BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 4 2006, pages 466–471
doi:10.1093/bioinformatics/bti824

Gene expression

Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information

Lexin Li

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

Received on July 14, 2005; revised on November 7, 2005; accepted on December 6, 2005
Advance Access publication December 8, 2005
Associate Editor: John Quackenbush

ABSTRACT

Motivation: It is important to predict the outcome of patients with diffuse large-B-cell lymphoma after chemotherapy, since the survival rate after treatment of this common lymphoma disease is <50%. Both clinically based outcome predictors and the gene expression-based molecular factors have been proposed independently in disease prognosis. However combining the high-dimensional genomic data and the clinically relevant information to predict disease outcome is challenging.

Results: We describe an integrated clinicogenomic modeling approach that combines gene expression profiles and the clinically based International Prognostic Index (IPI) for personalized prediction in disease outcome. Dimension reduction methods are proposed to produce linear combinations of gene expressions, while taking into account clinical IPI information. The extracted summary measures capture all the regression information of the censored survival phenotype given both genomic and clinical data, and are employed as covariates in the subsequent survival model formulation. A case study of diffuse large-B-cell lymphoma data, as well as Monte Carlo simulations, both demonstrate that the proposed integrative modeling improves the prediction accuracy, delivering predictions more accurate than those achieved by using either clinical data or molecular predictors alone.

Availability: R programs are available at http://www4.stat.ncsu.edu/~li/survipi/

Contact: li@stat.ncsu.edu

Supplementary information: Supplementary data are available at http://www4.stat.ncsu.edu/~li/survipi/bioinfo-supp.pdf

1 INTRODUCTION

Diffuse large-B-cell lymphoma (DLBCL) is the most common type of lymphoma in adults and has an annual incidence of more than 25 000 cases in the United States (Jaffe, 1998). It is potentially curable by anthracycline-based chemotherapy. However, only ∼35–40% of patients are cured with this standard therapy. It is thus important to predict the outcome of the treatment and to identify DLBCL patients who are unlikely to be cured. The International Prognostic Index (IPI) has been developed for this purpose. It is based on clinical characteristics such as age, tumor stage, serum lactate dehydrogenase concentration, performance status and a number of extranodal disease sites. Although it is a well-established predictor of the survival of DLBCL patients, the outcome in patients who have identical IPI values still varies considerably. Using the IPI alone as the outcome predictor is unsatisfactory (Rosenwald *et al*., 2002).

Thanks to recent development of DNA microarray technology, there have been extensive investigations studying the relationship between prognosis and the molecular features of DLBCL. Predictive models were built employing genome-wide gene expression profiles, coupled with biological insights into the disease (Lossos *et al*., 2004), and machine learning techniques (Alizadeh *et al*., 2000; Shipp *et al*., 2002; Rosenwald *et al*., 2002). New statistical methods have been developed to address the high dimensionality and low sample size issues of the microarray data. Examples include partial least squares (Nguyen and Rocke, 2002; Bair and Tibshirani, 2004; Li and Gui, 2004), supervised principal components (Bair *et al*., 2004), penalized Cox model (Gui and Li, 2005a,b), and sufficient dimension reduction (Li and Li, 2004). All the studies have demonstrated the capability of using genomic information, in the form of gene expression patterns, to define clinically relevant molecular factors in disease prognosis.

However all the above molecular methods and the clinically based international prognostic methods were applied independently to predict survival after chemotherapy for DLBCL. On the other hand, there is a growing body of research aimed at assessing whether integrative analysis approaches may yield more accurate predictions than those obtained based on the use of of clinical or molecular information alone (see, e.g. Pittman *et al*., 2004). In this article, we propose an integrated clinicogenomic modeling approach that combines gene expression profiles and the clinically-based IPI. We demonstrate its superiority to the methods using molecular or clinical information alone. More specifically, dimension reduction techniques, including principal components analysis and sliced inverse regression (SIR), are employed to produce linear combinations of gene expressions. Such summary measures, in conjunction with IPI information, capture all the regression information of the survival phenotype given both genomic and clinical data and are obtained without imposing any probabilistic model during the dimension reduction process. The extracted linear combinations of gene expressions and the IPI are then employed as covariates in the subsequent survival model formulation. A case study of DLBCL, obtained from Rosenwald *et al*. (2002), demonstrates that the proposed integrative modeling improves the prediction accuracy, delivering predictions more accurate than those achieved by using either clinical data or genomic predictors alone. Effectiveness of the proposed method was also verified by another independent DLBCL data from Shipp *et al*. (2002) and Monte Carlo simulations.

The remainder of the article is organized as follows. In Section 2 we first introduce the framework of sufficient dimension reduction,

 © The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of North Texas on December 30, 2016

including principal components analysis and SIR. We then discuss SIR in conjunction with preserving clinical IPI information. Next we present an adaption of the dimension reduction methods to the censored survival data. We then follow with Monte Carlo simulations and the application to the real DLBCL datasets. The paper is concluded with a brief discussion.

2 METHODS

2.1 Sufficient dimension reduction

For a given outcome variable $Y \in \mathbb{R}$ and a vector of predictors $X \in \mathbb{R}^p$, the goal of sufficient dimension reduction is to find a $p \times d$ matrix $\eta = (\eta_1, \ldots, \eta_d)$, with $d \leq p$, such that

$$Y \perp\!\!\!\perp X \mid \eta^\top X, \tag{1}$$

where $\perp\!\!\!\perp$ stands for the statistical independence. It implies that the p-dimensional predictor vector X can be replaced by d-dimensional linear combinations $\eta^\top X$ without losing any information on regression of Y given X, because given $\eta^\top X$, X contains no further information about Y. The subsequent model formulation can then be restricted to the extracted $\eta^\top X$. Since in practice d is often far less than p, and in many applications, d is as small as 1 or 2, substantial dimension reduction is achieved.

It is seen that η in (1) is not unique, because multiplying η by any full rank matrix would result in (1) still holding. Therefore, we seek the linear subspace that is spanned by the columns of η. Such a space is called a dimension reduction subspace (Cook, 1998). The intersection of all the dimension reduction subspaces is often itself a dimension reduction subspace. By definition, it is unique, and is the smallest space that preserves all regression information of Y given X. It is called the central subspace, denoted by $\mathcal{S}_{Y|X}$, and is the main object of interest in our dimension reduction inquiry.

There are a number of methods to estimate $\mathcal{S}_{Y|X}$ without imposing any probabilistic models for Y given X, for instance, SIR (Li, 1991) and sliced average variance estimation (SAVE, Cook and Weisberg, 1991). In this article we focus on SIR. It is shown that, under the linearity condition discussed below, the inverse mean $E(X \mid Y)$ resides in the central subspace (Li, 1991; Cook, 1998). Thus the population solution of SIR amounts to the following eigen-decomposition

$$\Sigma_{X|Y}\eta_j = \lambda_j\, \Sigma_X \eta_j, \tag{2}$$

where Σ_X is the covariance matrix of X and $\Sigma_{X|Y}$ is the covariance matrix of $E(X \mid Y)$. The eigenvectors $\eta_1, \ldots, \eta_d$ corresponding to the d non-zero eigenvalues consist of a basis for the central subspace.

Given n independent realizations $\{(X_i, Y_i), i = 1, \ldots, n\}$ of (X, Y), SIR first partitions the range of Y into h slices so that each Y_i belongs to one of h slices. The sample estimate of $E(X \mid Y)$ is then obtained by averaging over all the X_is whose corresponding Y_is belong to the same slice. The usual sample covariance matrices $\hat{\Sigma}_{X|Y}$ and $\hat{\Sigma}_X$ are then computed and substituted in (2), resulting in the SIR sample estimates. In this procedure, h is a tuning parameter, but it has been shown by various studies that the choice of h does not usually affect the SIR estimates as long as $h > d$ (Li, 1991; Cook, 1998).

For gene expression data, the sample size n is often far smaller than the number of predictors, i.e. the individual genes. This would cause the sample covariance $\hat{\Sigma}_X$ to be non-invertible, while the sample estimation of (2) requires $\hat{\Sigma}_X$ to be invertible. To circumvent this problem, we combine principal components analysis with SIR. That is, we first obtain q principal components, $\gamma^{\mathrm{T}}X$, where γ is a $p \times q$ matrix with $q < n$. γ is taken as the first q eigenvectors of the covariance matrix Σ_X, corresponding to the largest q eigenvalues, to capture the maximum variability among the predictor space. We then apply SIR on the extracted principal components $\gamma^{\mathrm{T}}X$. This two-stage dimension reduction is justified by the assumption that $\mathcal{S}_{Y|X} \subseteq \mathrm{Span}(\gamma)$, where $\mathrm{Span}(\gamma)$ denotes a space spanned by the columns of γ, and we find this condition often holds in practice. The same strategy has been employed by Chiaromonte and Martinelli (2002) and Li and Li (2004). q is a tuning parameter for principal components, and it has been shown by Li and Li (2004) that the result of dimension reduction is not overly sensitive to the choice of q. More discussion of choosing q will be presented in Section 3.

It is noteworthy to point out that SIR does not impose any model assumptions on the distribution of $Y \mid X$. Instead, it requires the linearity condition, an assumption placed on the marginal distribution of X, which states that $E(X \mid \eta^T X = u) = A_0 + A_1 u$, where $A_0 \in \mathbb{R}^p$ and A_1 is a $p \times d$ matrix. Elliptical symmetry of the marginal distribution of X is sufficient for the linearity condition to hold (Eaton, 1986), and in particular, it holds when X is multivariate normal. Hall and Li (1993) demonstrated that the linearity condition is not a severe restriction. In addition, the condition may be induced by predictor transformation, re-weighting (Cook and Nachtsheim, 1994) or clustering (Li *et al.*, 2004).

2.2 Partial sliced inverse regression

To incorporate both genomic and clinical information in sufficient dimension reduction, we consider partial dimension reduction. Let X denote the p-dimensional gene vector, and W denote the additional clinical information. The goal of partial dimension reduction is to find a $p \times d$ matrix η, with $d \leq p$, such that

$$Y \perp\!\!\!\perp X \mid (\eta^T X, W).$$

With η identified, the subsequent modeling can be focused on $\eta^\top X$ and W without any loss of regression information of Y on X and W. We define the partial central subspace in a way similar to that of the central subspace, and denote it by $\mathcal{S}^{(W)}_{Y|X}$ (Chiaromonte *et al.*, 2002).

In the DLBCL study, the clinically based IPI takes the discrete values Low, Intermediate and High, indicating three distinct risk groups. Correspondingly, let W take values in $\{1, 2, \ldots, C = 3\}$. Chiaromonte *et al.* (2002) showed that

$$\mathcal{S}^{(W)}_{Y|X} = \oplus_{w=1}^{C} \mathcal{S}_{Y_w|X_w}, \tag{3}$$

where (X_w, Y_w) indicates $(X, Y) \mid (W = w)$ and $\oplus$ denotes the direct sum between two subspaces. Equation (3) suggests a way of estimating the partial central subspace $\mathcal{S}^{(W)}_{Y|X}$ through the combination of individual central subspace $\mathcal{S}_{Y_w|X_w}$. Specifically, consider the eigen-decomposition

$$\left(\sum_{w=1}^{C} P(W = w)\mathrm{Cov}(E(X_w|Y_w))\right)\eta_j = \lambda_j\left(\sum_{w=1}^{C} P(W = w)\mathrm{Cov}(X_w)\right)\eta_j. \tag{4}$$

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of North Texas on December 30, 2016

By (3), the eigenvectors $\eta_1, \ldots, \eta_d$ in (4) that correspond to the d non-zero eigenvalues consist of a basis for the partial central subspace. The sample estimates are obtained by substituting the corresponding sample covariance estimates in (4), and $\hat{P}(W = w) = n_w/n$, where n_w is the number of observations with $W = w$. This method is referred to as partial SIR (PSIR, Chiaromonte *et al*., 2002).

2.3 Sufficient dimension reduction for survival data

Survival data are often subject to right censoring, owing to, for instance, termination of the follow-up, or drop-out of the patients. Let T denote the true survival time, and C the censoring time, i.e. the time at which the censoring event occurs. The observed survival time is $Y = T$ if $T \leq C$, and $Y = C$ otherwise. Accordingly, let δ be a binary censoring indicator with $\delta = 1$ when $T \leq C$, and $\delta = 0$ otherwise. For sufficient dimension reduction in survival data, the central subspace $\mathcal{S}_{T|X}$ of the regression of T on X is of essential interest. However, the central subspace $\mathcal{S}_{(Y,\delta)|X}$ of the bivariate response (Y, δ) given X is what we can estimate. Cook (2002) gave the sufficient condition to connect the two central subspaces. Letting η be a basis for $\mathcal{S}_{T|X}$, it is shown that, if $(T, C) \perp\!\!\!\perp X \mid \eta^\top X$, or equivalently, $C \perp\!\!\!\perp X \mid (\eta^\top X, T)$, then

$$\mathcal{S}_{(Y,\delta)|X} \subseteq \mathcal{S}_{(T,C)|X} \subseteq \mathcal{S}_{T|X}. \tag{5}$$

Moreover, it is expected that the equality in (5) will normally hold in practice, since proper containment requires carefully balanced conditions. It is also interesting to point out that under a special case when $(T, X) \perp\!\!\!\perp C$, the sufficient condition $C \perp\!\!\!\perp X \mid (\eta^\top X, T)$ holds.

With (5), we can employ SIR of bivariate response (Y, δ) to directly construct estimates of the central subspace of regression of T given X. This is accomplished by partitioning Y into two subsamples, with $\delta = 1$ and $\delta = 0$, respectively, then slicing Y within each subsample. The remaining eigen-decomposition is the same as a standard SIR. This procedure is called double slicing (Li *et al*., 1999; Li and Li, 2004). Similarly we can apply PSIR to estimate the partial central subspace of T given X and W using double slicing. A detailed description of the proposed algorithm of PSIR for survival data is given as a web supplement.

3 RESULTS

3.1 DLBCL data of Rosenwald *et al.*

The DLBCL dataset of Rosenwald *et al*. (2002) was employed as a main illustration of application of our proposed dimension reduction methods. These data consist of measurements of 7399 genes from 240 patients obtained from customized cDNA microarrays (lymphochip). Among those 240 patients, 222 patients had the IPI recorded, and they were stratified to three risk groups indicated as low, intermediate and high. A survival time was recorded for each patient, ranging between 0 and 21.8 years. Among them, 127 were deceased (uncensored) and 95 were alive (censored) at the end of the study. A more detailed description of the data can be found in Rosenwald *et al*. (2002).

Following Bair *et al*. (2004), we divided the patients into a training group of 148 samples and a testing group of 74 samples. Additionally, a nearest neighbor technique (Troyanskaya *et al*., 2001) was applied to fill in the missing values for the gene expression data.

We fitted and compared three models for this data. In model 1, we used only the clinically based IPI as the predictor and we fitted a Cox proportional hazards model. In model 2, we applied SIR to gene expression data without taking into account IPI information and fitted a Cox model based on the extracted first SIR component. In model 3, we applied the method of PSIR by incorporating both IPI data and gene expression profiles and built a Cox model with the IPI and the extracted first PSIR component as covariates. For both models 2 and 3, principal components were first identified based on the training data. Approximately 55–95% of total predictor variations were accounted for, with the number of principal components ranging from 20 to 120. We chose $q = 40$ PCs, which accounts for $\sim$70% of the total variation, for subsequent analysis. The choice of q will be discussed later. The extracted PCs were then employed as input variables for inverse regression estimation. Additionally, examining the marginal scatter plot of the 40 principal components reveals no strong violation of the linearity condition. Table S1 (web supplement) summarizes the fitted models based on the training samples. In the table, the discrete-valued IPI variable was coded as two indicator variables, IPI-Intermediate and IPI-High, with the first level of IPI, IPI-Low, as the baseline. It is seen from the table that all terms were significant, with P-values < 0.0001. This indicates that the patients' survival may be best predicted by incorporating both clinical and genomic information.

We next compared the three models in predicting patients overall survival. Figure 1 shows the Kaplan–Meier estimates of survival curves for three groups of patients defined by the fitted models, the low-risk patients, the intermediate-risk patients and the high-risk patients. For model 1, there are naturally three groups. For models 2 and 3, the cutoff values for the three risk groups were determined by the 33 and 66% quantiles of the estimated scores based on the training data. The same cutoff values were then applied when assigning the test samples into three risk groups. The log-rank test of difference among three survival curves is reported, with a smaller P-value indicating a better model fitting for the training data, and better prediction for the testing data. It is first noted that all three models achieved good separation of three risk groups for the training data. Among them, models 1 and 2 showed similar performance, while model 3 was slightly superior. The log-rank test yielded P-values of 2.14e−09, 1.47e−08 and 0, for models 1, 2 and 3, respectively, which confirms our visual examination. For the testing data, using the IPI score alone (model 1) provides reasonably good stratification of different risk groups of patients, with a P-value of 0.02470 for the log-rank test, verifying the value of this well-established clinical prognosis indicator. On the other hand, using gene expression independently of the IPI (model 2) yielded a better stratification than using IPI alone, with a P-value of 0.00171 for the log-rank test. This demonstrates the predictive power of the gene expression profiles of DLBCL, which agrees with the findings of Rosenwald *et al*. (2002), Bair and Tibshirani (2004), Bair, *et al*. (2004) and Li and Li (2004). By combining both clinical and genomic information, model 3 yielded a P-value of 0.00006 for the log-rank test, thus is demonstrated to have the best predictive performance in predicting future patients' survival risks.

To further evaluate and compare the predictive performance of those three models, we employ the time-dependent receiver-operator characteristics (ROC) curve for censored data and the area under the curve (AUC) as the criterion. These methods were developed by Heagerty *et al*. (2000) in the context of medical

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of North Texas on December 30, 2016

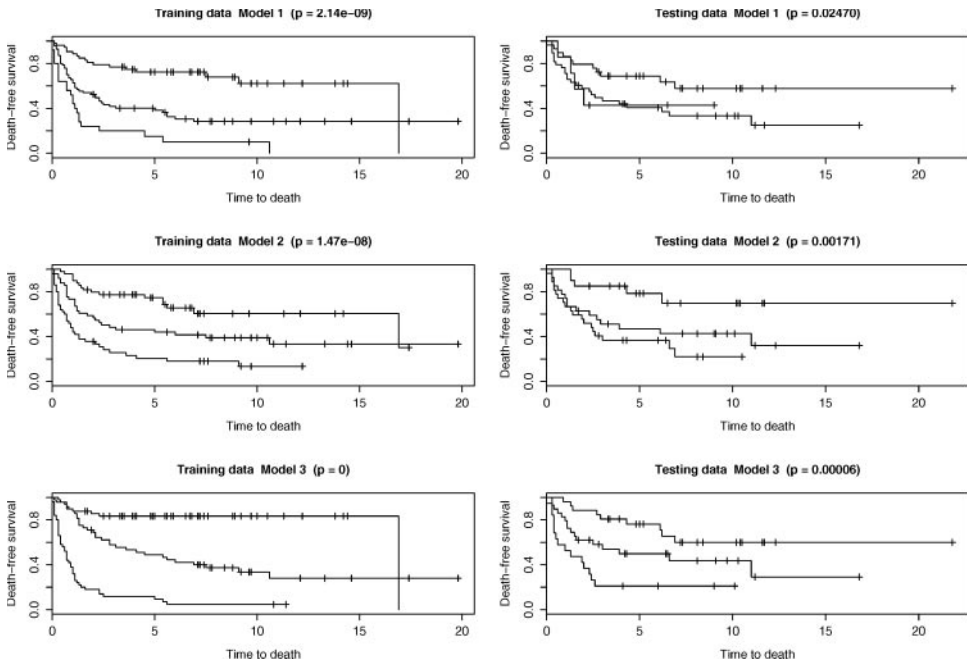

Fig. 1. Kaplan–Meier estimates of survival curves for three risk groups of patients defined by the fitted models. The left panels are for the training data, and the right panels are for the testing data. The *P*-values of the corresponding log-rank test are indicated. Model 1 uses clinical IPI information only, model 2 uses gene expression data with PC + SIR, and model 3 uses both clinical and genomic information with PC + PSIR.

diagnosis. The idea is to use sensitivity and specificity, which in this case are both time dependent, to measure the prognostic capacity of a given survival model. More specifically, for a given score function $f(x)$, we define the time-dependent sensitivity and specificity functions as

$$\text{sensitivity}(c, t \mid f(x)) = \Pr\{f(x) > c \mid \delta(t) = 1\},$$
$$\text{specificity}(c, t \mid f(x)) = \Pr\{f(x) \le c \mid \delta(t) = 0\}.$$

The corresponding ROC($t \mid f(x)$) curve, for any time t and the score function $f(x)$, is defined as the plot of {sensitivity($c, t \mid f(x)$)} versus {1 − specificity($c, t \mid f(x)$)}, with the cutoff point c varying. The area under the curve, AUC($t \mid f(x)$), is then defined as the area under the ROC($t|f(x)$) curve. Here $\delta(t)$ is the event indicator at time t. A nearest neighbor estimator for the bivariate distribution function is used for estimating these conditional probabilities accounting for possible censoring (Akritas, 1994). Following these definitions, a larger AUC at time t indicates a better predictability of time to event at time t, as measured by sensitivity and specificity evaluated at time t. Figure 2 shows the AUCs for the three models with the survival time ranging from 1 to 10 years. This plot reinforces our observations in Figure 1 that model 3 has the best performance both in fitting the training data and in predicting the future patients' survival risks among the three models. Additionally, models 2 and 3 had similar predictability of time to event when it is 7 years or longer. In summary, the model combining both clinical and genomic information delivered predictions more accurately than those made using clinical or genomic data alone.

The proposed dimension reduction methods employ both principal components and SIR. As a comparison, we also examined the performance of a Cox regression using PC alone. Specifically, a Cox model with 40 principal components as predictors (model a) and a Cox model with 40 PCs plus IPI (model b) were fitted to the training data and then evaluated on the testing data. The log-rank test for the testing data yielded the *P*-values of 0.00504 and 0.00397 for models a and b, respectively. Both are superior compared with the Cox model using IPI alone, but both are outperformed by the models employing SIR. This is further verified by the AUCs as

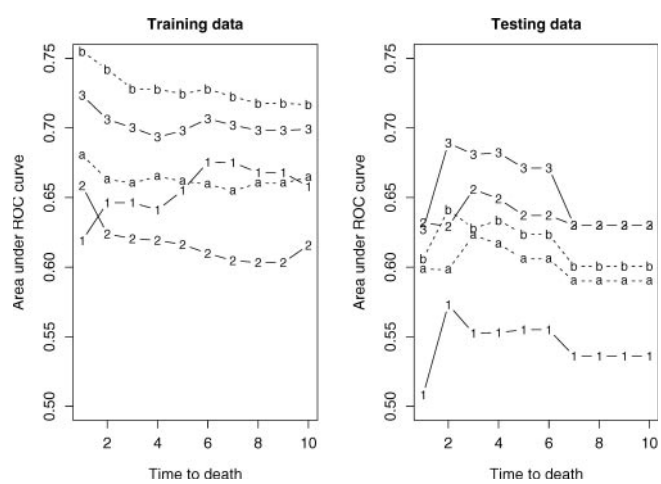

Fig. 2. Area under ROC curves at time 1 to 10 years for three models. The left panel is for the training data, and the right panel is for the testing data. Model 1 uses clinical IPI information only, model 2 uses gene expression data with PC + SIR, and model 3 uses both clinical and genomic information with PC + PSIR. As a comparison, model a uses gene expression data with PC only, and model b uses both clinical and genomic information with PC only.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of North Texas on December 30, 2016

shown in Figure 2. The Cox model with IPI and PC alone seems to overfit the training data, and our proposed integrated method employing PC, SIR and IPI dominates all other methods in prediction.

The choice of the number of principal components q was also examined using a 10-fold cross-validation. For both SIR and PSIR, we tried a sequence of q values ranging from 20 to 120. Models were fitted to 9/10 of the training data and were evaluated for the remaining 1/10 of the samples. Figure S1 (web supplement) shows the average AUCs of cross-validation for model 2 (left panel) and model 3 (right panel). It is noted that the AUCs are very close for all qs, with a possible overfitting for a very large value of q. This demonstrates the relative insensitivity of the dimension reduction methods to the choice of q. We also found through cross-validation that the performance of the proposed method is stable.

3.2 DLBCL data of Shipp *et al.*

To further test the robustness of our proposed methods with respect to cross-platform prediction, we applied the methods to the DLBCL data of Shipp *et al*. (2002). For 56 DLBCL patients, the IPI was recorded, and the expression values of 7129 genes were measured using Affymetrix microarrays. Among them, 26 patients died of lymphoma or experienced recurrent refractory or progressive disease (uncensored), and 30 patients remained disease-free (censored). The survival time or time to recurrence ranged between 3.2 and 182.4 months. Shipp *et al*. (2002) give a more detailed description of this dataset.

We first randomly partitioned all patients into a training group of 40 and a testing group of 16. We then applied the proposed dimension reduction methods jointly with a Cox model fitting. The patients were originally divided by Shipp *et al*. (2002) into four groups according to their IPI scores, Low, Low-Intermediate High-Intermediate and High. Owing to the small sample size in each group, we combined the patients in the first two groups as the Low-IPI group and the remaining as the High-IPI. We again compared models 1–3 as in the previous example, and we focused on the prediction performance of the fitted models. The results agree with those from the previous example. Model 1 which employs IPI as the sole covariate and model 2 using only expression predictor performed similarly, with the P-value for the log-rank test equal to 0.0013 and 0.0480, respectively. Model 3 that incorporates both clinical and genomic information performed best, yielding a P-value of 0.0006. We also point out that the sample size of this data is relatively small, while the proposed methods work best with large samples.

3.3 Monte Carlo simulations

We have conducted Monte Carlo simulations to evaluate the performance of the proposed dimension reduction methods and compared various models. Gene expression data with $p = 1000$ genes was simulated. The first 30 predictors are independent normal random variables with variance 5, and the remaining predictors are independent standard normal variables. Let X denote this 1000-dimensional gene vector. A binomial random variable W with two trials and success probability of 0.4 at each trial was also generated. Let W_j denote the level j of W with the first level of W as the reference, $j = 2, 3$. The survival time is related with X and W through a Cox proportional hazards model with the score function $f(x, w)$, where $x = \beta^{\mathrm{T}}X$, $w = W_2 + 1.5W_3$ and β has the first 30 elements taking the value of $1/\sqrt{30}$, and the remaining components equal to 0. Four different score functions were examined:

$$\begin{aligned}
&\text{Design 1:} \quad f(x,w) = -12w \\
&\text{Design 2:} \quad f(x,w) = 5x \\
&\text{Design 3:} \quad f(x,w) = 10w + 5x \\
&\text{Design 4:} \quad f(x,w) = 7w + 3x - x^2/2
\end{aligned}$$

A Weibull distribution with the shape parameter 10 and the scale parameter 1 was used for the baseline hazard function. Censoring time was simulated from a uniform distribution $U(0, 8)$, yielding a censoring proportion ranging between 10 and 35%. The observed survival time ranges between 0 and ~6 (years). The sample size was taken as $n = 200$.

For the simulated training data, four models were fitted. Model 1 used only W as the predictor; model 2 applied SIR to the X data and used the extracted first SIR component as the predictor; model 3 applied partial inverse regression to both W and X, and regressed on both W and the extracted first PSIR component; model 4 added the quadratic term of the PSIR component to model 3. The areas under the ROC curves were then evaluated on an independently generated testing data, and this procedure was repeated 100 times. The median and the median absolute deviation of the AUCs at time 1, 3 and 6 (years) are summarized in Table S2 (web supplement) for different designs and different models.

For Design 1, the survival time is based on W only. Model 1 showed the best performance, while model 2 failed in this case since it completely ignores the information of W. For Design 2, the survival time relates only with X, thus model 2 performed well, while model 1 failed. In both cases, models 3 and 4 were comparable with the best performing model. For Design 3, model 3, which incorporates both W and X information, was superior than either model 1 with W alone or model 2 with X data alone. For Design 4 where a quadratic term of X data is present, model 4 worked the best, as we would have expected. Overall, models taking into account both W and X information performed the best, especially when both data were involved in the true data generation.

4 DISCUSSION

We have proposed an integrated modeling approach, which combines genomic information, in terms of gene expression profiles, and the clinically based IPI, to predict the survival of patients with DLBCL after chemotherapy treatment. Dimension reduction techniques were employed to reduce the high dimensionality of gene expression data, meanwhile taking into account the clinical information. The survival phenotype was preserved, and the survival model was formulated based on the reduced-dimensional genomic and clinical covariates. Both the real data analysis and the Monte Carlo simulations demonstrated that the proposed integrative modeling improved the prediction accuracy over those methods using either clinical or genomic factors alone.

The focus of this article is on the prediction of patients' survival, and the proposed methods were designed for this purpose. In this case we assume there exist a large number of genes that jointly regulate the phenotype, and the extracted gene components after principal components and SIR may be regarded as 'supergene' factors (West *et al*., 2001). However, in many studies, it is of great interest to identify individual genes that are most significantly correlated with the phenotype and have the best predictive power.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of North Texas on December 30, 2016

Dimension reduction methods that integrate both outcome prediction and gene selection are currently under active investigation. An alternative method for integrative prediction and gene selection is the penalized estimation approach. Gui and Li (2005a,b) explored the L1-penalized Cox model and the threshold gradient descent method using gene expression data alone to predict the patients' survival. Those methods are computationally intensive but promising. However, no studies have been published using the penalized methods on combined genomic and clinical data; this line of research is being investigated.

ACKNOWLEDGEMENTS

This research was supported by NIH grant ES11269. The author thanks two anonymous referees, an associate editor and the editor for comments and suggestions that improved many aspects of this article.

Conflict of Interest: none declared.

REFERENCES

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large-B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Akritas,M.G. (1994) Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Stat*., **22**, 1299–1327.

Bair,E. and Tibshirani,R (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*., **2**, 511–522.

Bair,E. *et al*. (2004) Prediction by supervised principal components. *J. Am. Stat. Assoc*. (in press).

Chiaromonte,F. and Martinelli,J. (2002) Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci*., **176**, 123–144.

Chiaromonte,F *et al*. (2002) Sufficient dimension reduction in regressions with categorical predictors. *Ann. Stat*., **30**, 475–497.

Cook,R.D. (1998) *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, NY.

Cook,R.D. (2002) Dimension reduction and graphical exploration in regression including survival analysis. *Stat. Med*., **22**, 1399–1413.

Cook,R.D. and Nachtsheim,C.J. (1994) Re-weighting to achieve elliptically contoured covariates in regression. *J. Am. Stat. Assoc*., **89**, 592–600.

Cook,R.D. and Weisberg,S. (1991) Discussion of Li (1991). *J. Am. Stat. Assoc*., **86**, 328–332.

Eaton,M. (1986) A characterization of spherical distributions. *J. Multivariate Anal*., **20**, 272–276.

Gui,J. and Li,H. (2005a) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001–3008.

Gui,J. and Li,H. (2005b) Threshold gradient descent method for censored data regression, with applications in pharmacogenomics. *Pac. Symp. Biocomput*., **10**, 272–283.

Hall,P. and Li,K.C. (1993) On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist*., **21**, 867–889.

Heagerty,P.J. *et al*. (2000) Time dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**, 337–344.

Jaffe,E.S. (1998) Histopathology of the non-Hodgkin's lymphomas and Hodgkin's disease. In Canellos,G.P., Lister,T.A. and Sklar,J.L. (eds), *The Lymphomas*. W.B. Saunders, Philadelphia, pp. 77–106.

Li,H. and Gui,J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, **20**, i208–i215.

Li,K.C. (1991) Sliced inverse regression for dimension reduction (with discussion). *J. Am. Stat. Assoc*., **86**, 316–327.

Li,K.C. *et al*. (1999) Dimension reduction for censored regression data. *Ann. Stat*., **27**, 1–23.

Li,L. and Li,H. (2004) Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 3406–3412.

Li,L. *et al*. (2004) Cluster-based estimation for sufficient dimension reduction. *Comput. Stat. Data Anal*., **47**, 175–193.

Lossos,I.S. *et al*. (2004) Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New Engl. J. Med*., **350**, 1828–1837.

Nguyen,D.V. and Rocke,D.M. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625–1632.

Pittman,J. *et al*. (2004) Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl Acad. Sci. USA*, **101**, 8431–8436.

Rosenwald,A. *et al*. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New Engl. J. Med*., **346**, 1937–1947.

Shipp,M.A. *et al*. (2002) Diffuse large-B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nat. Med*., **8**, 68–74.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

West,M. *et al*. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of North Texas on December 30, 2016