Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2006 *Paper* 7

Survival Analysis of Longitudinal Microarrays

Natasa Rajicic* Dianne M. Finkelstein† David A. Schoenfeld‡

*MGH Biostatistics, nrajicic@partners.org
†
‡

Survival Analysis of Longitudinal Microarrays

Natasa Rajicic, Dianne M. Finkelstein, and David A. Schoenfeld

Abstract

Motivation: The development of methods for linking gene expressions to various clinical and phenotypic characteristics is an active area of genomic research. Scientists hope that such analysis may, for example, describe relationships between gene function and clinical events such as death or recovery. Methods are available for relating gene expression to measurements that are categorized or continuous, but there is less work in relating expressions to an observed event time such as time to death, response, or relapse. When gene expressions are measured over time, there are methods for differentiating temporal patterns. However, no methods have yet been proposed for the survival analysis of longitudinally collected microarrays. Results: We describe an approach for the survival analysis of longitudinal gene expression data. We construct a measure of association between the time to an event and gene expressions collected over time. The issue of high dimensionality and dependence when assessing statistical significance is addressed using permutations and control of the false discovery rate. Our proposed method is illustrated on a data set from a multi-center research study of inflammation and response to injury that aims to uncover the biological reasons why patients can have dramatically different outcomes after suffering a traumatic injury (www.gluegrant.org).

1

Survival Analysis of Longitudinal Microarrays

This is a pre-editing, author-produced PDF of an article accepted for publication in *Bioinformatics* following peer review.

Abstract

Motivation: The development of methods for linking gene expressions to various clinical and phenotypic characteristics is an active area of genomic research. Scientists hope that such analysis may, for example, describe relationships between gene function and clinical events such as death or recovery. Methods are available for relating gene expression to measurements that are categorized or continuous, but there is less work in relating expressions to an observed event time such as time to death, response, or relapse. When gene expressions are measured over time, there are methods for differentiating temporal patterns. However, no methods have yet been proposed for the survival analysis of longitudinally collected microarrays.

Results: We describe an approach for the survival analysis of longitudinal gene expression data. We construct a measure of association between the time to an event and gene expressions collected over time. The issue of high dimensionality and dependence when assessing statistical significance is addressed using permutations and control of the false discovery rate. Our proposed method is illustrated on a data set from a multi-center research study of inflammation and response to injury that aims to uncover the biological reasons why patients can have dramatically different outcomes after suffering a traumatic injury (www.gluegrant.org).

Contact: natasa.rajicic@pfizer.com

1.1 Introduction

Scientists are turning to the microarray technology for insights into the mechanisms of the human body that were previously poorly understood. We think of genes as units of heredity as they record the genetic makeup of organisms. Though it is believed that a large number of genes remain inactive for most of our lives, there are those genes for which the activity can be associated with various physiological or environmental effects. In simple terms, a gene is considered to be activated, or *ex-*

pressed, if its coded information is converted into proteins which are the main instigators of functions and processes in our bodies. An interesting problem in the analysis of the human genome is to relate changes in gene activity to clinical or phenotypic information. For example, scientists have been able to relate gene expression to the clinical implications of different types of cancer [Alizadeh *et al.*, 2000, Golub *et al.*, 1999, van't Veer *et al.*, 2002, van de Vijver *et al.*, 2002].

The nature of the data generated from a microarray experiment poses specific challenges to the statistical analysis. In a typical experiment, data from a relatively small number of subjects is available on thousands, even tens of thousands of genes, which makes many of the classical statistical procedures unapplicable. This is because the standard statistical methods are developed with the classical data type in mind, where the number of explanatory variables does not exceed the number of subjects on which data is collected. Time-to-an-event data poses additional challenges due to the presence of censoring.

We propose a method to study relationships between repeatedly collected gene expressions and time to an event of interest. Our problem is motivated by the data from Inflammation and Host Response to Injury research project (also referred to as the Glue grant, www.gluegrant.org). This multi-center and multi-disciplinary collaboration aims at better understanding of processes involved in the immune system's response to injury, as well as uncover the biological reasons why seemingly alike patients can have dramatically different outcomes after suffering a traumatic injury or burns. Doctors hope that identifying the genetic factors will help predict the course of recovery of severely injured patients.

Numerous methods have been proposed and developed for relating gene microarrays to either continuous or categorical measurements such as comparison of treatment groups or the level of a known biomarker. Ring & Ross (2002) offer a comprehensive review of methods that use microarrays for tumor classification. In comparison, fewer methods have been suggested for the analysis of gene expressions in

relation to time to an event (e.g., death, response, or relapse), or for detecting differences in genes over time [Luan & Li, 2002, Yeung *et al.*, 2003, Storey *et al.*, 2005]. Methods for the use of longitudinal microarray in either describing or predicting survival outcomes are currently unavailable.

Methods for relating gene activity to the occurrence of an event have been proposed when microarrays are collected at a single point in time (e.g., baseline). One approach is to first use an unsupervised classification method, e.g. hierarchical clustering, to generate two or more groups of patient samples [Rosenwald *et al.*, 2002, Makretsov *et al.*, 2004]. The survival distributions within such generated clusters are then compared using the logrank test and displayed by Kaplan-Meier curves. A second, related approach is to first cluster genes based on their expressions across different patient samples, and then use cluster averages of the gene expressions as explanatory variables in a Cox proportional hazard regression model [Li & Luan, 2003]. However, both of these approaches do not capture the marginal relationship between gene expressions and time to an event [Sorlie *et al.*, 2001, Jung *et al.*, 2005]. One may end up with gene classes that do not represent any meaningful grouping in terms of the survival, or the results may vary due to a particular clustering algorithm used.

A number of published approaches apply partial least squares (PLS) method to generate linear combinations of gene expressions as predictors in the proportional hazards model [Nguyen & Rocke, 2002]. PLS is a method related to principal components analysis (PCA), but while PCA creates combinations that maximize the explained variability among predictors only, PLS aims at maximizing correlations between predictors and the response variable. Bair & Tibshirani (2004) first calculate the Cox score for each gene (a statistic based on a proportional hazards partial likelihood) in order to select a subset of genes, then employ the PLS method to a reduced set of genes to arrive at the best model. Park *et al.* (2002) first reformulate the survival outcomes problem into a generalized linear (Poisson) regression, then apply the PLS

algorithm to derive a parsimonious model. In a related approach, [Li & Gui, 2004] and [Gui & Li, 2004] reduce the dimensionality of the microarray predictor space by either partial or penalized Cox regression. While in all of these approaches the high-dimensionality of the microarray is reduced, direct interpretation of the fitted parameters in terms of the individual genes is not possible. In contrast, we are interested in developing a method to be used as a first step in identifying individual genes for further investigation.

When thousands of genes are measured in a single experiment, a single question of interest can be formulated as simultaneous testing of numerous individual hypotheses. Testing many statistical hypotheses at once increases the possibility of a Type I error, as a significant result may occur purely by chance, regardless of the nature's true state. The control of the false discovery rate (FDR) has become a widely used method of error control in the analysis of gene microarrays [Storey & Tibshirani, 2003]. The control of FDR involves an estimate of the proportion of falsely positive genes among all genes found *positive* (i.e., exhibit differential expression in different samples or states under investigation). Westfall & Young (1989) promoted the use of permutations to allow for dependencies among test statistics. In this paper, we propose a permutation-based method that is related to the method of [Storey & Tibshirani, 2003] and to the popular SAM method [Tusher *et al.*, 2001].

A gene-specific test statistic is defined in Section 2.7. A multiple testing algorithm that controls the number of false positive findings is described in Section 2.2.3. The results of a series of simulations are presented in Section 1.3, while the analysis on the data from the study of inflammation and response to trauma is presented in Section 1.4. All programs for the analysis presented in this paper were done using R statistical package [R, 2005] and can be obtained from the first author (N.R.) or from **http://hedwig.mgh.harvard.edu/biostatistics/software.php**.

1.2 Methods

1.2.1 Notation

For each of the n patients enrolled in a clinical study, we record the elapsed time from the beginning of the study to the occurrence of the event of interest (e.g., death or recovery). Since we are interested in examining whether there is an association between the time to an event and changes in gene expression, we also observe patients' microarray over time. A microarray here is a collection of p expression values on p different genes. For a given gene, let $X_i(t)$ denote expressions for patient i at time t. Time to occurrence of the event is recorded using two indicator variables, δ and Y. Say subject i had an event at time t, then $\delta_i(t) = 1$ and $\delta_i(t) = 0$ for times prior to time t. Similarly, $Y_i(t) = 1$ if i is still *at risk* at time t, meaning that the patient is under observation and has not experienced the event by time t. A set of subjects remaining at risk at time t is of size $Y(t) = \sum_i Y_i(t)$. Let $n(t) = \sum_i \delta_i(t)$ denote a total number of subjects who experienced an event at time t. For patients for which the event does not occur for the duration of the study, or who for other reasons have discontinued the study followup, we say that their event time is censored. We further make a distinction between the observed event times, $\tau_k,\ k = 1 \ldots m$, and scheduled (i.e. planned) visit times, $t_j,\ j = 1 \ldots J$, as the planned timing of visits may not coincide with the observed event times. Here, m is the total number of observed events, and J is the total number of scheduled visits. Since events are considered to be '*terminal*', each subject can experience it only once during the study follow-up, and the total number of observed events, m, is less or equal to the total number of subjects, $m \leq n$.

1.2.2 Test statistic

To test the association between thousands of longitudinally collected gene expressions and time to an event of interest, we want to use a test statistic that is not only intuitive but simple to calculate. This is primarily because we intend to use a permutation-based testing procedure which is computationally intensive. The approach we take is to first calculate one test statistic per each gene, then determine the significance of each association using permutations. We begin by examining a general class of nonparametric tests for survival data, formulated by [Jones & Crowley, 1989]. We assume that for each subject i at risk at time t, it is possible to define a (quantitative) value $Z_i(t)$ that represents subject's covariate measurement, and denote by $\bar{Z}(t)$ the average value of Z for subjects at risk at time t. It is also assumed that the increasing covariate values correspond to either increasing or decreasing chances of event occurrence. Let m denote the total number of *observed* events, then the test statistic can be written as:

$$T(\omega, Z) = \sum_t \omega(t) \sum_i \delta_i(t)[Z_i(t) - \bar{Z}(t)]\,. \tag{1.1}$$

Note that $\omega(t)$ are optional weights chosen to emphasize either early or late events. The score statistic based on the partial likelihood from a Cox regression model [Cox, 1972] is a member of this general class. Jones & Crowley investigate various choices for weights and labels $(\omega(t), Z_i(t))$, where using $(1, X_i(t))$ results in the following test statistic:

$$T = \sum_t \sum_i \delta_i(t)[X_i(t) - \bar{X}(t)]\,. \tag{1.2}$$

Here, $\bar{X}(t)$ is the average gene expression for subjects at risk at time t, $\bar{X}(t) = (1/Y(t)) \sum_i Y_i(t)\, X_i(t)$. This test statistic captures the difference between the observed covariate values for subjects that had an event at a given time-point, and the average covariate value for subjects still at risk an instant before the event occurred. The differences are then summed up over all observed event times.

In our approach, however, we cannot implement (1.2) without further modification.

This is because the structure of our problem presents several challenges that need to be addressed. We want the outer summation in both (1.1) and (1.2) to be over the unique observed event times, τ_k, which results in summands involving current gene expression values at the time of an observed event. Ideally, if we had observed expression values for all subjects currently at risk at time of an observed event, the statistic would be easily and correctly calculated. In clinical trials, however, data are often collected according to some schedule of study visits. The data on time-varying covariates for all subjects at risk may not be available at the time an event occurred but rather at more than one prior scheduled visit times. For example, in the Glue study, gene expression is obtained on 7 scheduled visits over a period of 28 days but respiratory recovery event can occur and be recorded on any day during that time. A simple approach to deal with intermittent covariate data is the "last observation carried forward" approach (LOCF), where the most recent available observation is used in place of the missing data. If the event had occurred at some time k between the two scheduled visits t_{j-1} and t_j, so that $X_i(\tau_k)$ is not available, microarray collected at time t_{j-1} would be used in place of $X_i(\tau_k)$. While the last observation carried forward approach would be simple to implement, it has been traditionally heavily criticized as it produces biased results. We therefore explore a different approach to handling intermittent microarray data.

1.2.3 Semi-parametric test of association

Another way to deal with the intermittently available data in a time-to-event study is to model unknown values using measurements available up to that time. Taking into account our limited knowledge about the longitudinal behavior of microarrays, we search for ways to model the gene expression over time without assuming strict distributional properties. In order to do this, we follow the approach outlined in [Tsiatis & Davidian, 2001], who extend and apply the concept of a conditional score [Stefanski & Carroll, 1987] to the joint modelling of the longitudinal and event

data. Namely, we would like to consider a random effects model for the longitudinal expression data, as this model provides a way to incorporate subject-specific random intercept and slopes: $X_i(t) = \alpha_{0i} + \alpha_{1i}t,\ i = 1 \ldots n$. In this way, we also address the issue of between-subject gene expression variability which can be substantial (Cheng Li, personal communication).

However, the model estimate of the unknown gene expression value $X_i(t)$ at time t would require a distributional assumption for $\alpha_i = [\alpha_{0i}, \alpha_{1i}]$. As noted earlier, due to the unknown longitudinal behavior of genes, we would like to avoid specifying the exact distribution of the random effects. Fortunately, if we know the sufficient statistic for α_i, we can use it to avoid making assumptions on the distribution of the random effects. This is because conditioning on the sufficient statistic, $S_i(t)$, would remove the dependence of the conditional distribution on the random effects α_i. Namely, the sufficient statistic for α_i, conditional on subject i being at risk at time t, (i.e., $Y_i(t) = 1$), is

$$S_i(t) = \beta\sigma^2(t)\delta_i(t) + \hat{X}_i(t).$$

Here, $\sigma^2(t)$ accounts for the uncertainty when using $\hat{X}_i(t)$ as an estimate of the unknown covariate value $X_i(t)$. We describe the estimation of $\sigma^2(t)$ below. More importantly, assuming subject i is at risk at time t, and using all available data up to and including t, $\hat{X}_i(t)$ can simple be the ordinary least square estimate. The joint likelihood of the events $\delta_i(t)$ and the model estimate $\hat{X}_i(t)$ can then be factored into two parts, one of which does not involve a random variable α_i, and an other that does not involve information on the event:

$$L(\delta_i(t), \hat{X}_i(t)|\alpha_i) = L(\delta_i(t)|S_i(t)) \times L(\hat{X}_i(t)|\alpha_i).$$

The conditional likelihood $L(\delta_i(t)|S_i(t))$ does not depend on the random effect α_i. It arguably contains all the relevant information about the parameter of interest β, and

thus can be used to construct estimating equations for β [Tsiatis & Davidian, 2001]:

$$\sum_{k=1}^{m}\sum_{i}\delta_i(\tau_k)\big[S_i(\tau_k)-\frac{\sum_{i=1}^{n}S_i(\tau_k)Y_i(\tau_k)e_i}{\sum_i Y_i(\tau_k)e_i}\big], \tag{1.3}$$

where $e_i = exp(\beta S_i(\tau_k) - \beta^2\sigma^2(\tau_k)/2)$. Here we emphasize that the outer summation is taken over the observed event times $\tau_k, k = 1, \ldots, m$.

We now proceed to construct a score test for testing $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$. Let $T = U(0)/\sqrt{Var[U(0)]}$ be the test statistic for such test. The numerator, $U(0)$ is found by evaluating (1.3) when H_0 is true:

$$U(0) = \sum_{k=1}^{m}\sum_{i=1}^{n}\delta_i(t)[\hat{X}_i(\tau_k) - \frac{\sum_i Y_i(\tau_k)\hat{X}_i(\tau_k)}{Y(\tau_k)}]. \tag{1.4}$$

Also note that $U(0)$ has a form familiar to that in (1.2). The difference between the two statistics is that (3.1) estimates the unknown value of the covariate at the observed event time using the available covariate history.

The estimate of variance is obtained by finding a first derivative of (1.3), evaluated for $\beta = 0$:

$$\widehat{Var}[U(0)] = \sum_{k=1}^{m}\frac{n(t)}{Y(t)-1}[\hat{\sigma}^2(t)(Y(t)-n(t)) + (Y(t)-1)\hat{V}(t)]. \tag{1.5}$$

The estimate $\hat{\sigma}^2(t)$ is a product of two quantities. First is the estimate of the unknown variability in measuring $X_i(t)$ at time t, which is estimated by the pooled estimate of the residual sums of squares over all subjects. The second quantity is the estimate of variance of the predicted value $\hat{X}_i(t)$ using available values up to and including time t. Specifically, variance of the predicted value $\hat{X}_i(t)$ at time t is $1/m_i(t)+(t-\bar{t})^2/SS_i(t)$, where $m_i(t)$ is the number of available observations for subject i up to time t, and $SS_i(t)$ is the corresponding sum of squared differences from the mean, using values up to and including time t. As before, $Y(t)$ is the total number of subjects at risk at time t, and $n(t)$ is the number of events at time t. $\hat{V}(t)$ is the sample variance of the ordinary least square estimates, $\hat{X}_i(t)$, among subjects at risk at t. The resulting score test closely resembles the score test from the Cox's proportional hazards model for

non-time-varying covariates or for completely known time-varying covariate histories. Our proposed test can also be viewed as a test of form $H_0 : \lambda(t|X) = \lambda(t),$ for all X, where $\lambda(t)$ is a hazard function.

1.2.4 Calculation of the significance levels

The statistic in (3.1) is best suited for examining a single covariate, whereas we need to test thousands of genes (covariates) to determine which gene's changes over time are associated with a clinical event. The testing procedure controls for the number of false positive findings, as this is a standard approach for the genomewide studies. Simply put, to determine whether the calculated test statistic T_g is significant or not, we want to consider the number of falsely positive findings (FP) among the total number called significant (TP) when that test statistic is used as a cutoff value,

$$\frac{\text{number of false positive findings for T}_\text{g}}{\text{number of total positive findings for T}_\text{g}} = \frac{FP(T_g)}{TP(T_g)}.$$

The expected value of this ratio is defined as the False Discovery Rate (FDR) for statistic T_g (Storey & Tibshirani, 2003),

$$FDR(T_g) = E\left[\frac{FP(T_g)}{TP(T_g)}\right] \approx \frac{E[FP(T_g)]}{E[TP(T_g)]}.$$

One simple way to obtain an estimate of $FDR(T_g)$ is to directly estimate the numerator and the denominator [Xie *et al.*, 2005]. We call this estimator the *False Positive Ratio* (FPR) in order to emphasize it's derivation. To estimate the denominator, we use the total number of the test statistics called significant when T_g is used as a cut-off value, i.e., $\#(|T| > |T_g|)$. The numerator is estimated using permutation-based estimate of the null distribution of the test statistics. Given the nature of our longitudinal data with survival endpoints, the actual permutation needs to be clearly defined. At each observed event time, we permute the event indicators among subjects at risk at that time. In other words, the number of subjects with events is kept fixed at each event time, with their event indicators randomly exchanged among those currently at risk.

Let $T = [T_1, \ldots, T_p]$ be test statistics calculated on each of the p-genes in the original data. Here, $I(\cdot)$ is the usual indicator function, where $I(a) = 1$, if a true.

1. At each observed event time k, permute indicators of events among those subjects still at risk at that time. This is equivalent to choosing $n(t)$ elements out of $Y(t)$, at time t;

2. Using such perturbed data, calculate a set of p test statistics, $T^* = [T_1^*, \ldots, T_p^*]$;

3. Compare each original T_g with all permutation-based T^* and call the number of *false positives* the number among T^* that are greater than T_g,

$$\widehat{FP}(T_g) = \sum_{T^*=[T_1^*,\ldots,T_p^*]} I(|T^*| > |T_g|);$$

4. Repeat steps 1-3 many times, say, hundred times. For each gene g, $g \in \{1, \ldots p\}$, this produces a sequence of hundred numbers. Denote by $\overline{\widehat{FP}}(T_g)$ the mean value of such sequence for test statistic T_g;

5. For each gene, the estimated proportion of false positives is the ratio of $\overline{\widehat{FP}}(T_g)$ over the total number of statistics called significant when T_g is used as a cut-off value. Thus, the estimate of the false positive ratio (FPR) for T_g is:

$$\widehat{FPR}(T_g) = \frac{\overline{\widehat{FP}}(T_g)}{\sum_{T=[T_1,\ldots,T_p]} I(|T| > |T_g|)}.$$

If a test statistic has an estimated proportion of false positives below a desired, pre-specified level, say 10%, then the hypothesis is rejected and the observed test statistic is declared statistically significant. Our testing procedure is similar to the approach proposed by [Storey & Tibshirani, 2003], when the estimated proportion of null hypotheses $\hat{\pi}_0$ is set to 1, and the results are described in terms of the test statistic (rather than the appropriately defined p-value). The presented algorithm can also be viewed as a form of the Empirical Bayes calculation of the FDR [Efron *et al.*, 2001].

We calculate the permutation distribution by permuting the event indicators among subjects at risk at each time. Alternatively, the true permutation distribution for the test statistics would ostensibly be found by permuting the event times among the patients. The problem with this approach is that no samples were collected after the event for each patient, and if they were, the gene expression values after the event may have been affected by the event which would preclude their use. One way to fill-in such missing data, would be to define a distance metric in order to select among subjects with complete covariate series those that are 'close' or 'similar' to the subject with the missing observation. The algorithm will then proceed as follows: a) cluster subjects according to their microarray at time $t-1$, b) note the cluster membership of the subject with the missing t array, and c) impute the missing array by calculating some sample measure (e.g. mean array) using the remaining members of the cluster and their expressions at time t. The testing algorithm can then continue with the Step 2 above. One potential limitation with this approach is that the small size of a cluster of subjects determined to be 'close' or 'similar' to the subject with the missing observation may introduce bias when calculating the imputed value.

1.3 Simulations

We performed a series of simulations to assess validity and performance of our proposed method. The following describes an algorithm to generate longitudinal expressions along with survival outcomes that emulate the data-generating mechanisms presented by the actual problem.

1. Obtain an estimate of a variance-covariance matrix, $\hat{\mathbf{\Sigma}}$, of the random effects by fitting a random effects model to a selection of genes from the actual data;

2. Sample from a bivariate normal distribution with zero mean and variance-covariance matrix obtained in Step 1 to get a set of random effects (intercepts and slopes): $(a_{0i}, a_{1i}) \sim N_2(\mathbf{0}, \hat{\mathbf{\Sigma}})$, for $i = 1, \ldots, n$;

3. Generate individual gene trajectories for each subject and for all genes, using the generated random effects: $X_{ij} = a_{0i} + a_{1i}t_j + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$;

4. Choose a value of the association parameter b. Assuming an exponential hazard function $\lambda(t) = \lambda_0\, e^{bX(t)}$, the parameter b captures the strength of association between time to an event and the time-dependent covariate trajectory $X(t)$. Using the exponential form of the hazard function $\lambda(t)$, and an estimate of an underlying event hazard, λ_0, which uses the number of observed events, we derive an inverse of the cumulative hazard distribution:

$$\Lambda^{-1}(u) = \frac{1}{ba_1} log[\frac{-log(u)\, ba_1}{\hat{\lambda}_0\, e^{ba_0}} + 1];$$

5. By knowing the form of the inverse cumulative hazard function, we can use the Probability Integral Transformation to sample survival times, T; We first generate replicates of the uniformly distributed random variable $U \sim Unif(0, 1)$, then generate survival times as $T = \Lambda^{-1}[-log(U)]$.

In the final step of the algorithm, generated survival times that exceed 28 days are considered censored. The produced survival times are 'linked' to the trajectories generated in Step 3 through the random effects α. Namely, since the same random effects generated in Steps 1-2 are used in Step 5 to generate survival times T, subjects with comparable random effects get assigned similar event times (e.g., early or late).

Using the above algorithm, we generated 600 samples of data. Each sample consists of 100 subjects with 500 longitudinal gene expressions over 7 time-points. Fifty out of 500 genes were set to be significantly associated with the time to an event. Testing was done for three choices of the association parameter b, as well as two values for the measurement error, σ_ϵ^2. Within each simulation, the false positive ratio of 10% was used as a cutoff value for determining significance. Simulations were executed using R statistical software and results presented in Table 1.1.

Table 1.1: Simulation results
$n = 100$ subjects; $p = 500$ genes; 600 replications

median (IQR)	b= 2.5	b= 1.5	b= 0 prop.positive
σ^2_ϵ= 0.10			
# positive	56 (53, 58)	55 (53, 57)	1(0,1)
prop. false +	0.107 (0.056, 0.136)	0.090 (0.057, 0.137)	–
σ^2_ϵ= 0.20			
# positive	55 (52.75, 57)	53 (48, 55)	1(0,2)
prop. false +	0.092 (0.056, 0.137)	0.090 (0.063, 0.125)	–

To understand these results, let us examine the case when the association parameter b is set to 2.5, the measurement error of individual genes is $\sigma^2_\epsilon = 0.20$, and 50 out of a total of 500 simulated genes have trajectories associated with the time to an event (i.e., 10% of genes are significant). A median number of genes found positive over 600 simulations is 55 genes. The median false positive proportion over 600 simulations is 0.092, with an interquartile range of $(0.056, 0.137)$. Similar results are found for the remainder of the cells. When $b = 0$, all genes are expected to be non-significant under $H_0 : \beta = 0$. If the association parameter b is set to zero, any significant genes should be found purely by chance, and we would expect the total number of significant genes to be zero.

Inspection of the simulation results shows that our method performs reasonably well. The proportion of false positive findings remains close to the pre-specified 10% mark for both choices of the association parameter b and the two levels of measurement error. Also note that the proportion of false positives remains similar across the two columns. A change of the pre-specified association parameter b does not dramatically change the proportion of false positive. However, the assumed level of the measurement error seems to influence the estimated number of false positive findings. Estimates of false positive proportions are higher for $\sigma^2_\epsilon = 0.10$.

1.4 Results of Trauma Data Analysis

We apply our method to the data from Inflammation and Host Response to Injury research project (the Glue study). This collaborative program examines the biological reasons why patients can have dramatically different clinical outcomes after experiencing a traumatic injury. Among many scientific questions posed by the Glue investigators is whether we can identify genes whose temporal changes relate to the time until a specific clinical event. It is reasonable to assume that genes exhibiting greater variation are more likely to be associated with the time to an event. Patients in the Glue study are followed for 28 days from the time they experience a serious traumatic injury. This is an example of a right-censored data, which we assume for the developments of our method. Genomic data collected on days 0, 1, 4, 7, 14, 21, and 28 are generated using commercially available oligonucleotide array technology [Affymetrix Inc., 2001]. Each microarray includes expressions on 54,674 probe sets (which we will call 'genes' for the purposes of our analysis). Gene expressions were extracted from oligonucleotide probes by employing a *PM-only* analysis of [Li & Wong, 2001] and normalized across arrays to achieve comparable levels, using the 'Invariant Set' method in the dChip software [Li & Wong, 2003]. Finally, the gene expression values were log-transformed prior to any calculations.

To reduce the overwhelming dimensionality of a microarray, we first excluded those genes labelled 'Absent' over all arrays by the Affymetrix software. We then performed a simple filtering of genes and included only those genes whose estimated coefficient of variation (CV) exceeds a certain threshold. While more complex filtering can be used, one can also proceed without filtering at all. Our choice of threshold is somewhat ad-hoc as we aimed at having a couple of thousands of genes to work with, instead of over fifty four thousand, in this hypothesis-generating approach. This brought the number of genes to under four thousand ($p = 3,914$). Data on 56 subjects with complete entries were included in the analysis. For the purposes of

survival analysis, we define the event of interest as "respiratory recovery" which occurs when a patient no longer depends on a machine to breathe. The response is thus defined as the time from injury (and entry into the ICU) to getting off the ventilator. Of 56 patients, 52 experienced recovery of respiratory function prior to the end of 28-day follow-up. Four patients remained on the ventilator by day 28 and thus had recovery time censored at 28 days.

The test statistic that measures the association between each gene and the time to an event is calculated for each gene separately. We performed described permutation-based testing procedure to determine significance of each test statistic. Of 3,914 investigated genes, 154 were identified as statistically significant when we used $.10$ as a cut-off value for determining significance of each individual gene. As a comparison, a total of 694 genes were identified as significantly associated with the time to a recovery when their test statistics are compared to the $10th$-percentiles of the normal distribution.

A sample gene ontology for a selection of genes for which change in expression over time is associated with the time to respiratory recovery is given in Table 1.2. The sample genes are grouped in those that exhibited positive association with the time to respiratory response, and those with a negative association. For example, increased expression for NM_153701 (interleukin 12 receptor) is associated with the shorter time to recovery.

Results for a sample of four significant genes are presented in Figure 1, with plots of individual patient trajectories over time. The four panels are identified by accession numbers. In order to further illustrate our results, patients are distinguished by whether they had a 'late' respiratory recovery (those occurring after day 16). Red lines represent a patients that either experienced a recovery after day 16 or their times were censored at day 28. The two plots on the right-hand side are for genes negatively related to the time to a recovery. A decrease in gene expression on these two genes is related to a shorter time to recovery. The opposite is true for the other

Table 1.2: Description of a subset of significant genes

Name ID	Function or biological process (positive association)
NM_153701	lack of expression related to immunodeficiency
AF324888	regulation of muscle contraction; signal transduction
NM_002800	fatty acid biosynthesis and oxidation
NM_000593	oligopeptide transport; immune response; protein transport
Name ID	**Function or biological process (negative association)**
NM_002668	has a role in chemotactic processes via CCR1
NM_004994	linked to increased invasiveness of cancer cells
NM_001629	associated with myocardial infarction and stroke

two plots where an increase in gene expression relates to a shorter time to recovery.

The tens of thousands of gene expressions measured repeatedly over time on tens or hundreds of patients can create a considerable computational difficulty due to the enormity of the resulting data sets. To further help reduce the time needed for lengthy computations, we collaborated with an application specialist at the Massachusetts General Hospital Biostatistics Unit in order to obtain, develop, test, and employ a parallel computing system [Lazar & Schoenfeld, 2004]. The 30+ node computer cluster helped us greatly reduce the time needed for lengthy computations.

1.5 Discussion

We provide a method for the survival analysis of longitudinally collected microarrays. We address the issues of intermittently collected microarray data as well as the unknown longitudinal behavior of a single gene expression. A limitation of our approach is that the one-dimensional construction of the test statistic does not necessarily address the high-dimensionality of the problem as it does not use information across all genes simultaneously.

More complex longitudinal models can easily be incorporated into our approach. A

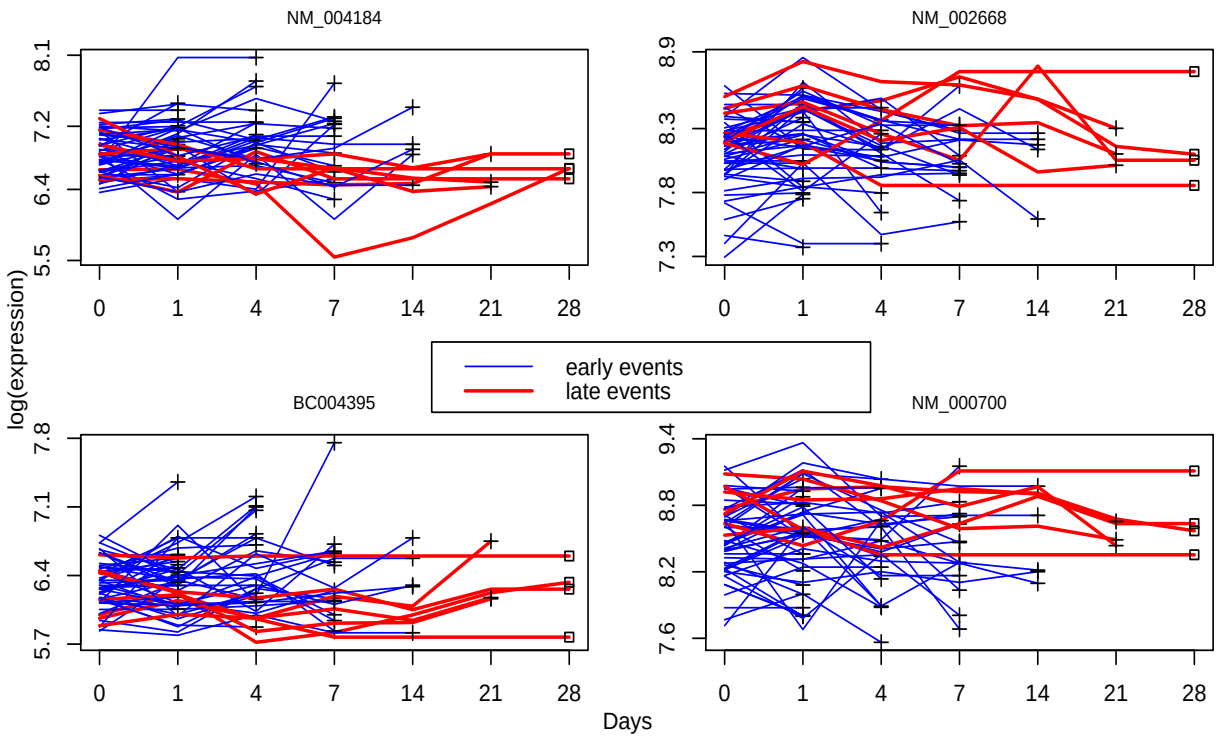

Figure 1.1: Plot of four selected genes

Solid lines represent patients with recovery events that occurred after day 16, or patients censored at day 28. The test statistics corresponding to the two plots to the right are negative, indicating an inverse relationship between gene expressions and the time to a recovery. Conversely, the test statistics corresponding to the two plots to the left are positive, an increase in gene expression is associated with a shorter time to recovery.

minimum set of assumptions regarding the functional relationship between longitudinal gene expression and timing of the events will depend on an individual biological problem at hand. For example, a natural extension would be to implement the approach of [Song *et al.*, 2002], which requires only the assumption that the random effects have a smooth density. Another modification, which may be relevant in some applications, is to devise a multiple imputation procedure for the unknown covariate values. While this will certainly add to the overall computational complexity, it would be interesting to explore whether it can be incorporated so to take advantage of the computations already in place and the high-dimensionality of the data. Finally, in order to make the proposed test statistics more robust to potential outliers, the actual values of gene expressions may have be replaced by ranks or some function of the ranks.

In the Glue study, patients were closely monitored at all times for a period of 28 days. The study subjects either experience an event or their time is censored at the end of the study follow-up. This is an example of Type I censoring where there is no possibility of missing data due to a dropout. However, in a typical clinical trial where study participants are followed for a longer period of time, it is likely that the censoring due to early dropout would be an issue. It is straightforward to accommodate this type of censoring in our approach.

The continual advancement of the microarray technology will ultimately result in many large studies routinely including longitudinal genomic observations as part of the study follow-up. The longitudinal microarray and event time data will thus become more common. Also, other applications of high-dimensional data such as proteomics and metabolomics data will arise. Therefore, further development of efficient methodologies to handle these high-dimensional event time data sets will be needed.

Acknowledgement

Supported by a Large-Scale Collaborative Project Award (U54-GM62119) from The National Institute of General Medical Sciences, National Institutes of Health (NIH), and NIH R01 Grant # CA47048.

We wish to thank Dr. Cheng Li for valuable insights on the subject. We also extend our thanks to Drs. John Storey and John Quackenbush for their helpful suggestions and review. The application specialist, Peter Lazar, engineered and maintained the parallel system in R used in our analysis. Finally, we are especially grateful to the four reviewers whose careful review and constructive criticism greatly improved the quality of the manuscript.

The following is a list of participating investigators of the Inflammation and the Host Response to Injury research project: Brett D. Arnoldo, M.D., Henry V. Baker, Ph.D., Paul Bankey, M.D., Timothy R. Billiar, M.D., Bernard H. Brownstein, Ph.D., Steve E. Calvano, Ph.D., David G. Camp II, Ph.D., Celeste Campbell-Finnerty, Ph.D., George Casella, Ph.D., Irshad H. Chaudry, Ph.D., J. Perren Cobb, M.D., Ronald W. Davis, Ph.D., Asit K. De, Ph.D., Constance Elson, Ph.D., Bradley Freeman, M.D., Richard L. Gamelli, M.D., Nicole S. Gibran, M.D., Brian G. Harbrecht, M.D., Douglas L. Hayden, M.A., Laura Hennessy, R.N,. David N. Herndon, M.D., Jureta W. Horton, Ph.D., Marc G. Jeschke, M.D., Ph.D., Jeffrey Johnson, M.D., Matthew B. Klein, M.D., James A. Lederer, Ph.D., Stephen F. Lowry, M.D., Ronald V. Maier, M.D., John A. Mannick, M.D., Philip H. Mason, Ph.D., Grace P. McDonald-Smith, M.Ed., Carol L. Miller-Graziano, Ph.D., Michael N. Mindrinos, Ph.D., Joseph P. Minei, M.D., Lyle L. Moldawer, Ph.D., Ernest E. Moore, M.D., Avery B. Nathens, M.D., Ph.D., M.P.H., Grant E. O'Keefe, M.D., M.P.H., Laurence G. Rahme, Ph.D., Daniel G. Remick, Jr. M.D., David A. Schoenfeld, Ph.D., Michael B. Shapiro, M.D., Geoffrey M. Silver, M.D., Richard D. Smith, Ph.D., John Storey, Ph.D., Ronald G. Tompkins, M.D., Sc.D., Mehmet Toner, Ph.D., H. Shaw Warren, M.D., Michael A. West, M.D., Wenzhong Xiao, Ph.D.

Bibliography

[Affymetrix Inc., 2001] Affymetrix, Inc. 2001. Microarray Suite 5.0. Santa Clara, CA.

[Alizadeh *et al.*, 2000] Alizadeh AA, Eisen MB, Davis RE, et al. 2000. Distinct types of diffuse large Bcell lymphoma identified by gene expression profiling. *Nature* 403, 503-11.

[Bair & Tibshirani, 2004] Bair, E., Tibshirani R. 2004. Semi-supervised methods for predicting patient survival from gene expression papers. *PLoS Biology* 2, 5011-22.

[Cox, 1972] Cox DR. 1972. Regression models and life tables. *J R Stat Soc* B34, 187-20.

[Davidov & Zelen, 1998] Davidov O, Zelen M. 1998. Urn sampling and the proportional hazard model. *Lifetime Data Analysis* 4, 309-27.

[Efron *et al.*, 2001] Efron B, Tibshirani R, Storey JD, and Tusher V. 2001. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 96, 1151-60.

[Golub *et al.*, 1999] Golub TR, Slonim DK, Tamayo P, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-37.

[Gui & Li, 2004] Gui L, Li H. 2005. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001-08.

[Jones & Crowley, 1989] Jones M, Crowley J. 1989. A general class of nonparametric tests for survival analysis. *Biometrics* 45, 15770.

[Jung *et al.*, 2005] Jung SH, Owzar K, George SL. 2005. A multiple testing procedure to associate gene expression levels with survival. *Stat Med*, 24(20):3077-88.

[Lazar & Schoenfeld, 2004] Lazar P & Schoenfeld D. 2004. Self-contained parallel system for R. MGH Biostatistics, Boston, MA. http://cran.r-project.org/src/contrib/Descriptions/biopara.html

[Li & Luan, 2003] Li H, Luan Y. Kernel Cox regression models for linking gene expression profiles to censored survival data. 2003. University of California, Davis, Proc Pac Symp Biocomp, 65-76.

[Li & Gui, 2004] Li H, Gui L. 2004. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20(Suppl. 1).

[Li & Wong, 2001] Li C., Wong WH. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci*, 98(1), 31-36.

[Li & Wong, 2003] Li C, Wong WH. 2003. DNA-Chip Analyzer (dChip). In The analysis of gene expression data: methods and software. Edited by Parmigiani G, Garrett ES, Irizarry R, and Zeger SL. New York: Springer, p. 120.41.

[Luan & Li, 2002] Luan Y, Li H. 2002. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474-82.

[Makretsov *et al.*, 2004] Makretsov NA, Huntsman DG, Nielsen TO, et al. 2004. Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. *Clin Cancer Res*, 10: 6143-51.

[Nguyen & Rocke, 2002] Nguyen D, Rocke D. 2002. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics* 18(12): 1625-32.

[Park *et al.*, 2002] Park P, Tian L, Kohane I. Linking gene expression data with patient survival times using partial least squares. 2002. Bioinformatics, 18(1):S120-27.

[R, 2005] R Development Core Team. R: A language and environment for statistical computing. 2005. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

[Ring & Ross, 2002] Ring BZ, Ross DT. 2002. Microarrays and molecular markers for tumor classification. *Genome Biology*, 3:comment 2005.1-6.

[Rosenwald *et al.*, 2002] Rosenwald A, Wright G, Chan WC, *et al.* 2002. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, 346:1937-47.

[Song *et al.*, 2002] Song X, Tsiatis AA, Davidian M. 2002.A Semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58:742-53.

[Sorlie *et al.*, 2001] Sorlie T, Perou CM, Tibshirani R *et al.* 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci*, 98(19):10869-74.

[Stefanski & Carroll, 1987] Stefanski LA, Carroll RJ. 1987. Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika* 74, 703-16.

[Storey & Tibshirani, 2003] Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci*, 100(16):9440-45.

[Storey *et al.*, 2005] Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. 2005 Significance analysis of time course microarray experiments. *Proc Natl Acad Sci*, 102(36):12837-42.

[Tsiatis & Davidian, 2001] Tsiatis AA, Davidian M. 2001. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2):447-458.

[Tusher *et al.*, 2001] Tusher V, Tibshirani R, Chow G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci*, 98:5116-21.

[van't Veer *et al.*, 2002] van't Veer LJ, Dai H, van de Vijver MJ, et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530-36.

[van de Vijver *et al.*, 2002] van de Vijver MJ, He YD, van 't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347:1999-2009.

[Westfall & Young, 1989] Westfall P, Young S. 1989. p-Value adjustments for multiple tests in multivariate binomial models. *J Am Stat Assoc*, 84(407):780-86.

[Xie *et al.*, 2005] Xie Y, Pan W, Khodursky AB. 2005. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21(23):4280-4288.

[Yeung *et al.*, 2003] Yeung KY, Medvedovic M, Bumgarner RE. 2003. Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(5):R34.