BIOINFORMATICS

Vol. 19 no. 6 2003, pages 686–693
DOI: 10.1093/bioinformatics/btg079

# Directed indices for exploring gene expression data

*Michael LeBlanc [1,*], Charles Kooperberg [1], Thomas M. Grogan [2] and Thomas P. Miller [2]*

*[1]Fred Hutchinson Cancer Research Center, PO Box 19024, Seattle, WA 98109, USA and [2]Arizona Cancer Center, 1515 N. Campbell Ave, Tucson, AZ 85724, USA*

Received on May 25, 2002; revised on August 20, 2002; November 16, 2002; accepted on November 29, 2002

## ABSTRACT

**Motivation:** Large expression studies with clinical outcome data are becoming available for analysis. An important goal is to identify genes or clusters of genes where expression is related to patient outcome. While clustering methods are useful data exploration tools, they do not directly allow one to relate the expression data to clinical outcome. Alternatively, methods which rank genes based on their univariate significance do not incorporate gene function or relationships to genes that have been previously identified. In addition, after sifting through potentially thousands of genes, summary estimates (e.g. regression coefficients or error rates) algorithms should address the potentially large bias introduced by gene selection.

**Results:** We developed a gene index technique that generalizes methods that rank genes by their univariate associations to patient outcome. Genes are ordered based on simultaneously linking their expression both to patient outcome and to a specific gene of interest. The technique can also be used to suggest profiles of gene expression related to patient outcome. A cross-validation method is shown to be important for reducing bias due to adaptive gene selection. The methods are illustrated on a recently collected gene expression data set based on 160 patients with diffuse large cell lymphoma (DLCL).

**Availability:** A program written in the R language implementing the gene index can be obtained at http://www.crab.org/papers/

**Contact:** mikel@crab.org

## INTRODUCTION

There is an expection that subsets of thousands of gene expression measurements may be meaningfully associated with patient outcome and will help researchers understand disease biology and progression. Sufficiently powered clinical/expression studies should yield insights into the associations of gene expression to patient outcome.

Unsupervised statistical methods have been useful for studying the joint associations of gene expression data. Clustering techniques, which depend on all pair-wise associations between expression measurements, have been widely used. For clinical correlative studies, the clusters obtained from unsupervised methods have then been related to patient outcome. Alternatively, techniques have been proposed for investigating relationships between individual gene expressions and outcome (e.g. Tusher *et al.*, 2001). Other approaches have suggested supervised methods that simultaneously cluster genes and link to patient outcome (e.g. Tibshirani *et al.*, 2002).

Interpretations of large clusters of genes can be difficult, since the understanding of the biology of the given system is often quite limited. Therefore, we investigate a more directed and hopefully more interpretable strategy for investigating genes that jointly relate to patient outcome and to a specific 'reference gene' of interest. This reference gene could be the gene identified to be most strongly related to outcome, or, more likely, it may be suggested from external data such as a protein analysis using immunohistochemistry or other experimental work.

The methods are illustrated with a large expression data set consisting of 160 patients with DLCL identified through the Lymphoma and Leukemia Molecular Profiling Project (LLMPP) (Rosenwald *et al.*, 2002). This data set includes a much larger number of patients with aggressive non-Hodgkin's lymphoma (NHL) than the approximately 50 cases described in Alizadeh *et al.* (2000).

## METHODS

### Gene Indices

Assume values $x_{ij}$ for $j = 0, \ldots, p$ genes (clones) and an outcome measure $y_i$ for $i = 1, \ldots, n$ samples. For cDNA arrays these are typically the logarithm of expression ratios and for oligo type arrays they are proportional to the logarithm of expression. Our examples are based on cDNA arrays and we assume interest focuses on correlating expression with patient survival. However, there is nothing

*To whom correspondence should be addressed.

specific to survival data in the following development. The method also applies to other ordered or categorical patient outcome variables. Let $\boldsymbol{y}$ and $\boldsymbol{x}_j$, $j = 0, \ldots, p$, denote the $n$-tuples for the $n$ observations. The expression measurements, $\boldsymbol{x}_j$, are standardized within gene to have mean zero and variance one. Let $\boldsymbol{x}_0$ denote the expression measurements for the reference gene.

We consider two measures of association. Let $\widehat{\beta}_j$ be the association between outcome $\boldsymbol{y}$ and $\boldsymbol{x}_j$. We take $\widehat{\beta}_j$ to be the regression coefficient in the model relating $\boldsymbol{y}$ to $\boldsymbol{x}_j$, appropriate for the outcome of interest. Since the expression measurements are assumed to be standardized, the regression coefficients are proportional to coefficients standardized by their standard error. Let $\widehat{\rho}_{j0}$ represent the correlation between the expression measure $\boldsymbol{x}_j$ and the expression for the reference gene $\boldsymbol{x}_0$. The correlation can be standard Pearson correlation. However, we take $\widehat{\rho}_{j0}$ to be a correlation conditional on outcome to account for the potential marginal association of each expression variable with the outcome. For example, suppose there were two correlated expression variables and two classes, Class A observed for high levels of both variables 1 and 2 and Class B observed for low levels of variables 1 and 2. The Pearson correlation would be large, but for individuals within a class there may be essentially no correlation between the expression measurements. Set $\rho_{j0|y} = \text{cor}(\boldsymbol{x}_j, \boldsymbol{x}_0|y)$ and $\rho_{j0} = \int \text{cor}(\boldsymbol{x}_j, \boldsymbol{x}_0|y)dF(y)$, where $F(y)$ is the distribution of the outcome. For two class data an estimate of $\rho_{j0}$ is a weighted average of the correlation between features within each class. For linear regression, one can estimate $\rho_{j0}$ by regressing each gene expression variable $\boldsymbol{x}_j$ on $y$ and calculating the correlation between the residuals for gene $j$ and the reference gene 0. We discuss an extension of this correlation measure for survival data in the Survival Data Section .

'Target' values are defined for each of these measures; let $\beta^*$ denote the target association to the outcome, and $\rho^*$ be the target correlation to the reference gene. Examples of target correlation values to outcome could be, $\beta^* = -h$ (for some large positive $h$ ) and $\beta^* = h$, which corresponds to picking genes with a large negative and positive associations, respectively, with patient outcome. Alternatively, one could choose the target to be some moderate value, for instance, something close to $\widehat{\beta}_0$. The target correlation to the reference gene is most likely high positive, $\rho^* = 1$, or negative, $\rho^* = -1$, correlation. However, there may be some cases where there is interest in genes that are uncorrelated to the reference gene, $\rho^* = 0$.

It may also be desirable to find genes within the same (possibly functional) class. Let $c_j$ denote the class label for the $j$th gene and $c^*$ the (target) class. Typically, this is the class of the reference gene, $c_0$. In the lymphoma example described later, we identify the major histocompatibility (MHC) Class II complex. The reference target class need not be same class as the reference gene: for the lymphoma data, one may want to use a specific MHC Class II gene as the reference gene, but focus on relationships to genes within the MHC Class I group.

An univariate ordering based on a weighted combination of distances to the target parameters is constructed. To summarize, the three components of the gene index are:

(1) Correlation with patient outcome.

(2) Correlation between other gene expression and reference gene expression.

(3) Class membership of genes.

The gene index (GIN) for gene $j$ is defined as

$$Q(j) = \nu_o D_o(\widehat{\beta}_j, \beta^*) + \nu_e D_e(\widehat{\rho}_{j0}, \rho^*) + \nu_c D_c(c_j, c^*) \quad (1)$$

where the functions $D_o$, $D_e$ and $D_c$ measure the departure of $\widehat{\beta}_j$ from the target outcome association, $\widehat{\rho}_{j0}$ from the target correlation and $c_j$ from the target gene class, respectively. We use squared difference for the functions $D_o$ and $D_e$, and $D_c = 0$ if $c_j = c^*$ and 1 otherwise. There is no natural scale for (1), since it involves both correlations to outcome and between expressions, so the components are standardized to have standard deviation one. The three parameters $(\nu_o, \nu_e, \nu_c)$ specify the relative weight attached to each component. The special case of $(\nu_o, \nu_e, \nu_c) = (1, 0, 0)$ is the ranking based only on single gene associations to the outcome, and $(\nu_o, \nu_e, \nu_c) = (0, 1, 0)$ ranks genes in terms of correlation to the reference gene. In general, the GIN is a quadratic in the template coefficient $\beta^*$ and correlation $\rho^*$ that is modulated by class membership if $\nu_c \neq 0$. Figure 1 shows contours of equal GIN for a hypothetical data set (with one class of genes highlighted). The impact of a non-zero $\nu_c$ would be to move contours for genes that are not within the same class as the reference gene. In practice, it is reasonable to constrain constrain $\nu_o$, $\nu_e$ and $\nu_c$ to be non-negative and sum to one.

To summarize the ordering, we plot the marginal estimates $\widehat{\beta}_j$ versus the rank of the GIN (1) resulting in a scatterplot of marginal associations, corresponding to a list of genes. These lists of genes that can be flexibly ordered by different weights $(\nu_o, \nu_e, \nu_c)$. Of course, if interest was primarily in gene associations, one could plot correlations $\widehat{\rho}_{j0}$ versus the rank of the GIN.

## Survival Data

We illustrate our approach with the DLCL data set, described in detail in the Lymphoma Example Section. Initial investigation of this data set suggested there were associations between major histocompatibility (MHC) Class II gene expression and survival. The MHC Class II DR gene was particularly interesting, as there is a

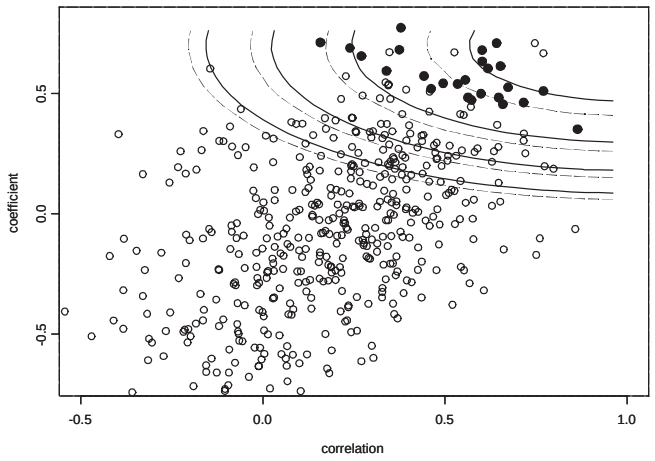


**Fig. 1.** Contour plot of the gene index (GIN) Equation 1 for hypothetical data where the target coefficient is $\beta^* = 0.7$ and target correlation is $\rho^* = 1$. The filled circles denote genes in the same class as the reference gene. The dashed contour lines apply to genes of the same class as the reference gene and the solid to other genes, assuming the gene class weight $\nu_c > 0$.

lymphoma biology publication (Miller *et al.*, 1988) based on protein data showing an important role for that gene in terms of disease progression. We choose a DR beta clone as our reference gene and considered several different weights for the GIN. Some genes are represented by multiple clones on the arrays used for the DLCL data set; we treat them separate in this analysis and to simplify text, we will refer to them as 'gene' rather than 'gene/clone'.

For a survival outcome, it was natural to choose $\widehat{\beta}_j$ to be the Cox (1972) regression coefficient for gene $j$. The survival outcome is typically coded as $y_j = (t_i, \delta_i)$ where $t_i$ is time under observation and $\delta_i$ indicates if the patient was alive ($\delta_i = 0$) or dead ($\delta_i = 1$) at time $t_i$. The hazard function for a model including a single gene is

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_j x_{ij})$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. If $\beta_j > 0$ higher values of expression are associated with worse survival (hazard ratios greater than 1) than those with lower expression, and if $\beta_j < 0$ larger values of expression are associated with better survival (hazard ratios less than 1). We standardized the expression measurements to have mean zero and variance one and we removed some genes with a large amount of missing data or low information. We chose a 1-step approximation to the Cox (1972) partial likelihood estimate, to facilitate rapid parallel computation across the genes (e.g. LeBlanc and Crowley, 1999). To estimate the correlation between genes conditional on the patients uncensored outcome, we use the individual components of the partial likelihood score vector. This correlation only depends on the ranks of the survival times. We also did the GIN analysis using standard Pearson correlation (results not shown) and the results were quite similar.

For this example a template association of survival $\beta^* = -0.5$, close to the estimate for the DR clone was chosen. An alternative would be to set $\beta^*$ to a large negative number to select the genes most strongly (negative) associated with survival. Figure 2 plots the coefficients $\widehat{\beta}_j$ for the 100 genes with the smallest value of the GIN for different choices of $(\nu_o, \nu_e, \nu_c)$. Filled circles indicate the MHC Class II genes. In Figure 2A only the outcome association has non-zero weight ($(\nu_o, \nu_e, \nu_c) = (1, 0, 0)$). Since, the expression measurements are standardized, the order on the horizontal axis is almost the same for the usual ranking of genes based on univariate significance of the null hypothesis $\beta = \beta^*$. Many of the largest negative Cox regression coefficients are from the MHC Class II genes, but there are also large negative effects for some genes not in that class. In Figure 2B we only weight the correlation of the gene expressions to the DR clone of interest ($(\nu_o, \nu_e, \nu_c) = (0, 1, 0)$). It is clear that the most highly correlated genes are also MHC Class II genes (in fact some are just different clones for the same genes). Figure 2C only weights the class of the genes corresponding to the $D_c(c, c^*)$, component ($(\nu_o, \nu_e, \nu_c) = (0, 0, 1)$), so it brings every MHC Class II expression variable to the left side to the figure. (Within a group, ties are resolved at random.) Finally, in Figure 2D we show an example of a weighted combination: in particular $(\nu_e, \nu_o, \nu_c) = (0.6, 0.2, 0.2)$ which places 60% of weight on outcome and 20% on both the correlation with DR and on MHC Class II membership. Now, most of the MHC Class II genes are in view, with some other strongly negatively prognostic expression genes. Of course other weights would change the relative importance of outcome versus the correlation of DR beta clone expression to other genes. In addition, one could then view the list of corresponding genes as ordered in the horizontal axis of the plot. This list is included as Table 1 in a supplementary document at http://www.crab.org/papers/.

If desired one could also choose weights $(\nu_e, \nu_o, \nu_c)$ to optimize some objective function related to the GIN, for example to chose weights that lead to the largest average of $\widehat{\beta}_j$ for the $q$ genes with smallest GIN. Since such a selection is adaptive, $K$-fold cross-validation could reduce selection bias, as described below.

We estimate parameter estimates $\widehat{\beta}_j$, rather than test-statistics for the hypothesis $\beta_j = 0$. As we standardize expression measurements, there is a close connection between those two summary statistics. As an alternative, one could plot individual score test statistics versus the rank of the GIN and score statistics (or a reduction in prediction error) could be used directly in the GIN as the first component to focus on the genes most predictive of

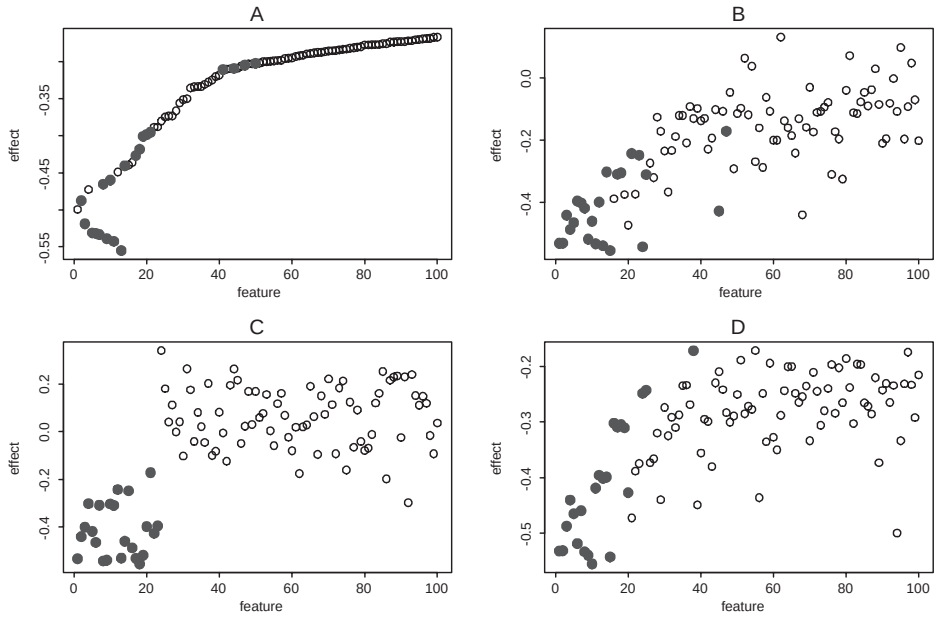


**Fig. 2.** Illustrations of the marginal effects against the rank of the GIN for the DLCL data using different weighting schemes. The filled circles denote MHC Class II genes. Plot A: weight only outcome association, $(\nu_o, \nu_e, \nu_c) = (1, 0, 0)$, with template $\beta^* = -0.5$. Plot B: weight only on correlation with DR clone, $(\nu_o, \nu_e, \nu_c) = (0, 1, 0)$. Plot C: weight only on class variable (MHC Class II), $(\nu_o, \nu_e, \nu_c) = (0, 0, 1)$. Plot D: combination weight of the three components, $(\nu_o, \nu_e, \nu_c) = (0.6, 0.2, 0.2)$.

outcome. Our software also calculates standardized test statistics rather than parameter estimates $\widehat{\beta}_j$.

### Bias Correction by Cross-validation

The small number of genes with the smallest value of the GIN (1) are selected from potentially thousands of candidate genes. Therefore, estimates of the GIN and the marginal associations $\widehat{\beta}_j$ associated with the smallest ranks are closer to the template effect $\beta^*$ than one would expect for the associations calculated on a new data set. Let $r(i)$ denote the index $j$ of the gene with the $i$th smallest value of the GIN. Ideally we would compute 'true' association $\widehat{\beta}_j$ on a large test set $(y_{new}, x_{new\,r(i)})$. As often a large test set is not available we use $K$-fold cross-validation to adjust for selection bias. The adjusted estimator is

$$\widehat{\beta}^{adj}(r(i)) = \widehat{\beta}(r(i)) + \widehat{\Delta}(r(i)),$$

where $\widehat{\Delta}(r(i))$ is the selection adjustment. The algorithm to determine $\widehat{\Delta}(r(i))$ is:

(1) $\mathcal{L}$, is divided at random into $K$ test samples $\mathcal{L}_k$ of about equal size. Let $\mathcal{L}_{(k)} = \mathcal{L} - \mathcal{L}_k, k = 1, \ldots, K$ be the training samples.

(2) Construct the GIN on $\mathcal{L}_{(k)}$.

(3) Calculate $\widehat{\beta}^k$ on $\mathcal{L}_k$.

(4) Loop over (2)–(3): $k = 1, \ldots, K$.

(5) Calculate the smooth adjustment

$$\widehat{\Delta}(r(i)) = S(\widehat{\beta}^k(r(i)) - \widehat{\beta}(r(i)))$$

$S(\cdot)$ represents a scatterplot smoother (we use 'loess' Cleveland and Devlin, 1988) estimate of the selection bias.

Typically, we choose $K = 5$ or 10, and average over a small number of repeated K-fold cross-validations to reduce variance. We choose a relatively small $K$ (rather than leave 1 out ($K = n$)) to reduce computation. See Hastie *et al.* (2001) for a discussion on how to choose $K$ for cross-validation in general. We note that if there are one or more known strong prognostic clinical factors cross-validation could be stratified on a prognostic index to reduce variance.

Figure 3 gives an illustration of the smoothly corrected 5-fold cross-validated (averaged 5 times) for the lymphoma data, using the DR beta reference gene and $(\nu_o, \nu_e, \nu_c) = (0.6, 0.2, 0.2)$. We refer to a plot like Figure 3 as a marginal GIN plot. The amount of selection bias depends on several aspects, including the weighting in the GIN, the number of genes and the strengths of association between genes and outcome. Figure 3 shows that the most extreme effects do not appear to have large bias. However, many of the cross-validated effects on the right two thirds have been shrunken substantially to

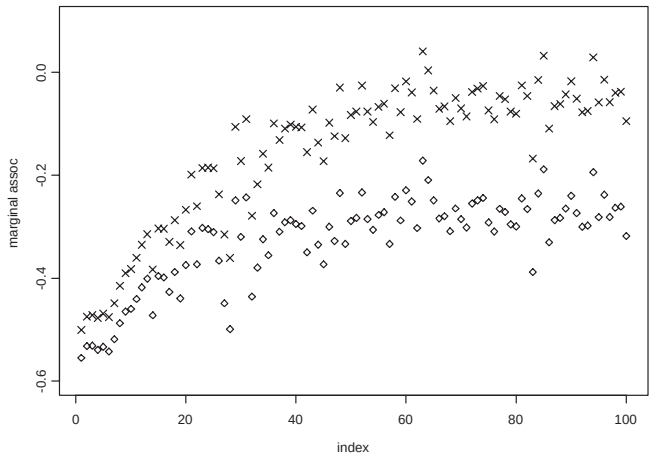


**Fig. 3.** Marginal GIN plot of the observed associations on the training data (diamonds) and associations corrected using smooth cross-validation (crosses) to adjust for the selection bias.

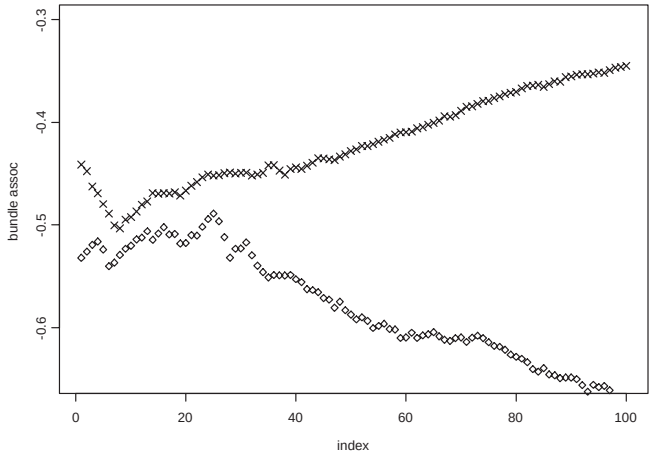


**Fig. 4.** Bundled GIN plot for the association with survival for bundled predictors based on the training data (diamonds) and by cross-validation (crosses) using MHC DR Clone as reference gene. The 'bundle association' is the Cox (1972) regression coefficient on the mean expression in for the genes in the bundle standardized to have variance equal to one. Large selection bias is present at the right side of the plot.

zero suggesting limited outcome associations with those variables.

As expected, we have noted in other plots (not shown) that a larger weight $\nu_o$ associated with the outcome association leads to larger selection bias in $\widehat{\beta}$. We confirmed this in a small simulation study (results not shown).

## Constructing Gene Summaries via Bundling

Selection of the genes with the smallest GIN was intended to order gene expression to explore individual effects of genes. However, the ordering could also be used to bundle genes. A simple bundle could be the mean of the reference gene expression and the $q$ closest genes with the smallest GIN

$$\overline{x}_q = \frac{1}{q}\sum_{i=0}^{q} x_{r(i)}.$$

The measure of association $\widehat{\beta}_{(0,q)}$ between $\overline{x}_q$ and the outcome $y$ is calculated for each $q$. As $q$ varies, the trajectory of $\widehat{\beta}_{(0,q)}$ can be used to investigate different combinations of the gene expression variables.

The bundled gene effects are also biased by the adaptive selection of the variables in the plot. Therefore, cross-validation can be used to obtained less optimistic bundle estimates. Since bundled features are already averages, the use of the smoother, as in the previous section, is not needed. In this case, one can take

$$\widehat{\beta}^{adj}_{(0,q)}(r(i)) = \text{Average}_{\{k\}}\widehat{\beta}^{k}_{(0,q)}(r(i)),$$

where $\widehat{\beta}^{k}_{(0,q)}(r(i))$ is the $k$th test sample estimate, based on the closest $q$ genes determined from the $k$th training sample. In the survival setting, we use a stratified partial likelihood Cox (1972) to calculate the 'average' $\widehat{\beta}^{adj}_{(0,q)}$. We standardize the mean expressions for each bundle to have variance one, to make the $\widehat{\beta}_{(0,q)}$ comparable for different sized bundles.

Figure 4 shows the Cox regression coefficients for the mean expression in bundle trajectory constructed using GIN weights $(\nu_o, \nu_e, \nu_c) = (0.6, 0.2, 0.2)$, corresponding to association to outcome, correlation and class membership for the MHC Class II. The variance of the mean expression is standardized to have variance one, so that the prognostic performance between smaller and larger bundles can be compared. We call Figure 4 a bundled GIN plot. The target outcome association was taken to be a large negative value ($\beta^* = -2$). The goal is to pick the largest negative associations. The lower line of points are estimates from the training data. These estimates of $\widehat{\beta}_{(0,q)}$ become slightly smaller in magnitude until approximately 25 genes. This slight decrease in magnitude is explained by the strong correlation and the weight on MHC Class II membership in the GIN. Past that point the ordering is mostly driven by large negative associations with outcome and it appears as if constructing a large bundle of genes including other genes in addition to the MHC Class II genes substantially improves the magnitude of the association to survival. The maximum standardized coefficient is at 175 genes (not shown), after that point the coefficients get smaller in magnitude (as one would expect them to get close to zero as more unrelated genes are added to the bundle). The 5-fold cross-validated estimates are quite different, however. These estimates remain quite close to the training estimates for bundles up to about 25 genes (almost all MHC Class II genes) but then the effect wains more rapidly than the learning sample as additional genes are

added. We investigated this further on a relatively small test data set available for the the Lymphoma data set consisting of an additional 80 patients. We calculated the logarithm of partial likelihood ratio on the test sample, based variables selected from learning data, and denote this as *TestLRatio*. We calculated these measures for a bundle of 25 genes (*TestLRatio* $= 1.4$) and bundles leading leading to the largest magnitude coefficient on the learning sample (175 genes) (*TestLRatio* $= 1.1$). While performance is close, there is somewhat poorer performance for the large model. A much larger test sample would be needed for a definitive conclusion. With similar performance between summaries based on different numbers of genes, a model with smaller number of genes is to be preferred due to simplicity of interpretation and because potential follow-up or validation studies may involve gene expression measurements by means other than microarrays.

We note that some researchers have constructed classifiers and predictors based on large numbers of genes. (e.g. Kato *et al.*, 2002, Khan *et al.*, 2001). Our simple example suggests there is at least the potential for selection bias to lead to erroneous conclusions that a large cluster of genes related best to outcome; while, in fact, these additional genes are just adding noise.

## LYMPHOMA EXAMPLE

We use data from previously untreated patients with the most common type of lymphoma, diffuse large cell lymphoma (DLCL). Various clinical features are known to be quite strongly associated with patient survival. A subset of these features, stage, performance status, lactate dehydroginase levels, presence of extra nodal disease have been combined to form the International Prognostic Index (IPI) (Shipp *et al.*, 1993). The IPI has been widely adopted in reporting of non-Hodgkin's lymphoma clinical studies, for stratification of new randomized studies and even for selection of patients for more aggressive therapies. However, while the IPI model has clinical utility it provides little insight into disease biology.

In contrast to the IPI there have been many biological/molecular studies of DLCL with correlations to patient outcome. Some examples include HLA-DR, proliferation or transcription (KI-67, C-MYC) and apoptosis (BCL-2) (Miller *et al.*, 1988; Grogan *et al.*, 1988; Silvestrini *et al.*, 1993; Gascoyne *et al.*, 1997; Kramer *et al.*, 1998). Typically these studies involved one or a small number of molecular factors. Recently a large high dimensional gene expression data set has been developed for DLCL as part of the LLMPP consortium. A primary goal of the LLMPP was to define the classification of human lymphoid malignancies in molecular terms. A second major goal was to define molecular correlates of clinical parameters which can be used in prognosis and in the selection of appropriate therapy for these patients. Our goal is to explore genes which appear to be predictive of patient survival but that are also related to genes documented in previous studies. We considered the sample of data from 160 patients that was used as the training data in Rosenwald *et al.* (2002).

Frozen tissue specimens and patient outcome data were collected from the seven collaborating groups. The mRNA from frozen DLCL biopsies was used to profile gene expression on Lymphochips specifically designed for expression analysis for lymphomas. The patients were mostly treated with CHOP or CHOP like regimens, which has been demonstrated in a randomized clinical trial to be standard (or preferred) therapy for advanced disease patients (Fisher *et al.*, 1993). The patients in the study all had Stage I–IV disease.

For the purposes of this example, we do not combine expression measurements for genes and instead leave each expression clone as separate variable. We filter the genes based on amount of missing elements (removing clones with $> 20\%$ missing) and minimum information (removing clones with marginal partial likelihood information in the lowest 20% of all clones) to obtain 3822 genes for consideration. The minimum information selection, is essentially a minimum variance requirement on the gene expression measurements. We chose not to adjust for the IPI in the analysis and only focus on the molecular features.

First, we explore the data with respect to important single gene associations with survival times. We use a one-step approximation to the marginal Cox (1972) regression coefficients to speed computation (e.g. LeBlanc and Crowley, 1999). As a preliminary analysis, we calculated permutation sample estimates (based on 100 permutation samples) of the false discovery rate (FDR) (the fraction of genes called associated when they are truly not associated with survival) using score tests based on the Cox model. For the FDR calculation, we compare the observed distribution of the score test statistics to the distribution of the permutation sample test statistics. We do not calculate differences from the expected order statistics as in SAM procedure (Tusher *et al.*, 2001). There appear to be clear associations between expression data and outcome in this data with the estimated FDR corresponding to the upper and lower 0.5% of the expression outcome associations being 0.24 and 0.13 respectively. Our further investigation will be based on the GIN below.

The higher levels of the MHC class of gene expression are found to correspond to better survival. We return to the analysis using the DR (beta) clone as the reference gene described in Survival Data Section and now consider labels corresponding to genes. For the lower right panel of Figure 2 the names of the top 40 genes are presented in Table 1 in a supplementary document at http://www.crab.org/papers/.

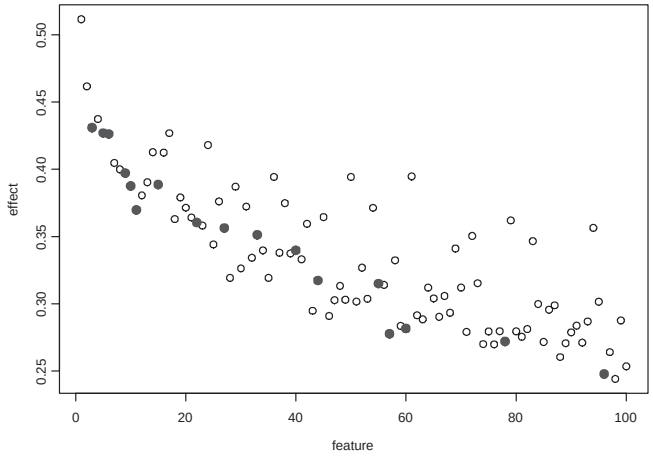


**Fig. 5.** Marginal GIN association plot corresponding to C-MYC. The ribosomal genes are denoted by the filled circles.

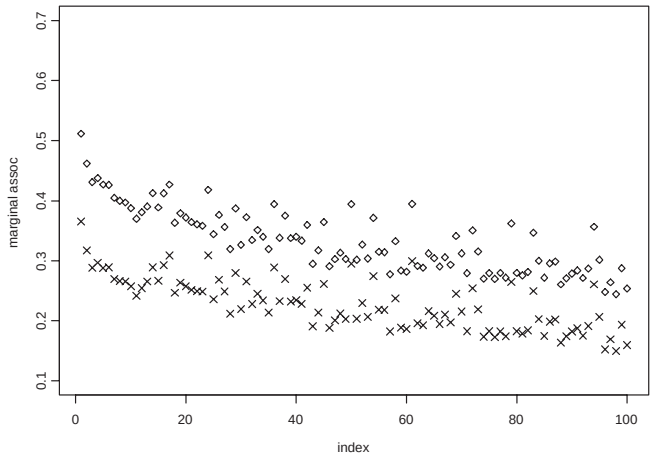


**Fig. 6.** Marginal GIN plot with associations with survival based on the training data (diamonds) and by cross-validation (crosses).

The top four entries on the list are DR beta and alpha clones. In addition, to the list of MHC Class II genes, there are also MHC Class I antigens lower in the list, including MHC Class I A2 and G. The invariant chain, which we did not label as MHC Class II, also appears in the list and is known to participate in the MHC Class II presentation.

As a second example, we took the C-MYC gene as the reference gene (Kramer *et al.*, 1998). The corresponding protein is a transcription factor that is required for proliferation. We used a weighting scheme for the GIN of $(\nu_o, \nu_e, \nu_c) = (0.75, 0.25, 0)$: we assign 75% weight to the outcome and 25% to expression correlation. The list of genes is given in Table 2 in the supplementary document. While large positive associations are identified, no clear picture emerges. There is a striking number of ribosomal genes in the list. Ribosomal genes which are involved in protein synthesis are known to be impacted by C-MYC (e.g. Chappell *et al.*, 2000). We denote the ribosomal genes on Figure 5.

The training set and cross-validated estimates of the association of individual genes are presented in Figure 6, and the corresponding estimates for bundles are presented in the supplementary document. The cross-validated estimates are considerably smaller in magnitude and emphasize the presence of a strong selection bias.

Both the results on the training data and those by cross-validation suggests that bundling a large number genes does not strengthen the C-MYC association with survival over a bundle of a small number (4–5) genes. The rapid increase in magnitude of the association for the first few genes was likely because C-MYC was not close to the strongest marginal association with survival. Therefore, it was improved upon addition of a small number of genes. However, adding other genes lower in the list, including the many ribosomal genes, does not seem to help strengthen the association with survival.

We have just presented two potential reference genes for this type of exploration. For DLCL there are a substantial list of alternatives that have been investigated in previous studies such as proliferation Ki-67, apoptotis BCL-2 and T/B cell signature genes.

## DISCUSSION

The gene index (GIN) is a simple empirical tool to aid in the statistical analysis of gene expression data. The GIN combines associations to patient outcome, expression correlations, and gene class membership to provide a rich class of indices to explore. The tool allows one to link previously studied genes to the discovery of new gene/outcome associations. Since the user needs to pick a reference gene, we think the ranking by the GIN can be easier to understand than hierarchical clustering, which leads to more symmetric joint interpretations. Simulation studies (results not reported due to space requirements) show that including the gene correlations in the GIN can increase the probability of selecting genes truly related to outcome when there are correlations between expression variables. Since the ordering is adaptively based on the marginal effects to patient outcome, there is potential for a large selection bias in the estimated effects. We have used cross-validation to adjust for this selection process. We note that gene selection bias (‘gene data dredging’) is a problem for other gene expression procedures that select among thousands of genes using small numbers of patient samples. While those data analyses will sometimes include the use of permutation sampling to ‘test’ for some association with patient outcome, unadjusted summaries such as estimated survival curves or error rates are also presented after selecting the genes. The simulated data and lymphoma data results suggest even with a relatively large

number of patients (>100 samples) the interpretation of prognostic results should incorporate reasonable methods to adjust for that gene selection bias.

We believe the current proposal will be best suited to expression studies with relatively large numbers of patient samples (e.g. >100 samples), but this is true for many other methodological proposals that attempt to link noisy patient outcome, such as survival or clinical response to expression. For other applications, where relationships are stronger, smaller numbers of samples would be sufficient. For instance, if the outcome variable was histological type, one would expect to have a stronger relationship of expression to outcome. The number of genes is less a concern, given the method functions on marginal effects and correlations. The GIN approach also is applicable for clinical studies where one wants to adjust for known clinical factors by using the adjusted score residuals after fitting those factors in a regressionmodel.

In essence, our proposal is to combine gene-gene similarity (correlation between genes) with gene-sample similarity (correlation with patient outcome) and gene-functional group similarity (class membership). Clearly these three similarities can be defined in different ways, and they can be combined in different (convex) ways. We believe our proposal is a reasonable way to do so. Generalizations are easy to imagine.

## ACKNOWLEDGEMENTS

The authors wish to thank Richard Fisher and Richard Simon for helpful comments. The Lymphoma and Leukemia Molecular Profiling Project (LLMPP) (PI L. Staudt) for interactions and for use of the gene expression data. This work was supported for M. LeBlanc by NIH CA90998, for C. Kooperberg by NIH CA74841, for both M. LeBlanc and C. Kooperberg by a pilot grant from Fred Hutchinson Cancer Research Center.

## REFERENCES

Alizadeh,A., Eisen,M., Davis,R.E., Ma,C.A., Lossos,I., Rosenwald,A., Boldrick,J., Sabet,H., Tran,T., Yu,X., Powell,J. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Chappell,S.A., LeQuesne,J.P., Paulin,F.E., deSchoolmeester,M.L., Stoneley,M., Soutar,R.L., Ralston,S.H., Helfrich,M.H. and Willis,A.E. (2000) A mutation in the c-myc-IRES leads to enhanced internal ribosome entry in multiple myeloma: a novel mechanism of oncogene de-regulation. *Oncogene*, **19**, 4437–4440.

Cleveland,W.S. and Devlin,S.J. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Amer. Stat. Assoc.*, **83**, 596–610.

Cox,D.R. (1972) Regression models and life-tables (with discussion). *J. Roy. Stat. Soc. B*, **34**, 187–220.

Fisher,R.I., Gaynor,E.R., Dahlberg,S., Oken,M.M., Grogan,T.M., Mize,E.M., Glick,J.H., Coltman,C.A. and Miller,T.P. (1993) Comparison of a standard regimen (CHOP) with three intensive chemotherapy regimens for advanced non-Hodgkin's lymphoma. *NEJM*, **328**, 1002–1006.

Gascoyne,R.D., Adomat,S.A., Krajewski,S., Krajewska,M., Horsman,D.E., Tolcher,A.W., O'Reilly,S.E., Hoskins,P., Coldman,A.J., Reed,J.C. and Connors,J. (1997) Prognostic significance of Bcl-2 protein expression and Bcl-2 gene rearrangement in diffuse aggressive non-Hodgkin's lymphoma. *Blood*, **90**, 244–251.

Grogan,T.M., Lippman,S.M., Spier,C.M., Slymen,D.J., Rybski,J.A., Rangel,C.S., Richter,L.C. and Miller,T.P. (1988) Independent prognostic significance of a nuclear proliferation antigen in diffuse large cell lymphomas as determined by the monoclonal antibody Ki-67. *Blood*, **71**, 1157–1160.

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

Kato,K., Muro,S., Takemasa,I., Matoba,R. and Moden,M. (2002) Distint molecular basis of malignacy and metastatic potential in human colorectal carcinomaa. *Human Genome Meeting, abstract*.

Khan,J., Wei,S., Ringnér,M., Saal,L., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C., Peterson,C. and Meltzer,P. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Kramer,M.H.H., Hermans,J., Wijburg,E., Philippo,K., Geelen,E., van Krieken,J.H.J.M., de Jong,D., Maartense,E., Schuuring,E. and Kluin,P.M. (1998) Clinical relevance of BCL2, BCL6, and MYC rearrangements in diffuse large cell lymphoma. *Blood*, **92**, 3152–3162.

LeBlanc,M. and Crowley,J. (1999) Adaptive regression splines in the Cox Model. *Biometrics*, **55**, 204–213.

Miller,T.P., Lippman,S.M., Spier,C.M., Slymen,D.J. and Grogan,T.M. (1988) HLA-DR (Ia) immune phenotype predicts outcome for patients with diffuse large cell lymphoma. *J. Clin. Inv.*, **82**, 370–372.

Rosenwald,A., Wright,G., Chan,W.C., Connors,J.M., Campo,E., Fisher,R.I., Gascoyne,R.D., Muller-Hermelink,H.K., Smeland,E.B., Giltnane,J.M. *et al.* (2002) Molecular Diagnosis and Clinical Outcome Prediction in Diffuse Large B-cell Lymphoma. *NEJM*, **346**, 1937–1947.

Shipp,M.A. Harrington,D.P. *et al.* (1993) A predictive model for Aggressive Non-Hodgkin's Lymphoma. *NEJM*, **329**, 987–994.

Silvestrini,R., Costa,A., Boracchi,P., Giardini,R. and Rilke,F. (1993) Cell proliferation as a long-term prognostic factor in diffuse large-cell lymphomas. *Int. J. Cancer*, **54**, 231–236.

Tibshirani,R., Hastie,T., Narasimham,B., Eisen,M., Sherlock,G., Brown,P. and Botstein,D. (2002) Exploratory screening of genes and clusters from microarray experiments. *Statist. Sinica*, **12**, 47–60.

Tusher,V., Tibshirani,R. and Chu,C. (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.