was isolated from the ventricles of thirty mice and hybridized one heart per set of arrays as described in Redfern *et al.* (2000). The thirty mice were divided into four groups. The control group was comprised of eight mice that were treated exactly the same as the eight-weeks experimental group except that they did not have the Ro1 transgene. A group of six transgenic mice expressed Ro1 for two weeks, which is approximately the amount of time required to reach maximal expression of Ro1 (Redfern *et al.*, 1999). These mice did not show symptoms of disease. A group of nine transgenic mice expressed Ro1 for eight weeks and exhibited cardiomyopathy symptoms. The recovery group of seven transgenic mice expressed Ro1 for eight weeks before expression was turned off for four weeks. In subsequent graphics, we label these groups as "C," "2," "8," and "R," respectively.

The Ro1 transgene is based on the human kappa opioid receptor. A probe set that targets the mouse kappa opioid receptor occurs on the Mu6500 array. This probe set cross-hybridizes to the Ro1 transgene and can be used as a measure of Ro1 expression, although the contribution of endogenous mouse kappa opioid receptor to the measured expression level cannot be ruled out. To determine which gene expression changes were due to the expression of the Ro1 transgene, we want to find genes that correlate (positively or negatively) with the Ro1 expression profile as displayed in Fig. 1. Genes that "explain" this expression profile are potential candidates to provide additional markers, therapeutic targets, and clues to the mechanism of disease.

Average difference values for gene expression were obtained using the Affymetrix GeneChip 3.1 software. As discussed in Section 4, there are numerous preprocessing steps and approaches to the extraction of expression summaries. The results that follow utilize standardized average differences (mean 0, variance 1) since such standardization is imposed for some of the methods considered subsequently (lasso, least angle regression).

2. GENE HARVESTING

Gene harvesting was developed by Hastie *et al.* (2001a) to explicitly tackle the challenges posed by regression in the microrarray context. The central strategy is to initially cluster genes via hierarchical clustering and then to consider the average expression profiles from all of the clusters in the resulting dendrogram as potential (an additional $p - 1$) covariates for the regression modeling. This modeling is effected by use of a forward stepwise algorithm with a prescribed number of terms. The number of terms

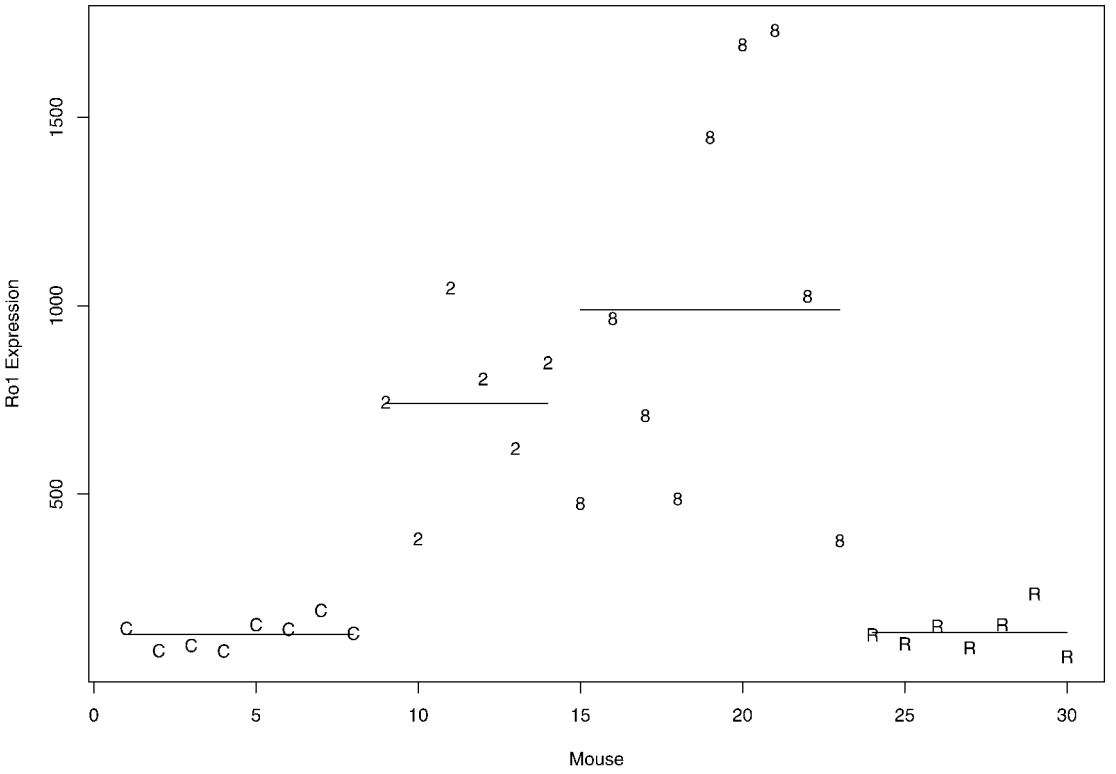

FIG. 1. Ro1 expression for the 30 mice. Symbols designate control ("C"), two week ("2"), eight week ("8"), and recovery ("R") groups. Horizontal lines are group expression averages.

actually retained is determined by cross-validation; this number constitutes the most important "tuning parameter" of the procedure. Provision is also made for between-gene interactions and nonlinear effects.

The authors claim two advantages for this approach. First, because of the familiarity of hierarchical clustering (e.g., Eisen *et al.* 1998) in *unsupervised* analyses of microarray expression data, the usage of clusters as covariates will be convenient for interpretation. Second, by using clusters as covariates, selection of correlated sets of genes is favored, which in turn potentially reduces overfitting. Implicit in this motivation is that regression procedures that yield lists of individual genes are deficient as there will "always be a story" linking an isolated gene to outcome. Ostensibly, credence is gained by finding groups of functionally related genes that are linked to outcome. However, as we demonstrate by way of application to the cardiomyopathy data, not only are these advantages not always realized, but harvesting can also give rise to artifactual results. We note that the abovementioned concerns regarding the use of derived (here cluster average) summaries not capturing outcome variation, and/or being fixed across differing outcomes, are mitigated by retention of the original genes as covariates in addition to the derived cluster average covariates.

Before presenting results, we give a brief overview of the gene harvesting algorithm. For the cardiomyopathy study, available data consists of the $n \times p$ matrix of gene expression values $X = [x_{ij}]$ where x_{ij} is the expression level of the j^{th} gene ($j = 1, \ldots, p = 6{,}319$) for the i^{th} mouse ($i = 1, \ldots, n = 30$). Each mouse also provides an outcome (Ro1) measure y_i. A hierarchical clustering algorithm is applied to the expression matrix and, for each of the resulting clusters c_k, $k = 1, \ldots, 2p-1$, the average expression profile $\bar{x}_{c_k} = (\bar{x}_{1,c_k}, \bar{x}_{2,c_k}, \ldots, \bar{x}_{n,c_k})$ where $\bar{x}_{i,c_k} = 1/|c_k| \sum_{j \in c_k} x_{ij}$ is obtained. Note that we have included the individual genes (the tips/leaves of the dendrogram) as clusters (of size 1) in this formulation—their average expression profile coinciding with the individual gene profile.

This set of $2p - 1$ average expression profiles constitutes the covariate set ($\mathcal{C}$). A forward stepwise regression is performed as follows. Initially, the only term in the model ($\mathcal{M}$) is the constant function 1; i.e., an intercept term. At each subsequent stage, candidates for inclusion consist of all products between a term in $\mathcal{M}$ and a term in $\mathcal{C}$. The term chosen for inclusion is that which most improves the fit as measured

here by the residual sum of squares (RSS, see below). The process continues until some prespecified maximum number of terms, m, have been added to the model. The number of terms retained is subsequently determined by cross-validation. Hastie *et al.* (2001a) restrict terms to second-order interaction terms; i.e., product terms are limited to pairwise products. This is partly motivated by interpretational considerations and borrows from the multivariate additive regression spline (MARS) methodology of Friedman (1991).

The gene harvesting model for continuous response is then

$$\hat{y}_i = \beta_0 \cdot 1 + \sum_{k \in S_1} \beta_k \bar{x}_{i,c_k} + \sum_{k,k' \in S_2} \beta_{k,k'} \bar{x}_{i,c_k} \bar{x}_{i,c_{k'}}. \tag{1}$$

Here, S_1 constitutes the set of clusters that enter singly while S_2 is the set of clusters that enter as product terms. So, $m = |S_1| + |S_2|$. The coefficients $\beta_k, \beta_{k,k'}$ are obtained by minimizing the residual sum of squares,

$$\mathrm{RSS}(\beta_k, \beta_{k,k'}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2}$$

Alternative loss functions to RSS are used for more general outcome types; for example, partial log-likelihood is used in conjunction with censored survival time outcomes. Further details on general outcome types and other aspects and applications of the gene harvesting algorithm are provided by Hastie *et al.* (2001a). Connections with the forward selection scheme of Keleş *et al.* (2002) are indicated in Section 4.

Table 1 provides results of applying gene harvesting to the cardiomyopathy data with $m = 6$. In Table 1(a), the hierarchical clustering was performed using average linkage (as used by Eisen *et al.* [1998] and often termed UPGMA), while in Table 1(b) single linkage was used. In both instances, the distance metric was Euclidean distance. We note that for single linkage, hierarchical clustering is invariant under monotone changes of the distance metric so that, for example, identical results would be obtained using correlation distance. While this property does not hold for average linkage, results using correlation distance were similar.

Immediately striking is the dramatic differences in gene harvesting results according to type of hierarchical clustering employed. This is compounded by further examination of the first, large (687 gene)

TABLE 1. RO1 GENE HARVESTING RESULTS[a]

Step	*Node*	*Parent*	*Score*	*Size*
1	6295	0	22.40	687
2	1380	6295	19.67	6
3	663	0	15.62	2
4	3374	663	10.69	3
5	1702	0	12.92	2
6	6268	663	11.27	83

(A)

Step	*Node*	*Parent*	*Score*	*Size*
1	g3655	0	21.97	1
2	2050	g3655	20.62	3
3	900	g3655	16.91	1
4	g1324	g3655	16.01	1
5	g1105	g3655	24.34	1
6	g230	g3655	12.44	1

(B)

[a](A) Average linkage. (B) Single linkage.

cluster selected under average linkage. None of the genes contained in this cluster are chosen under single linkage. Single linkage tends to select much smaller clusters, primarily singletons. Indeed, average linkage has arguably been too successful in selecting large clusters—it is problematic to characterize or infer relationships amongst a group of 687 genes! We note that these results were obtained without biasing the procedure to select large clusters as is advocated.

However, more consequential problems emerge when we pursue model selection. Figure 2 displays cross-validated and training residual variances for the average linkage results. Not only do the cross-validation results indicate that the best harvesting model (solid curve) only includes an intercept term (i.e., $m = 0$), but that this is far superior to all other models. Now, while it is the case that cross-validation is highly variable in this setting (as reflected by standard errors which are not shown for clarity; see Section 4), this is nonetheless a disturbing result. The face value interpretation is that none of the original 6,319 genes or the 6,318 gene clusters is *predictive* of Ro1. This conclusion is at odds with previous experiments, and analysis that shows that the cardiomyopathy phenotype is due to expression of the Ro1 transgene and that the expression of known markers of cardiomyopathy are up-regulated in the sick mice (Redfern *et al.*, 2000). We note that, in general, analogous null results may well be indicative of lack of signal in the data.

Figure 2 also displays cross-validated residual variances when the gene harvesting procedure is restricted to employing only the original genes (dotted curve for which corresponding standard errors are given). Now the results do withstand cross-validation to the extent that one term is retained under a "one standard error rule"; i.e., the model with one term has minimum residual variance and no smaller, competing model has residual variance within one standard error (as computed under the one term model) of this minimum value; see Breiman *et al.* (1984) for the basis of such rules. More importantly, cross-validated residual variances under the restricted approach are appreciably smaller than under the full gene harvesting procedure. We next examine the reasons for this poor performance of gene harvesting.

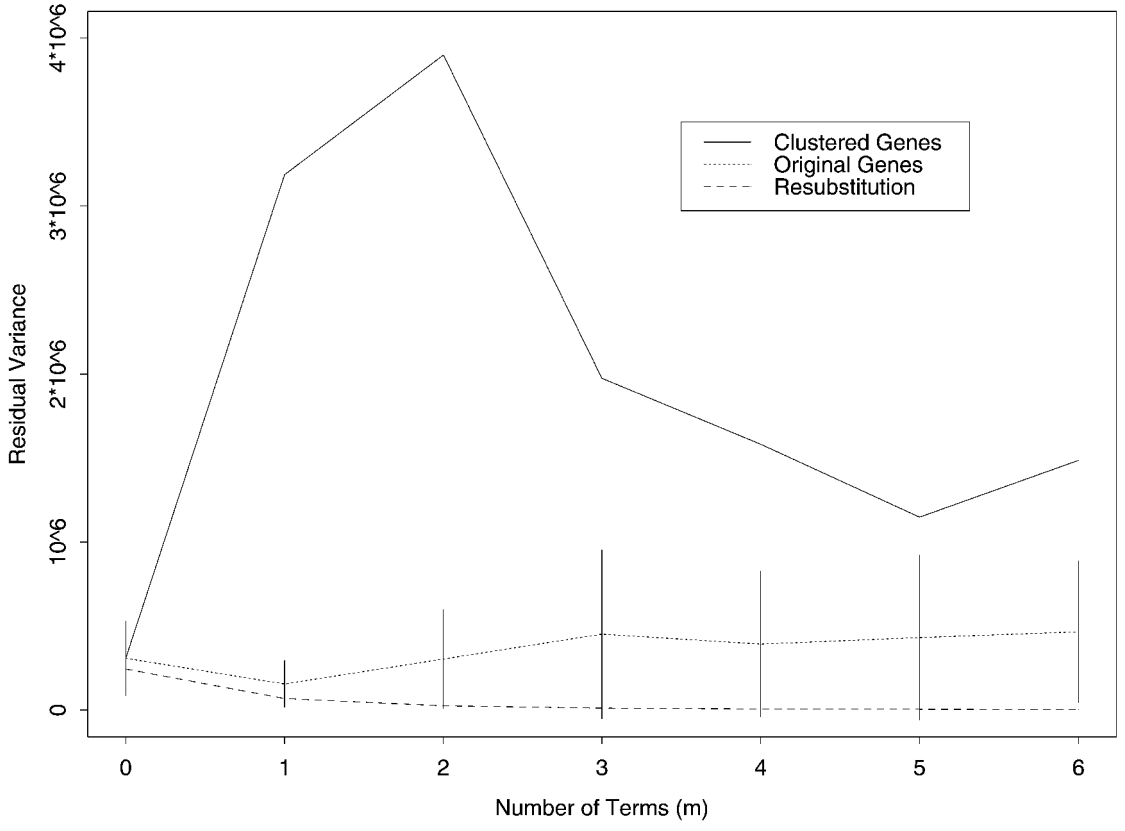

FIG. 2. Cross-validated residual variances for gene harvesting as a function of number of retained terms ($m = 0$ designates solely an intercept term): — full gene harvesting with clusters; $\cdots$ reduced gene harvesting with singletons; --- resubstition or training. The vertical bars are standard errors for the reduced harvesting approach.

A putative reason for the poorer performance of full gene harvesting is the expanded search space used in the forward stepwise selection that results from the addition of the 6,318 gene cluster average profiles as covariates. One way of assessing this is through assessments of model complexity. In Section 1.1, we referred to generalized degrees of freedom (Ye, 1998) that provide one such assessment. Here, we employ a related measure, effective number of parameters (*enp*, which we will also refer to as degrees of freedom) as derived from the covariance inflation criterion (CIC) (Tibshirani and Knight, 1999). Both measures are designed to capture the *cost* of adaptive (here the forward stepwise selection) methods. They differ primarily in whether simulation (Ye, 1998) or permutation (Tibshirani and Knight, 1999) is employed.

These costs are considerable. For the full gene harvesting procedure, the effective number of parameters for the inclusion of 1 through 5 terms are approximately 14, 18, 22, 25, and 27, respectively. It is immediately apparent that, for a sample size of $n = 30$, at most one or two terms is reasonable. Interestingly, similar *enp* values are obtained when we restrict ourselves to using only single rather than product terms, or using only individual genes rather than genes and clusters. These findings can be understood in light of Fig. 3, which concentrates on *enp* for selecting just one term. What is varied is the number of genes used in the harvesting approach. Filtering of genes was done in two ways, both blind to association with Ro1 outcome: genes were retained at random or genes were retained in order of their variation—the smaller gene sets contain the most variable genes. Here, results were invariant to retention scheme since we are applying harvesting with standardized expression values. What is notable from Fig. 3 is the slow rate of change in *enp* for large changes in proportion of genes retained above 30%. In reducing the complete data set ($p = 6{,}319$) to a 50% sample ($p = 3{,}160$), we gain only about 0.5 degrees of freedom, while reduction to a 30% sample ($p = 1{,}896$) buys about 2 degrees of freedom. It is this slow rate of change that accounts for the comparability of *enp* values using full harvesting or only individual genes for the entire dataset. The rate of change is rapid for small ($< 10\%$) proportions of genes retained, but the costs are still considerable relative to sample size. For example, selection of one term using a 1% sample ($p = 63$) costs 6.4 degrees of freedom.

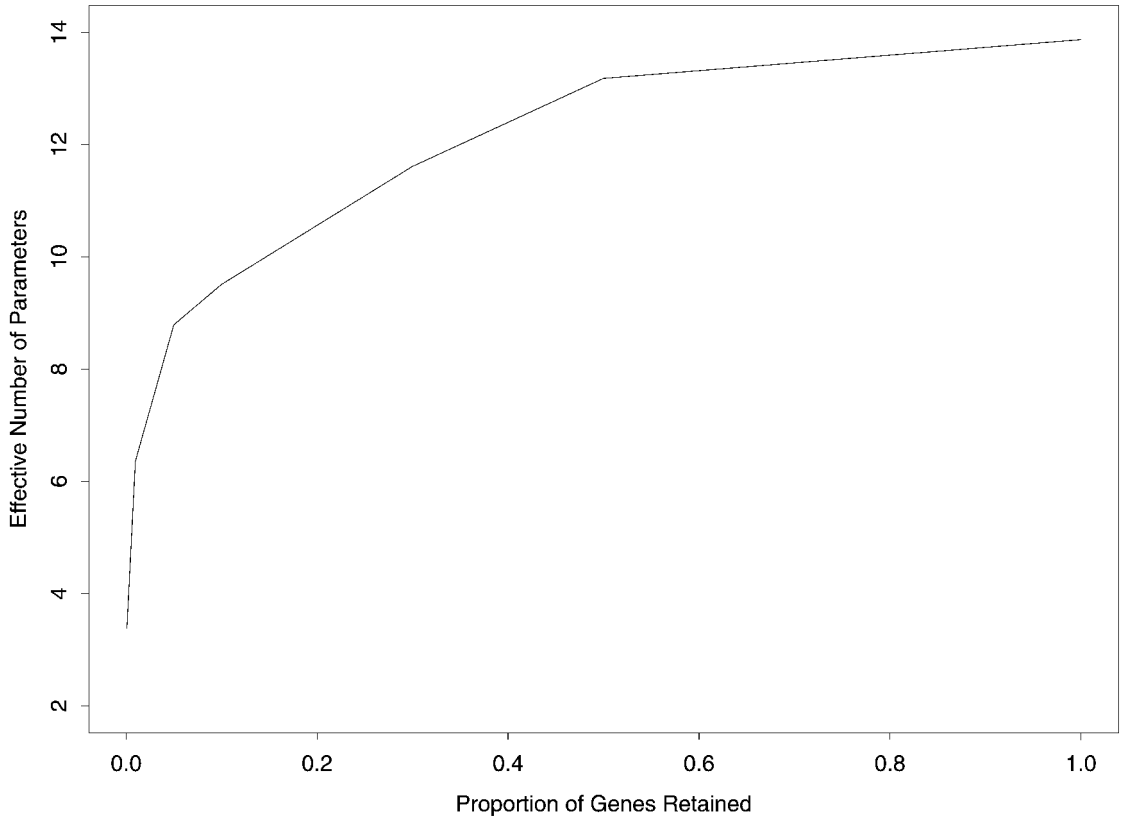

FIG. 3. Effective number of parameters for the first term of gene harvesting as a function of the proportion of genes retained.

While such determination of *enp* helps calibrate costs of adaptive procedures, and so can inform model size in the $n \gg p$ setting, the fact that *enp* values were similar for harvesting using clusters and singleton genes, whereas cross-validation displays substantial differences (Fig. 2) prompts further investigation. Additional scrutiny of the first term selected in the full harvesting procedure—the 687 gene cluster—is revealing. The heat map for this cluster is presented in Fig. 4. A seemingly coherent collection of expression profiles, characterized by reduced values for the mice in the eight-week group, constitutes the cluster. However, if we examine the actual correlations between the 687 genes in the cluster and the average expression profile for the cluster, the coherence is not so impressive. Figure 5a is a histogram of these 687 correlations. We note that 28 (4%) of the correlations are negative, and more than 50% are less than 0.5. An alternate view of cluster coherence can be obtained by examining the scores (essentially squared *t*-statistics) of the 687 genes when they are individually regressed against Ro1. The results are presented in Fig. 5b. The number of genes displaying no association with Ro1 is striking: 33% have *t*-statistics < 1 and approximately 75% have *t*-statistics < 2. Even the maximal individual squared *t*-statistic (13.94) is far removed from the score for the average expression profile (22.4).

What has occurred is the following. The hierarchical clustering procedure has yielded a sizable cluster whose *average* expression profile *happens* to be strongly associated with Ro1. This occurs despite the bulk of the cluster members (genes) exhibiting little or no association with Ro1. In view of this artifact, it is not surprising that no terms are selected on cross-validating and that the cross-validated residual variances are so large.

It is possible to constrain the harvesting procedure to mitigate against this behavior. In particular, by allowing only clusters meeting coherence criteria to be selected, these artifacts are avoided. As illustrated above, coherence can be captured by individual member genes being sufficiently correlated with the cluster average profile and/or the individual gene squared *t*-statistics being sufficiently close to the squared *t*-statistic for the cluster average profile.

Applying such a constrained harvesting algorithm with a correlation threshold of 0.3 (i.e., only clusters for which each individual member gene had a correlation of ≥ 0.3 with the cluster average profile were eligible for selection) produced the following interesting results. The term chosen first is an eight-gene cluster, itemized in Table 2 and depicted via a heat map in Fig. 6. This was the only term to be retained under cross-validation. The striking feature of the heat map is the appreciable down-regulation (red) of all genes for the nine mice in the eight week (induced cardiomyopathy) group when Ro1 expression is elevated. The constituent genes admit the following interpretation.

Lipoprotein lipase, ATP synthase gamma chain, and ATP synthase coupling factor 6 encode proteins involved in energy production for the cell. Lipoprotein lipase is the enzyme that cleaves fatty acids from triacylglycerol so that they can be further utilized in the fatty acid degradation pathway, a major source of energy in the cardiomyocyte. ATP synthase gamma chain and ATP synthase coupling factor 6 are subunits of the ATP synthase complex of the electron transport chain in the mitochondria. Myoglobin stores and delivers oxygen in muscle which is needed to generate ATP in the mitochondria. The down-regulation of delta-aminolevulinate dehydratase is potentially related to the down-regulation of the ATP synthase complex genes above because it catalyzes the second step in the biosynthesis of heme, a cofactor required by several proteins in the electron transport chain. Elongation factor 1 alpha 2 (Eef1a2) is a translation factor required for protein synthesis. The down-regulation of Eef1a2 is also consistent with the down-regulation of genes involved in energy production since protein synthesis is one of the most energy intensive processes in the cell. That is, if energy production is decreased, down-regulating protein synthesis is a typical and effective cell response to conserve energy. The functions of Skd3 and translationally controlled tumor protein are unknown. In summary, several of the genes in this cluster are consistent with the down-regulation of energy production during the induced cardiomyopathy.

By construction, constrained harvesting will alleviate the problem of artifacts associated with the original gene harvesting approach. But the question of how to specify correlation and/or score-based constraints remains open. We prefer to rely or prior prescription since devising appropriate estimation criteria for these parameters appears problematic. This concern is made moot by the following observation: the tendency under constrained harvesting is to select singleton genes. This was evident for a wide range of constraint thresholds, providing the correlation between the cluster average expression profile and individual genes in the cluster was 0.5. If, under such (appropriate) restriction, harvesting is going to be reduced to selecting singleton genes, then it becomes pertinent to consider alternate gene selection schemes in view of the recognized limitations of forward selection strategies. Accordingly, we next examine the utility of lasso, least angle regression, and support vector machines for regression in microarray gene expression settings.

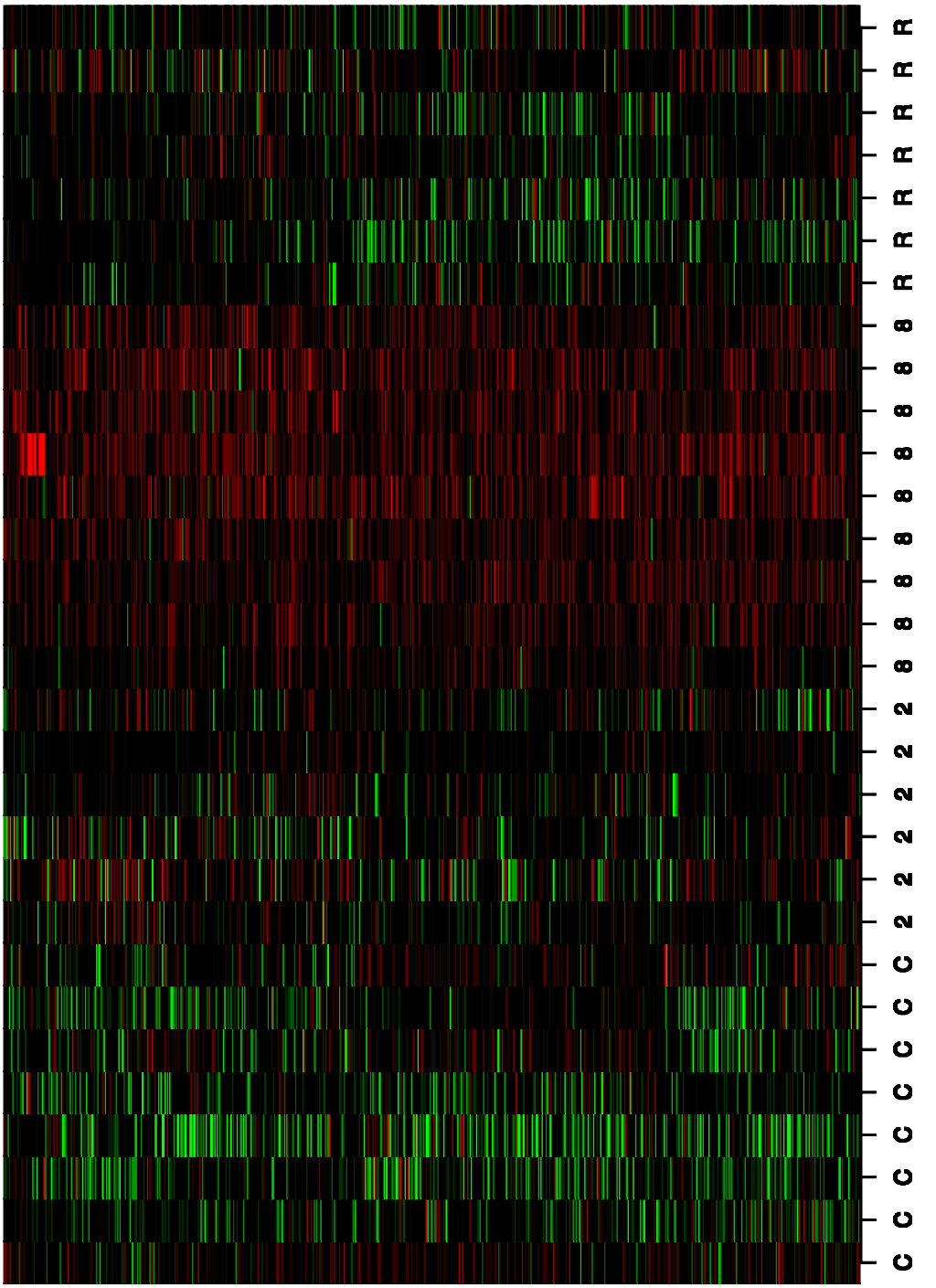

Figure 4. Heat map for the 687 genes (rows) comprising the cluster of the first term (node 6295) selected by gene harvesting. Group symbols (columns) are detailed in Fig. 1.

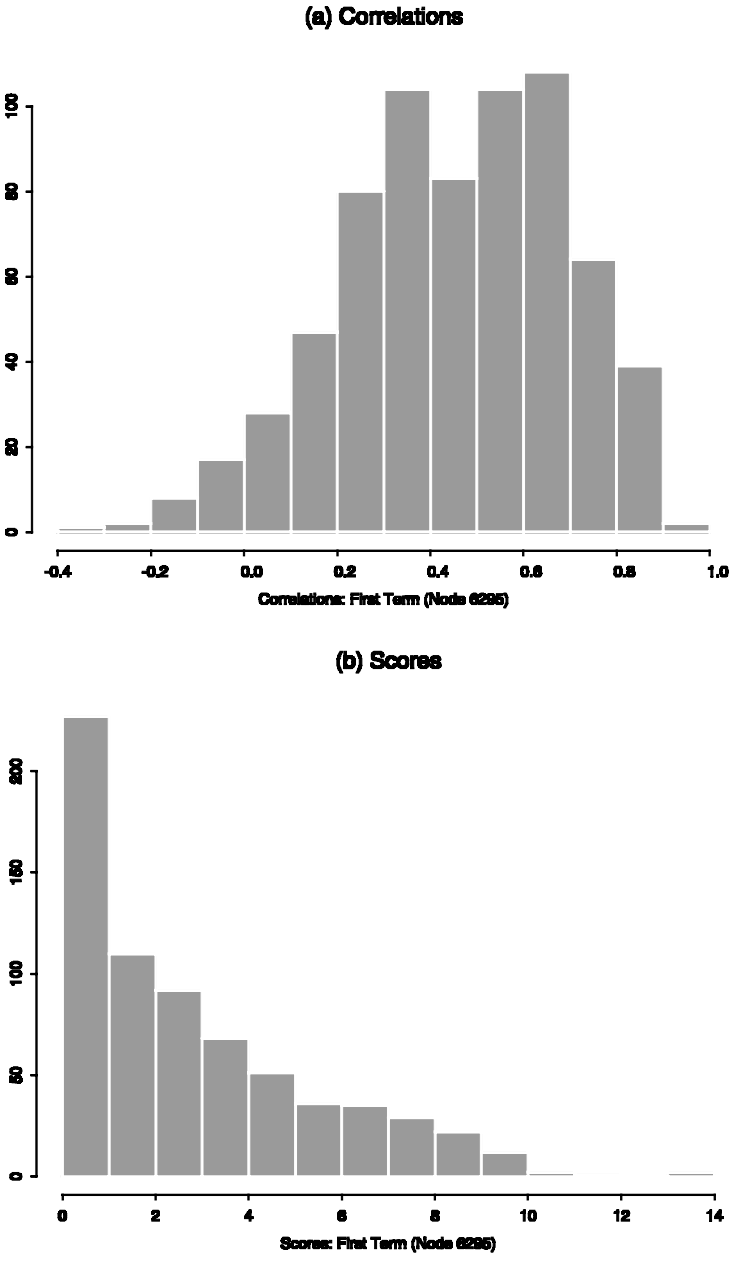

FIG. 5. Correlations (**a**) and scores (**b**) for the 687 genes constituting the first term (node 6295) selected by gene harvesting.

TABLE 2. CONSTRAINED HARVESTING SELECTED CLUSTER

Mu6500 probe set	*GenBank*	*Symbol*	*Description*
Msa.909.0	M60847	Lpl	Lipoprotein lipase
Msa.33808.0	AA114811	—	EST homologous to ATP synthase gamma chain
Msa.2424.0	X13752	Alad	Delta-aminolevulinate dehydratase
Msa.2412.0	X06407	Tpt1	Translationally-controlled tumor protein 1
Msa.22491.0	AA036584	—	EST homologous to ATP synthase coupling factor 6
Msa.2037.0	X04405	Mb	Myoglobin
Msa.1923.0	L26479	Eef1a2	Eukaryotic translation elongation factor 1 alpha 2
Msa.1435.0	U09874	Skd3	Suppressor of K+ transport defect 3

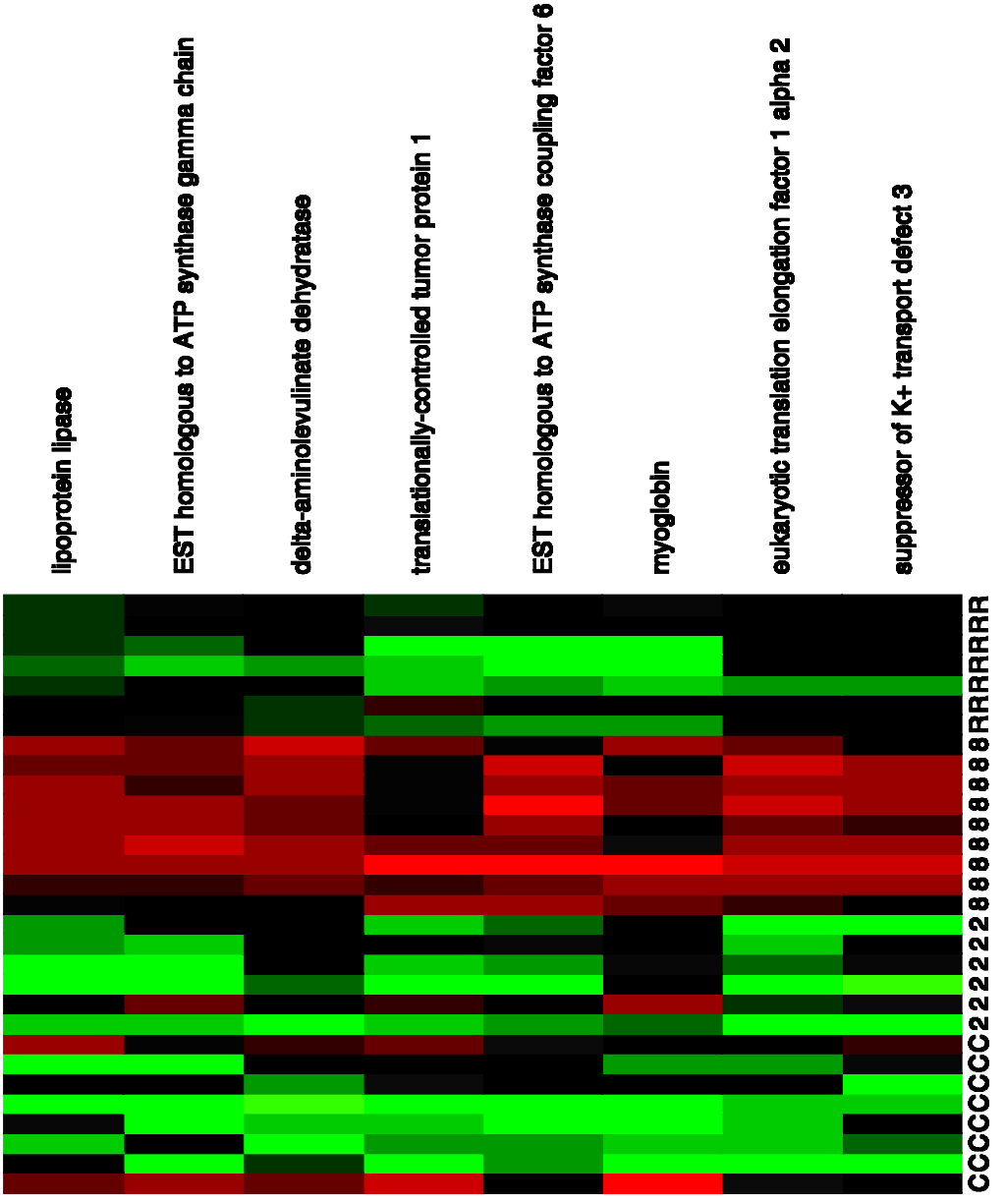

Figure 6. Heat map for the 8 genes (rows) comprising the cluster of the first term selected by constrained gene harvesting. Group symbols (columns) are detailed in Fig. 1.

3. REGULARIZED REGRESSION APPROACHES

As illustrated in the context of gene harvesting, the combination of $p \gg n$ and adaptive regression procedures does not mix well. While the flexibility of adaptive procedures is necessary to enable gene selection, additional constraints are needed to overcome costs/variability inherent in such approaches. Here we consider some regression methods that impose constraints by way of penalties/regularization. Indeed, even for classification approaches to microarray data, such regularization is often applied, albeit implicitly (e.g., Dudoit *et al.*, 2002).

3.1. Lasso

The lasso (least absolute shrinkage and selector operator) was proposed by Tibshirani (1996). The lasso combines the good features of ridge regression and subset regression procedures, which in turn were developed to overcome deficiencies with ordinary (OLS) least squares regression estimates. There are two primary shortcomings ascribed to OLS. Firstly, *prediction accuracy* is affected by the fact that OLS estimates, while enjoying low bias, frequently have large variance. Prediction accuracy can often be improved by shrinking or zeroing select coefficients. Secondly, *interpretation* is complicated by retention of large numbers of covariates. It is generally preferable to isolate a smaller subset of covariates that have the strongest effects. However, it is important in the microarray context to remain mindful of the fact that there will likely be many alternative such subsets having comparable prediction accuracies in view of the anticipated between-gene correlations.

Ridge regression (Hoerl and Kennard, 1970) achieves improved prediction accuracy via shrinkage. For simplicity, consider centered data (so we can ignore the intercept term) and the usual linear predictor $\mu = \mathbf{X}\beta = (\sum_{j=1}^{p} \beta_j x_{ij})$. Instead of minimizing just the usual residual sum of squares as per OLS, $\mathrm{RSS}(\beta) = ||y - \mu||^2 = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2$, ridge regression achieves coefficient shrinkage by constraining their size:

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \quad \text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t. \tag{3}$$

An equivalent formulation is afforded by L_2 penalized regression:

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{4}$$

there being a one-to-one correspondence between t in (3) and λ in (4). We note here that ridge regression coincides with one version of support vector machine regression, considered in section 3.3.

The difficulty, acute in the array setting, with ridge regression is that all coefficients are retained. Tibshirani (1996) demonstrates how replacing the L_2 penalty in (4) with an L_1 penalty

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t \tag{5}$$

results in some of the coefficients being exactly zero. The resultant estimates define the lasso estimates. Again, there is an equivalent penalized version:

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{6}$$

By varying t in (5), we obtain a continuous form of subset regression. This overcomes the inherent variability in subset regression due to its discreteness. Such discreteness arises since covariates are either retained or discarded. It is recommended that t be determined by cross-validation. Thus, the lasso seeks to

simultaneously capture the good properties of ridge and subset regression. Hastie *et al.* (2001b) contains extensive discussion.

For the microarray setting, implementation issues are forefront. The original algorithm proposed by Tibshirani (1996) does not handle the $p > n$ case and is consequently inapplicable. This limitation, along with efficiency concerns, motivated Osborne *et al.* (2000) to regard the lasso as a convex programming problem and to devise an algorithm based on homotopy methods. While the objectives of handling $p > n$ and improving efficiency were realized, the algorithm, at least as implemented in Splus (available from *lib.stat.cmu.edu/S/lasso2*), remains problematic for microarray studies. When applied to the Ro1, cardiomyopathy dataset run times on a Sun Microsystems E420R server with four 450MHz UltraSPARC-II processors and 4GB memory, user time (as provided by *unix.time()*) for a sequence of 30 bounds (t values) was 47 minutes. However, getting to this run required considerable trial and error to determine an appropriate range of bounds since specification of bounds that are too large produces errors. Furthermore, attempts to pursue model selection (picking a specific t or λ) based on cross-validation failed due to insufficient memory.

3.2. Least angle regression

The development of least angle regression (LARS) (Efron *et al.*, 2002), which can readily be specialized to provide all lasso solutions in a highly efficient fashion, represents a major breakthrough. LARS is a less greedy version of standard forward selection schemes. The simple yet elegant manner in which LARS can be adapted to yield lasso estimates as well as detailed description of properties of procedures, degrees of freedom, and attendant algorithms are provided by Efron *et al.* (2002). Code can be obtained from *www-stat.stanford.edu/~hastie/Papers.*

Results from applying LARS and the LARS version of the lasso to the Ro1 study are described below. Since these coincide through 18 steps, presentation is for LARS only. A plot of regression coefficient profiles is given in Fig. 7. Using the built-in cross-validation function and applying a "1-SE" rule suggests that five terms be retained. The corresponding genes are given in Table 3.

Each of these genes must be interpreted individually because they do not constitute a "cluster" as per clusters extracted by gene harvesting. Each of these genes is up-regulated in response to Ro1 induced cardiomyopathy. Ribophorin II is a subunit of the oligosaccharyltransferase complex in the endoplasmic reticulum that glycosylates proteins in the secretory pathway. Heat shock 70 kD protein 8 is a chaperone involved in protein folding in the cytoplasm. CD98 heavy chain is part of a heterodimer that makes up the L-type amino acid transporter in the plasma membrane. The Lon protease homolog is a mitochondrial enzyme that may be important for the folding and degradation of proteins in the mitochondrion. None of these four genes has been previously implicated in cardiomyopathy. However, the final gene in this list, fibronectin 1, is a structural component of the extracellular matrix that is part of the fibrotic response to cardiomyopathy in humans and the Ro1-expressing mice (Redfern et al., 2000).

According to the prescription given in Efron *et al.* (2002), this costs roughly five degrees of freedom: degrees of freedom $\approx$ number of terms (steps). However, care is needed in making comparisons with harvesting degrees of freedom for two reasons. First, both the empiric and theoretic setting for the LARS prescription had $n > p$. Second, the CIC determination requires an (external) estimate of $\hat{\sigma}^2$. So we applied CIC to LARS with the same $\hat{\sigma}^2$—and the prescription continued to hold.

Finally, we note the dramatic computational gains over the Osborne *et al.* (2000) implementation of the lasso. The user time for the same number of bounds (30) was 0.5 minutes for LARS and 0.7 minutes for the LARS implementation of the lasso. These represent appreciable improvements on the above-mentioned 47 minutes. Additionally, there were no memory issues in performing cross-validation based model selection.

3.3. Support vector machines

Support vector machines (SVMs) have been used for classification purposes in the microarray setting (Brown *et al.*, 2000). Regression modalities for SVMs are described in Cristianini and Shawe-Taylor (2000) and briefly overviewed here.

Given a set of basis functions $\{\phi_m\}_1^M$ (obtained via a kernel as described below) and a corresponding regression function (linear predictor) $f(\mathbf{x}_i) = \sum_{m=1}^{M} \beta_m \phi_m(\mathbf{x}_i) + \beta_0$ where $\mathbf{x}_i \in R^p$ is the expression

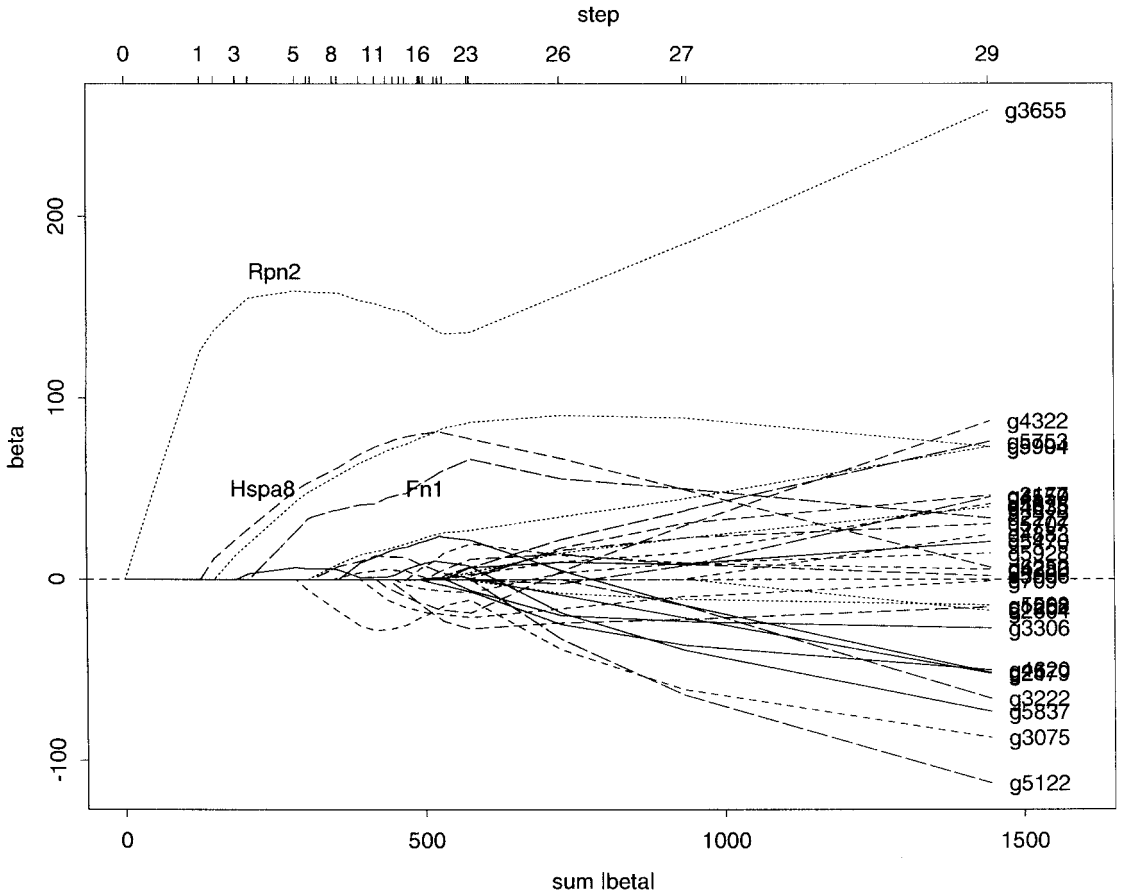

FIG. 7. Coefficient profiles for the 30 bounds (29 steps) of the LARS algorithm. Profiles of known genes for the 5-term model as chosen by cross-validation (see Table 3) are identified.

TABLE 3. LARS/LASSO SELECTED GENES

Mu6500 probe set	*GenBank*	*Symbol*	*Description*
Msa.2877.0	D31717	Rpn2	Ribophorin II
Msa.778.0_i	U73744	Hspa8	Heat shock 70kD protein 8
Msa.2134.0	U25708	—	CD98 heavy chain
Msa.26025.0	AA061310	—	EST homologous to lon protease homolog, mitochondrial
Msa.657.0	M18194	Fn1	Fibronectin 1

vector for the i^{th} mouse, SVM obtains coefficient (β) estimates via

$$\min_{\beta} \sum_{i=1}^{n} L^{\varepsilon}(y_i - f(\mathbf{x}_i)) + \frac{\lambda}{2} \| \beta \|^2 . \tag{7}$$

Here, L^{ε} designates ε insensitive loss whereby we ignore errors of absolute size less than ε. Thus, for example, $L_1^{\varepsilon}(y_i - f(\mathbf{x}_i)) = \max(0, |y_i - f(\mathbf{x}_i)| - \varepsilon)$. Since, as detailed below, we will take $\varepsilon = 0$, and use of L_2^0 loss coincides with ridge regression, we restrict attention to L_1 loss. For such a loss function, the equivalent primal optimization problem, following the introduction of slack variables ξ_i, ξ_i^*, is

$$\min_{\beta,\xi,\xi_*} \sum_{i=1}^{n} (\xi_i + \xi_i^*) + \frac{\lambda}{2} \| \beta \|^2 \quad \text{subject to} \quad y_i - f(\mathbf{x}_i) \le \varepsilon + \xi_i; \quad f(\mathbf{x}_i) - y_i \le \varepsilon + \xi_i^*; \quad \xi_i, \xi_i^* \ge 0.$$

The corresponding dual problem is readily solved:

$$\min_{\alpha,\alpha^*} \frac{1}{2}(\alpha-\alpha^*)^T Q(\alpha-\alpha^*) + \varepsilon \sum_{i=1}^{n}(\alpha_i+\alpha_i^*) - \sum_{i=1}^{n} y_i(\alpha_i-\alpha_i^*) \tag{8}$$

$$\text{subject to} \qquad \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0; \qquad 0 \le \alpha_i, \alpha_i^* \le 1/\lambda.$$

Here, $Q_{ij} = \sum_{m=1}^{M} \phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j) = \langle \phi_m(\mathbf{x}_i), \phi_m(\mathbf{x}_j) \rangle \equiv K(\mathbf{x}_i, \mathbf{x}_j)$. The solution is

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n}(\hat{\alpha}_i^* - \hat{\alpha}_i) K(\mathbf{x}_i, \mathbf{x}) + \beta_0. \tag{9}$$

The fact that (8) and (9) involve $\phi(\mathbf{x})$ only through inner products as given by the *kernel*, K, confers huge computational benefit. This is because all that needs to be stipulated is the kernel; the individual basis functions $\phi()$ are not required. Accordingly, it is possible to simply yet greatly enrich the underlying basis as illustrated by popular kernels including d^{th} *degree polynomial*: $K(x, y) = (1 + \langle x, y \rangle)^d$; *radial basis*: $K(x, y) = \exp(-||x - y||^2/c)$.

However, the added flexibility afforded by such basis expansion is typically going to be of limited utility in the microarray setting since already we have $p \gg n$. That is, while it may be conceptually appealing to include select d^{th} order between-gene interactions via a polynomial kernel, the gains from fitting *all* such terms with small n and interpretational objectives are unclear. For classification problems in the microarray setting, there have been corresponding calls for feature (basis) selection in using SVMs (Furey *et al.*, 2000; Guyon *et al.*, 2002; Lee and Lee, 2002). Of course, feature selection is central to gene harvesting and lars/lasso.

Use of $\varepsilon > 0$ results in only a subset of $\hat{\alpha}_i^* - \hat{\alpha}_i$ being nonzero. The associated i^{th} data point is termed a *support vector*. Again, for classification problems, numerous examples demonstrate the advantages of obtaining sparse solutions wherein only data points close to the decision boundary (the support vectors) are used to define the boundary. However, when $p \gg n$ and for regression problems, such sparsity in n is not desirable.

Accordingly, our application of SVMs to the cardiomyopathy data focuses on $\varepsilon = 0$ and emphasizes linear kernels. We did investigate using quadratic kernels, but even with an extensive grid search for λ, no models withstood cross-validation. Similarly, Guyon *et al.* (2002) restrict themselves to linear kernels for microarray classification. For a linear kernel ($\phi(\mathbf{x}) = \mathbf{x}$), we recover gene specific coefficients via

$$\hat{\beta} = \sum_{i=1}^{n}(\hat{\alpha}_i^* - \hat{\alpha}_i)\mathbf{x}_i. \tag{10}$$

Again, λ was determined using grid search combined with CV. Examination of the $\hat{\beta}$ distribution from (10) reveals outlying/extreme genes as presented in Table 4.

TABLE 4. SVM SELECTED GENES

Mu6500 probe set	*GenBank*	*Symbol*	*Description*
Msa.799.0	V00756	Ifrd1	Interferon-related developmental regulator 1
Msa.778.0_i	U73744	Hspa8	Heat shock 70kD protein 8
Msa.2972.0	U49350	Ctps	Cytidine 5′-triphosphate synthase
Msa.2134.0	U25708	—	CD98 heavy chain
Msa.2138.0	X15830	Sgne1	Secretory granule neuroendocrine protein 1, 7B 2 protein
Msa.3227.0	U36788	Hccs	Holocytochrome c synthetase
Msa.433.0	X69063	Ank1	Ankyrin 1, erythroid
Msa.2877.0	D31717	Rpn2	Rbophorin II

Given a total of $p = 6{,}319$, there is clearly considerable overlap with lars/lasso selections; three genes are common to both lists. The functions of the five new genes found with the SVM approach are described below, again keeping in mind that the genes are not a "cluster" and must be interpreted individually. Interferon-related developmental regulator 1 (Ifrd1) may be involved in myoblast differentiation (Guardavaccaro *et al.*, 1995) and up-regulated in an inflammatory response due to ischemia-reperfusion injury from cardiopulmonary bypass in a neonatal lamb model (Nelson *et al.*, 2002). Its up-regulation in the Ro1 expressing mice could indicate a common response pathway for ischemia-reperfusion injury and cardiomyopathy. Cytidine 5′-triphosphate synthase catalyzes the final step in the production of the nucleotide cytidine triphosphate (CTP) and is also involved in phosphadidyl-choline metabolism (Kent and Carman, 1999). Secretory granule neuroendocrine protein 1, 7B2 is involved in regulating pituitary hormone secretion and has been previously shown to be expressed only in neuroendocrine cells (Westphal *et al.*, 1999). Holocytochrome c synthetase links heme to cytochrome c, a protein involved in the electron transport chain. Its down-regulation is consistent with the down-regulation of delta-aminolevulinate dehydratase found in the gene harvesting cluster. Ankyrin 1 is a structural protein involved in anchoring the cytoskeleton to the plasma membrane. Its down-regulation is potentially related to gene expression changes in other cytoskeletal components seen in the Ro1-expressing mice (Redfern *et al.*, 2000). None of these genes has been previously implicated in cardiomyopathy, although Ifrd1, holocytochrome c synthetase, and ankyrin 1 are the most likely of the group to be related to the phenotype of the Ro1 mice based on their previously described functions.

4. MODEL SELECTION ISSUES

The problem of variable selection in the context of microrarray regression is of crucial importance—identification of gene expression changes associated with phenotypes of interest being a primary objective of microarray studies. However, the distinguishing characteristics of such studies ($p \gg n$, correlated gene expression) makes such selection inherently difficult. Here we discuss the two principal means for effecting gene (variable) selection, criterion based and prediction error based, from the microrarray regression perspective. Throughout we continue to assume squared error (L_2) loss. The question of multiple solutions (variable sets) is also addressed.

A variety of model selection criteria exists, including Akaike Information Criterion (AIC) (Akaike, 1973) which is equivalent to Mallows (1973) C_p under the usual Gaussian model, Bayesian Information Criterion (BIC) (Schwartz, 1979), and the Covariance Information (CIC) (Tibshirani and Knight, 1999). Common to all approaches is (i) penalization of resubstitution or training error estimates, and (ii) the need to estimate σ^2, the residual (error) variance. The penalization in (i) seeks to compensate for training error optimism, so as to recover unbiased estimates of prediction error. Despite differing derivations, the approaches primarily differ in the degree of penalization. To the extent that the criteria involve/allow estimation of "degrees-of-freedom" analogs, especially for greedy/adaptive procedures, this provides useful additional information. In particular, as illustrated in the context of gene harvesting, degree-of-freedom estimates were helpful in judging appropriate model size in the face of small n, and similarly for tree-structured methods (see Section 1.1).

However, estimating σ^2 is problematic. The frequent recommendation to base estimates on a full model (e.g., Tibshirani and Knight, 1999) will yield $\hat{\sigma}^2 = 0$ when $p > n$. This eliminates the penalty term in the above criteria, reducing them to (useless) resubstitution error measures. General strategies for specifying "nearly full" models in order to alleviate this difficulty are elusive. One possibility, specific to LARS/lasso, is to employ the largest model such that all coefficient profiles are monotone. The logic here is that departures from monotonicity result from between-variable correlation, which we seek to avoid in parsimonious model descriptions. The retained genes can be viewed as representing distinct pathways that are jointly predictive of outcome. Here, for example, this results (Fig. 7) in a full model with five genes. Using the corresponding estimate of σ^2 in conjunction with the AIC or CIC criteria results in choosing this full five-gene model and so coincides with the CV selection of Section 3.2. This strategy is clearly an ad hoc and nongeneralizable prescription. Furthermore, such prescriptions are consequential in that it is absolute (rather than relative) values of the respective criteria that are used to determine model size. Therefore, direct measures of prediction error, such as provided by cross-validation, are preferred.

The merits of basing model selection on prediction error determinations have been recently and convincingly advanced (e.g., Breiman, 2001a). But, CV can also be problematic in the $p \gg n$ setting, especially when n is small. The difficulties largely pertain to the variability of CV estimates of prediction error. These were showcased for leave-one-out (LOO) CV by Kim *et al.* (2002). In order to circumvent this variability concern (and secondarily to reduce computation), K-fold CV is advocated (Hastie *et al.*, 2001b). Here the data is partitioned into K roughly equal-sized samples, model building utilizes $K - 1$ of these, validation (i.e., computation of prediction error) the remaining (withheld) sample with cycling and aggregation over all (K) such possibilities. However, with $n = 30$ as in the Ro1 dataset, the popular choice of 10-fold CV amounts to leave-three-out, and prediction error estimates remain highly variable. Use of smaller K and or test/split sample approaches is limiting with respect to model building given large p, and all the more so for adaptive methods.

One promising refinement pertinent to gene harvesting, but more widely applicable, is the forward selection with Monte Carlo CV (FSCV) approach proposed by Keleş *et al.* (2002). Their regression model resembles the gene harvesting regression scheme (1) applied to individual genes (rather than clusters) with one other important distinction. Rather than basing model selection (number of terms) on CV applied to an a priori determined series of nested models, by minimizing average prediction error on the validation samples as is done for gene harvesting (as per typical cross-validation praxis), FSCV embeds cross-validation into the selection procedure. This is accomplished as follows. Data is partitioned into test and training sets. Using the training data, coefficients ($\hat{\beta}_j$) are obtained for all genes ($j = 1, \ldots, p$) by minimizing RSS (2). However, unlike standard forward selection or gene harvesting, the gene selected for entry is not that achieving minimal RSS on the training data. Instead, RSS is evaluated using the test data and the corresponding best gene (that achieves minimal RSS) is included. To accommodate variation introduced by sample splitting, the entire procedure is repeated K times and results synthesized. While FSCV provides test sample validation on a per-step basis, it clearly does not overcome the variability of CV error estimates.

The distinguishing attributes of microarray data ($p \gg n$, correlated expression) make the existence of so-called "Rashomon effects" (Breiman, 2001a)—many competing distinct models with comparably good performance—a foregone conclusion. Indeed, gene harvesting was partially motivated from this viewpoint: rather than eliciting multiple models, each containing an instance from a set of correlated genes, perform a priori clustering of genes so that such a set emerges from a single run.

There are several approaches to extracting multiple solutions. These include perturbing data, modifying criteria/algorithms, and extending obtained solutions. Data perturbation can be pursued in two distinct ways: the raw input values themselves can be modified and/or some operator (e.g., filtering, resampling) can be applied to a given set of inputs. With microarray expression data, there are a multitude of specifications and approaches that determine actual input data values, even after completion of the experiment. For spotted two-color arrays, background correction (e.g., Kooperberg *et al.*, 2002), normalization, and "unfolding" (Goryachev *et al.*, 2001) can be applied. For Affymetrix arrays, several algorithms exist for deriving expression values including Affymetrix GeneChip 3.1–4.0 software (used here), Affymetrix MAS 5.0, and dChip (Li and Wong, 2001), which involves model-fitting across probes in the gene set to derive expression values. Further, often related, processing concerns thresholding/truncating extreme expression measurements, scaling (logs, standardized as here), and filtering (e.g., elimination of genes not meeting variation criteria). Imputation or other handling of missing data provides another means whereby alternate data versions are realized. Illustration of some of these aspects for a selection of public microarray datasets is provided by Dudoit *et al.* (2002).

It should be noted that there are no singly best options/specifications for any of the perturbation schemes. So, by selective choice from amongst these processing possibilities, a variety of alternate data realizations can be obtained. Then, application of a given regression method with fixed specifications to each dataset will yield a range of models. Conversely, focusing on just a single data version, but changing tuning parameters, optimization criteria, starting values, estimation methods, and/or other components of the technique, will also yield a range of models.

To illustrate the breadth of possibilities, we make concrete some of the possibilities for gene harvesting. As already demonstrated, even within the hierarchical clustering world, the choice of algorithm (linkage type) is consequential. There are several possibilities for distance metric. A tuning parameter biasing toward selection of larger clusters is provided. We note that for unconstrained harvesting, the large 687-gene cluster was chosen first even when this parameter was set to zero. The regression scheme can

accommodate differing interaction orders and include nonlinear terms. Selection of an appropriate number of terms by cross-validation requires specification of fold number and standard error multiplier.

While for harvesting the same order model is chosen with either a 1-SE or 0-SE (i.e., pick the model with smallest cross-validated prediction error), for LARS/lasso, the use of the 0-SE criteria yields a 14-gene model. This contrasts with the five-gene model (Table 3) obtained under the 1-SE rule. Furthermore, when data preprocessing is effected using dChip (see above), the respective model sizes are 5 (0-SE) and 1 (1-SE). Comparisons of selected genes and their associated pathways are beyond the scope of this paper.

Finally, having obtained a particular solution (gene set), it is possible to generate multiple solutions in post hoc fashion by selecting/enumerating from genes that are similar to those in the chosen set. Here, similarity could be based on correlation, functional class, pathway, or annotation.

A rare example of proffering multiple solution sets in the microarray (classification) context was provided by Kim *et al.* (2002). Indeed, exhaustive evaluation of all two-gene classifiers (using a variant of penalized discriminant analysis) was undertaken in contrast with the greedy forward selection approach of gene harvesting. A simple genetic algorithm was employed to search for larger gene sets. However, this two-gene limit pertained even with substantial computing power, more refined genetic algorithms were seemingly prohibitive, and optimization of the tuning (penalty/spread) parameter was not attempted.

It is evident that with the current state of microarray technology and study dimensions, differing data processing and/or modeling approaches can give very different results. This is not necessarily bad—rather, such results can be viewed from a "sensitivity analysis" perspective. The real difficulty lies in making judicious choices among the myriad processing/analysis possibilities. Ultimately, it is the biology that matters. For the Ro1 study, we are most interested in extending selected genes to biological pathways. For example, the observed coordinated down-regulation of genes in fatty acid degradation, electron transport chain, and heme biosynthesis will lead us to further examine these pathways. But, of course, choice among differing preprocessing possibilities and/or analytic methods should not be based on convenience or interpretability of results. Finally, as is widely recognized, microarray results need to be validated experimentally by another complimentary method. At least with the present state of microarray technology, claims about individual genes and pathways require verification in order to meet accepted scientific standards.

5. DISCUSSION

In this paper, we have considered regression methods for relating gene expression profiles to continuous phenotypes. Evaluation of a recently proposed method, gene harvesting (Hastie *et al.*, 2001a), revealed that results were sensitive to the clustering algorithm employed and, more importantly, subject to artifact wherein heterogeneous gene clusters whose average expression profile happened to correlate with phenotype would be inappropriately deemed important. Correcting this behavior, by limiting the harvesting approach to homogenous gene clusters, produces an algorithm that tends to select singleton genes. However, the eight-gene cluster, chosen under particular correlation constraints, was highly interpretable.

Another recent development, the LARS algorithm (Efron *et al.*, 2002), offers improvements on the forward selection strategy, as used in gene harvesting, that are especially pertinent to the microarray setting. While it would be straightforward to augment the candidate covariate (gene) pool submitted to LARS with cluster average profiles, the above experience with harvesting suggests this will add little. Similarly, basis expansion akin to that of support vector machines could also be pursued. Again, however, as transpired with SVMs, this is unlikely to yield better prediction and/or interpretation. The microarray setting, where already we have $p \gg n$, mandates stringent regularization as opposed to basis expansion; see Hastie *et al.* (2001b).

By shrinking the size of regression coefficients, LARS (and lasso) provide less greedy versions of forward selection. This is important in the microarray setting where the typically small sample sizes curtail the usefulness of greedy procedures. While SVM also shrinks coefficients, it retains the entire coefficient vector (length p) whereas LARS zeroes out all but at most n, which is interpretationally advantageous. That we observed overlap between genes selected by LARS and SVM is likely due to a combination of these genes being most correlated with Ro1 and criteria similarity (apparent from dual problems) between the methods. The fact that LARS is also computationally highly efficient and has built-in cross-validation schemes for model size determination makes it a frontline technique for regression analysis of microarray studies.

Interpretation and selection concerns warrant further attention. Analogous to issues surrounding fold change and significance inadequacy for selecting differentially expressed genes (e.g., Newton *et al.*, 2001), so to in the regression setting is it necessary to consider expression levels and variation. Given that preprocessing to standardize expression is frequently employed (Dudoit *et al.*, 2002; Lee and Lee, 2002), there is the possibility of purely correlation-based procedures, such as LARS, to select genes whose expression level is below the noise level but whose variation correlates with phenotype.

Conversely, investigators are typically not interested in the usual regression interpretation of coefficients. Rather, it is selection and perhaps ranking of genes associated with phenotype that matter. In this regard, application of methods such as random forests (Breiman, 2001b) might prove valuable. In view of this, retreating from multivariate regression approaches to assessment of univariate (individual gene) regressions is purposeful. Tusher *et al.* (2001) devise methods and software that facilitate this and which provide protection against multiple testing concerns via control of false discovery rates. Nonetheless, as we have demonstrated, application of the regression methods presented elicits genes of biologic relevance. Further, there are additional potentially important genes amongst the novel (with respect to cardiomyopathy) genes extracted.

ACKNOWLEDGMENTS

We thank Trevor Hastie, Chih-Jen Lin, Chuck McCulloch, Adam Olshen, Rob Tibshirani, Berwin Turlach, Karen Vranizan, and Mu Zhu for software assistance and/or helpful comments. This work was supported by NIH grant AI40906.

REFERENCES

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *Second Int. Symp. Info. Theory*, 267–281.

Breiman, L. 2001a. Statistical modeling: The two cultures. *Statist. Sci.* 16, 199–215.

Breiman, L. 2001b. Random forests. *Machine Learning* 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees*, Wadsworth, Belmont, CA.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., and Haussler, D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS* 97, 262–267.

Cristianini, N., and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge.

Dudoit, S., Fridyland, J., and Speed, T.P. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* 97, 77–87.

Efron, B., Hastie, T.J., Johnstone, I., and Tibshirani, R.J. 2002. Least angle regression. Submitted.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1988. Cluster analysis and display of genome-wide expression patterns. *PNAS* 95, 14863–14868.

Friedman, J.H., 1991. Multivariate adaptive regression splines. *Ann. Statist.* 19, 1–67.

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

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

Goryachev, A.B., Macgregor, P.F., and Edwards, A.M. 2001. Unfolding of microarray data. *J. Comp. Biol.* 8, 443–461.

Guardavaccaro, D., Ciotti, M.T., Schafer, B.W., Montagnoli, A., and Tirone, F. 1995. Inhibition of differentiation in myoblasts deprived of the interferon-related protein PC4. *Cell Growth Differ.* 6, 159–169.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46, 389–422.

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. 2001a. Supervised harvesting of expression trees. *Genome Biol.* 2, 0003.1–0003.12.

Hastie, T., Tibshirani, R., and Friedman, J.H. 2001b. *The Elements of Statistical Learning*, Springer-Verlag, New York.

Hoerl, A.E., and Kennard, R.W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.

Keleş, S., van der Laan, M., and Eisen, M.B. 2002. Identification of regulatory elements using a feature selection algorithm. Submitted.

Kent, C., and Carman, G.M. 1999. Interactions among pathways for phosphadidylcholine metabolism, CTP synthesis and secretion through the Golgi apparatus. *Trends Biochem. Sci.* 24, 127–162.

Kim, S., Dougherty, E.R., Barrera, J., Chen, Y., Bittner, M.L., and Trent, J.M. 2002. Strong feature sets from small samples. *J. Comp. Biol.* 9, 127–146

Kooperberg, C., Fazzio, T.G., Delrow, J.J., and Tsukiyama, T. 2002. Improved background correction for spotted microarrays. *J. Comp. Biol.* 9, 55–66.

Lee, Y., and Lee, C.-K. 2002. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Submitted.

Li, C., and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS 98*, 31–36.

Li, H., and Hong, F. 2001. Cluster-Rasch models for microarray gene expression data. *Genome Biol.* 2, 0031.1–0031.13

Mallows, C.L. 1973. Some comments on C_p. *Technometrics* 15, 661–675.

Osborne, M., Presnell, B., and Turlach, B. 2000. On the LASSO and its dual. *J. Comp. Graph. Statist.* 9, 319–337.

Nelson, D.P., Wechsler, S.B., Miura, T., Stagg, A., Newburger, J.W., Mayer, J.E., and Neufeld, E.J. 2002. Myocardial immediate early gene activation after cardiopulmonary bypass with cardiac ischemia-reperfusion.*Ann. Thorac. Surg.* 73, 156–162.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.* 8, 37–52

Redfern, C.H., Coward, P., Degtyarev, M.Y., Lee, E.K., Kwa, A.T., Hennighausen, L., Bujard, H., Fishman, G.I., and Conklin, B.R. 1999. Conditional expression and signaling of a specifically designed Gi-coupled receptor in transgenic mice. *Nat. Biotechnol.* 17, 165–169.

Redfern, C.H., Degtyarev, M.Y., Kwa, A.T., Salomonis, N., Cotte, N., Nanevicz, T., Fidelman, N., Desai, K., Vranizan, K., Lee, E.K., Coward, P., Shah, N., Warrington, J.A., Fishman, G.I., Bernstein, D., Baker, A.J., and Conklin, B.R. 2000. Conditional expression of a Gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. *PNAS* 97, 4826–4831.

Schwartz, G. 1979. Estimating the dimension of a model. *Ann. Statist.* 6, 461–464.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B.* 58, 267–288.

Tibshirani, R., and Knight, K. 1999. The covariance inflation criterion for adaptive model selection. *J. Roy. Statist. Soc. B.* 61, 529–546.

Tusher, V., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98, 5116–5121.

Vapnik, V. 1998. *Statistical Learning Theory*, Wiley, New York.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks, J., and Nevins, J. 2001. Predicting the clinical status of human breast cancer using gene expression profiles. *PNAS* 98, 11462–11467.

Westphal, C.H., Muller, L., Zhou, A., Zhu, X., Bonner-Weir, S., Schambelan, M., Steiner, D.F., Lindberg, I., and Leder, P. 1999. The neuroendocrine protein 7B2 is required for peptide hormone processing in vivo and provides a novel mechanism for pituitary Cushing's disease. *Cell* 96, 689–700.

Ye, J. 1998. On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Assoc.* 93, 120–131.

Zhang, H., Yu, C.-Y., Singer, B., and Xiong, M. 2001. Recursive partitioning for tumor classification with gene expression microarray data. *PNAS* 98, 6730–6735.

Address correspondence to:
Mark R. Segal
Department of Epidemiology and Biostatistics
University of California
500 Parnassus Avenue, MU 420-W
San Francisco, CA 94143-0560

E-mail: mark@biostat.ucsf.edu