Statistical methods for microarray data

Goeman, Jelle Jurjen

CHAPTER 3

Testing Association of a Pathway with Survival

Abstract

A recent surge of interest in survival as the primary clinical endpoint of microarray studies has called for an extension of the Global Test methodology (Goeman et al., 2004) to survival. We present a score test for association of the expression profile of one or more groups of genes with a (possibly censored) survival time. Groups of genes may be pathways, areas of the genome, clusters from a cluster analysis or all genes on a chip. The test allows one to test hypotheses about the influence of these groups of genes on survival directly, without the intermediary of single gene testing. The test is based on the Cox proportional hazards model and is calculated using martingale residuals. It is possible to adjust the test for the presence of covariates. We also present a diagnostic graph to assist in the interpretation of the test result, visualizing the influence of genes. The test is applied to a tumour data set, revealing pathways from the Gene Ontology database that are associated with survival of patients. The global test for survival has been incorporated into the R-package globaltest (from version 3.0), available from http://www.bioconductor.org.

3.1 Introduction

A microarray experiment typically results in many thousands of measurements, each relating to the expression level of a single gene. Single genes, however, are often not the primary theoretical focus of the researcher, who might be

This is a pre-copy-editing, author-produced version of an article accepted for publication in *Bioinformatics* following peer review. The definitive publisher-authenticated version : J. J. Goeman, J. Oosting, A. M. Cleton-Jansen, J. Anninga, and J. C. van Houwelingen (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics 21*(9), 1950–1957 is available online at: `http://dx.doi.org/10.1093/bioinformatics/bti267`

more interested in certain pathways or genomic regions that are suspected to be biologically relevant.

For this reason we have introduced the Global Test for groups of genes (Goeman et al., 2004), which allows the unit of analysis of the microarray experiment to be shifted from the single gene level to the pathway level, where a "pathway" may be any set of genes, e.g. chosen using the *Gene Ontology* database or from earlier experiments. For every pathway, the Global Test can test (with a single test) whether the expression profile of that pathway is significantly associated with a clinical variable of interest. This allows researchers immediately to test theoretical hypotheses on the clinical importance of certain pathways. Even when such hypotheses are not directly available from biological theory or past research, the Global Test can significantly reduce the multiple testing problem, because there are typically much fewer pathways than genes.

In the original publication of the Global Test, the clinical variable could be either a continuous measurement or a 0/1 group indicator. Recently, however, there has been a surge of interest in survival time of patients as the primary clinical outcome in a microarray experiment. Many studies focus on prediction of survival, e.g. in breast cancer Van 't Veer et al. (2002), Van de Vijver et al. (2002) and Pawitan et al. (2004) and in lung cancer Wigle et al. (2002) and Beer et al. (2002). Other studies use multiple testing methods to find genes which are associated with survival (Jenssen et al., 2002).

The present paper extends the Global Test methodology to survival outcomes. It allows the researcher to test whether the expression profile of a given set of genes is associated with survival. More precisely, it tests whether individuals with a similar gene expression profile tend to have similar survival times. A significant pathway may be a mix of genes which are upregulated for patients with short survival time, genes which are downregulated for the same patients, and other genes that show no association with survival at all.

The test of the present paper is based on the Cox proportional hazards model. Therefore, it avoids the requirement of many analysis strategies to choose an arbitrary cut-off (e.g. five years survival), but uses all survival information that is present in the data. Technically, the test is derived from the goodness-of-fit test of the Cox model by Verweij et al. (1998). The original Global Test was derived in a similar way from a goodness-of-fit test for generalized linear models (Le Cessie and van Houwelingen, 1995). The two Global Tests are therefore highly comparable and allow quite similar interpretations.

In this paper we also show how the test can be adjusted for the presence of covariates (possible confounders or competing risk factors). This allows better use of the Global Test in observational studies. Furthermore, it allows the researcher to establish that the microarray really adds something to the predictive performance of known risk factors, showing that it is not simply an expensive way to measure risk factors already known. It also allows the test to be used on more complex designs than a simple one-sample follow up study. The approach will be illustrated on a data set of 17 osteosarcoma patients, testing pathways from the Gene Ontology database.

The new Global Test method presented in this paper has been incorporated into the R-package *globaltest*, version 3.0, which is available from BioConductor (Gentleman et al., 2004, www.bioconductor.org).

3.2 The model

The Global Test exploits the duality between association and prediction. By definition, if two things are associated, knowing one improves prediction of the other. Hence, if survival is associated with gene expression profile, this means that knowing the gene expression profile allows a better prediction of survival than not knowing the expression profile.

With this idea in mind we make a prediction model for prediction of survival from the gene expression measurements. The most convenient choice for such a model is the Cox proportional hazards model, which is the most widely used model for survival data in medical research. The Cox model uses the full empirical distribution of the survival times and it can handle censored data, i.e. samples for which the exact survival time is not known, but for which it is only known that the patient is still alive at a certain moment (Klein and Moeschberger, 1997). The use of the Cox model requires a true follow-up study design, meaning that patients were not selected on their survival times in any way. If such a patient selection was made, the methods of this paper may not be appropriate: in Van 't Veer et al. (2002), for example, where a selected group of early metastases was compared to a selected group which was at least five years metastasis-free, the original Global Test for a 0/1 outcome is preferable (Goeman et al., 2004).

Suppose the matrix of normalized gene expression measurements for the group of genes of interest is given by the $n \times m$ matrix X with elements x_{ij}, where n is the sample size and m the number of genes in the group. Suppose also that there is a number $p \geq 0$ of covariates for each patient, which we put in an $n \times p$ data matrix Z with elements z_{ij}. It will be assumed that $p < n$, but no such restriction is put on m.

Cox's proportional hazards model (Klein and Moeschberger, 1997, chapter 8) assumes the hazard function at time t for individual i to relate to the covariates as

$$h_i(t) = h(t)e^{c_i + r_i}, \tag{3.1}$$

where $h(t)$ is an unknown baseline hazard function and $c_i + r_i$ is a linear function of the covariates, which is split up in our case into $r_i = \sum_{k=1}^{m} \beta_k x_{ik}$, relating to the gene expressions, and $c_i = \sum_{l=1}^{p} \gamma_l z_{il}$, relating to the covariates. The hazard function determines the survival function $S_i(t)$, which gives the probability that individual i survives up to time t, through

$$S_i(t) = e^{-H_i(t)},$$

where $H_i(t) = \int_0^t h_i(s)\,\mathrm{d}s$ is the cumulative hazard up to time t.

In this model showing that the gene expressions are associated with survival is equivalent to rejecting the null hypothesis

$$H_0 : \beta_1 = \ldots = \beta_m = 0,$$

that all regression coefficients relating to the gene expressions are zero. If m were always small, we could test H_0 using classical tests which were developed for the Cox model. These tests do not work for general m, however (for an overview of these classical tests see Klein and Moeschberger, 1997, section 8.2).

To obtain a test that works whatever the value of m, we put an extra assumption on the regression coefficients $\beta_1, \ldots, \beta_m$. We assume that the regression coefficients of the genes are random and a priori independent with mean zero and common variance τ^2. The null hypothesis now becomes simply

$$H_0 : \tau^2 = 0,$$

so that the dimension of H_0 does not depend on m anymore. Note that the coefficients $\gamma_1, \ldots, \gamma_p$ of the covariates are not assumed to be random.

The Cox model with random coefficients is an empirical Bayesian model and is closely linked to penalized likelihood methods. It should be noted that we have not assumed a specific distributional form for the regression coefficients; the derivation of our test is invariant to the choice of the shape of this distribution. Choosing a Gaussian distribution results in a Cox ridge regression model (Pawitan et al., 2004); choosing a double exponential distribution results in a LASSO model (Tibshirani, 1997). Both models can also be used to predict survival times of patients.

In the context of testing it is most insightful to view the prior distribution of the regression coefficients as the *focus of the power* of the test. The test that will be derived in the next section will be a score test, which has the property that it has optimal power against alternatives with small values of the parameter τ^2. This property stems from the fact that the score test is equivalent to the likelihood ratio test in the limit where the alternative $\tau^2 \to 0$ (Cox and Hinkley, 1974). Alternatives with small values of τ^2 tend to have small values of $\sum \beta_i^2$, so that the test can be said to be optimal on average against alternatives with small values of $\sum \beta_i^2$. These alternatives are mainly alternatives which have all or most regression coefficients non-zero but small. The test can therefore be said to be optimized against alternatives in which all or most genes have some association with the outcome. This alternative is precisely the situation in which we are interested, because we want to say something about the pathway as a whole.

Alternative tests can easily be derived for regression coefficients with a more complex covariance structure. If the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ is assumed a priori to have mean zero and covariance matrix $\tau^2\Sigma$, the resulting test of H_0 would be optimal against alternative with small values of $\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}$. The standard choice of $\Sigma = I_m$ distributes power equally over all directions of $\boldsymbol{\beta}$, while a different choice will have more power against deviations from H_0 in directions which correspond to the larger eigenvalues of Σ. This property could be exploited in the derivation of a test for a specific purpose or to incorporate prior knowledge. In this paper we shall restrict ourselves to $\Sigma = I_m$.

3.3 Derivation of the test

Testing association of a group of genes with survival can therefore be done by testing H_0 in the empirical Bayesian model (3.1) with random regression coefficients. In this section we will derive the test statistic for this test. A score test for the same model has also been studied by Verweij et al. (1998) in the context of testing the fit of the Cox model. Their derivation was based on the partial likelihood of the Cox model. In this paper we give an alternative derivation based of the full likelihood and a simpler martingale argument.

We derive the test in stages. First suppose that all parameters except τ^2 are known, i.e. the regression coefficients $\gamma_1, \ldots, \gamma_p$ and the baseline hazard function $h(t)$ are known. In this simplified situation it will be relatively easy to derive the score test, which can be generalized to the situation with unknown parameters later in this section.

The basic score test

By definition a score test is based on the derivative of the log-likelihood at the value of the parameter to be tested. Suppose for each individual i we have observed a survival time t_i and a status indicator d_i, where $d_i = 1$ indicates death (the patient died at t_i) and $d_i = 0$ censoring (the patient was lost to follow-up at t_i). The loglikelihood of τ^2 in the model (3.1) is

$$L(\tau^2) = \log\big\{\mathrm{E}_{\mathbf{r}}\big[\exp\big(\sum_{i=1}^{n} f_i(r_i)\big)\big]\big\}, \tag{3.2}$$

where

$$f_i(r_i) = d_i[\log\{h(t_i)\} + c_i + r_i] - H(t_i)e^{c_i+r_i}$$

is the contribution to the loglikelihood of individual i for fixed r_i, and $H(t) = \int_0^t h(s)\,\mathrm{d}s$ is the cumulative baseline hazard.

From the assumptions on the distribution of $\beta_1, \ldots, \beta_m$ we can derive the distribution of $\mathbf{r} = (r_1, \ldots, r_n)'$, the vector of the linear effects of the gene expressions. This $\mathbf{r}$ has mean zero and covariance matrix $\tau^2 R$, where $R = XX'$. For the general likelihood (3.2) and an $\mathbf{r}$ of this form, Le Cessie and van Houwelingen (1995) have used a Taylor approximation to derive that

$$\frac{\partial L(0)}{\partial \tau^2} = \frac{1}{2}\left(\sum_i R_{ii} \frac{\partial^2 f_i(0)}{(\partial r_i)^2} + \sum_{i,j} R_{ij} \frac{\partial f_i(0)}{\partial r_i} \frac{\partial f_j(0)}{\partial r_j} \right).$$

For the Cox model this becomes

$$\frac{\partial L(0)}{\partial \tau^2} = \frac{1}{2}\left(\sum_{i,j} R_{ij}(d_i - u_i)(d_j - u_j) - \sum_i R_{ii} u_i \right), \qquad (3.3)$$

where $u_i = e^{c_i} H(t_i)$, $i = 1, \ldots, n$, is the hazard incurred by individual i up to time t_i. Note that $d_i - u_i$ is the martingale residual of individual i at time t_i (Klein and Moeschberger, 1997, section 11.3).

For known $H(t)$ and known $c_1, \ldots, c_n$, the expression (3.3) can be standardized to have unit variance and used as the score test statistic. When these parameters are unknown, we must plug in maximum likelihood estimates for them under the null model in which $\tau^2 = 0$. Standardizing the score test is traditionally done using the Fisher Information, calculated from the second derivatives of the loglikelihood. In this case these calculations are very unpleasant, and it turns out to be simpler to standardize using the estimated variance of the test statistic.

Using estimated baseline hazard

We shall first plug in the estimate for the cumulative hazard $H(t)$, but still assume that $\gamma_1, \ldots, \gamma_p$ and hence $c_1, \ldots, c_n$ are known. As the maximum likelihood estimate of $H(t)$ we can take the Breslow estimator (Klein and Moeschberger, 1997, section 8.6)

$$\hat{H}(t_i) = \sum_{t_j \le t_i} \frac{d_j}{\sum_{t_k \ge t_j} e^{c_k}}, \qquad i = 1, \ldots, n,$$

and write $\hat{u}_i = e^{c_i} \hat{H}(t_i), i = 1, \ldots, n$.

Using twice the estimated derivative of the log-likelihood (3.3) as the test statistic and writing it in matrix notation we get the test statistic

$$T = (\mathbf{d} - \hat{\mathbf{u}})' R (\mathbf{d} - \hat{\mathbf{u}}) - \mathrm{trace}(R\hat{U}) \tag{3.4}$$

where $\mathbf{d} = (d_1, \dots, d_n)'$, $\hat{\mathbf{u}} = (\hat{u}_1, \dots, \hat{u}_n)'$ and $\hat{U} = \mathrm{diag}(\hat{\mathbf{u}})$, an $n \times n$ diagonal matrix with $\hat{U}_{ii} = \hat{u}_i$.

The derivation of estimates for the mean and variance of T is quite technical and will be given in the separate subsection on page 41. The estimated mean is

$$\hat{\mathrm{E}}(T) = -\mathrm{trace}(RPP'), \tag{3.5}$$

where P is an $n \times n$ matrix with i, j-th element

$$p_{ij} = \mathbf{1}_{\{t_i \geq t_j\}} \frac{d_j e^{c_i}}{\sum_k \mathbf{1}_{\{t_k \geq t_j\}} e^{c_k}},$$

where $\mathbf{1}_{\{\cdot\}}$ indicates an indicator function. Each p_{ij} is the increment of the cumulative hazard incurred by individual i at time t_j, so that $\sum_i p_{ij} = d_j$ and $\sum_j p_{ij} = \hat{u}_i$.

The estimated variance of T is

$$\widehat{\mathrm{Var}}(T) = \sum_{j=1}^{n} \mathbf{p}_j{}' \,\mathrm{diag}(\mathbf{t}_j \mathbf{t}_j'), \tag{3.6}$$

where $\mathbf{p}_j$ is the j-th column of P and $\mathbf{t}_j = (I - \mathbf{1}\mathbf{p}_j{}')[\mathrm{diag}(R) + 2R(\mathbf{m}_j - \mathbf{p}_j)]$. The diag of a square matrix is the column vector of its diagonal elements; $\mathbf{1}$ is an n-vector of ones, and $\mathbf{m}_j$ is the j-th column of the matrix $M = (D - P)B$, where $D = \mathrm{diag}(\mathbf{d})$ is a diagonal matrix with $D_{ii} = d_i$ and B is an $n \times n$ matrix with elements $b_{ij} = \mathbf{1}_{\{t_i < t_j\}}$. The elements m_{ij} of M can be interpreted as the estimated martingale residual of individual i just before time t_j.

For purposes of interpretation it is often easier to take

$$T_0 = (\mathbf{d} - \hat{\mathbf{u}})' R (\mathbf{d} - \hat{\mathbf{u}})$$

as the unstandardized test statistic. It has $\hat{\mathrm{E}} T_0 = \mathrm{trace}(R\hat{U} - PP')$ and $\widehat{\mathrm{Var}}(T_0) = \widehat{\mathrm{Var}}(T)$, so that it leads to the same standardized test statistic:

$$Q = \frac{T - \hat{\mathrm{E}}T}{\widehat{\mathrm{Var}}(T)} = \frac{T_0 - \hat{\mathrm{E}}T_0}{\widehat{\mathrm{Var}}(T_0)}.$$

Using estimated regression coefficients

In general the regression coefficients $\gamma_1, \ldots, \gamma_p$ of the covariates are not known but must be estimated. Replacing $\gamma_1, \ldots, \gamma_p$ by their maximum likelihood estimates will still give a valid score test for H_0, but with a different distribution of the test statistic. We use the following approximation to this distribution which is derived by Verweij et al. (1998).

The estimated martingale residuals $\mathbf{d} - \tilde{\mathbf{u}}$ based on the estimated $\hat{\gamma}_1, \ldots, \hat{\gamma}_p$ can be approximated in a first order Taylor approximation by

$$\mathbf{d} - \tilde{\mathbf{u}} \approx (I - V)(\mathbf{d} - \hat{\mathbf{u}}) \tag{3.7}$$

with $V = WZ(ZWZ')^{-1}Z'$, $W = \hat{U} - PP'$ and Z the $n \times p$ data matrix of the fixed covariates. Therefore the unstandardized test statistic T_0 can be approximated as

$$T_0 \approx (\mathbf{d} - \hat{\mathbf{u}})'\tilde{R}(\mathbf{d} - \hat{\mathbf{u}})$$

with $\tilde{R} = (I - V)'R(I - V)$. The expectation of T_0 can be estimated using the formulae in section 3.3. They are approximately

$$\hat{\mathrm{E}}T_0 \approx \mathrm{trace}(\tilde{R}W)$$

and

$$\widehat{\mathrm{Var}}(T_0) \approx \sum_{j=1}^{n} \mathbf{p}_j{}' \,\mathrm{diag}(\tilde{\mathbf{t}}_j \tilde{\mathbf{t}}_j'),$$

with $\tilde{\mathbf{t}}_j = (I - \mathbf{1}\mathbf{p}_j{}')[\mathrm{diag}(\tilde{R}) + 2\tilde{R}(\mathbf{m}_j - \mathbf{p}_j)]$. To evaluate $\hat{\mathrm{E}}T_0$ and $\widehat{\mathrm{Var}}(T_0)$ we replace the parameter values of $\gamma_1, \ldots, \gamma_p$ by their estimates. Simulations in Verweij et al. (1998) show this approximation to be quite accurate.

The distribution of the test statistic

There are two ways to calculate the p-value of the test: by asymptotic theory and by permutation arguments. We outline both options and their advantages.

In equation (3.3) it will be shown that the centered test statistic $T - \hat{\mathrm{E}}T$ can be written as a linear combination of n martingales. Therefore by the martingale central limit theorem (Andersen et al., 1993) the distribution of the standardized Q converges to a standard normal distribution as $n \to \infty$. This fact motivates the use of a normal approximation to the distribution of Q for calculating the one-sided p-value (see also simulation results by Verweij et al., 1998).

For small samples the asymptotic distribution may not be reliable enough. An alternative is to calculate Q for all, or a random sample of many (10,000), permutations of the martingale residuals of the n samples. This randomly redistributes the vectors of gene expression measurements over the individuals, while keeping the relationship between the fixed covariates and survival the same. The resulting distribution is another approximation to the null distribution of Q, which can be used to find the p-value. Use of the permutation null distribution requires the assumption that there is no relationship between the gene expressions on the one hand and the covariates and the censoring mechanism on the other hand: permuting destroys these associations. This makes the permutation null distribution less useful when covariates are present.

The main advantage of the permutation-based p-value is that it gives an "exact" p-value, which is guaranteed to keep the alpha level, provided enough permutations are used. This is especially useful for smaller sample sizes, where we may not trust the normality of the distribution of Q. The advantage of the asymptotic theory p-value—aside from being much quicker to calculate—is that it has more power: the permutation based p-value does not use the full null distribution, but the null distribution conditional on the set of observed martingale residuals. With this conditioning the test loses some power, as the set of observed residuals is informative on the parameter τ^2.

Counting process calculations

In this technical section we calculate the mean and variance of the test statistic T under the null hypothesis for known $c_1, \ldots, c_n$ but estimated $H(t)$, as given in (3.5) and (3.6). For this we will use a counting process notation (Andersen et al., 1993; Fleming and Harrington, 1991). The strategy we will use is common in martingale theory: we write our test statistic T as the limit of a process $T(t)$ as $t \to \infty$ and decompose $T(t)$ into a martingale and a compensator. The limit of the compensator is the estimator of the mean of T and the limit of the predictable variation process is the estimate of the variance. For an alternative derivation, see Verweij et al. (1998).

Let $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))'$ be the vector of at-risk processes of individuals $1, \ldots, n$ and $\mathbf{N}(t) = (N_1(t), \ldots, N_n(t))'$ the vector of their counting processes. Then $\mathbf{N}$ has intensity process $\mathbf{\Lambda} = C\mathbf{Y}(t)H(t)$, where C is a diagonal matrix with $C_{ii} = e^{c_i}$, $i = 1 \ldots, n$. Write $N(t) = \mathbf{1}'\mathbf{N}(t)$, the total counting process.

In the counting process notation, $\mathbf{d} = \mathbf{N}(\infty)$ and $\hat{\mathbf{u}} = \hat{\mathbf{\Lambda}}(\infty)$ with $\hat{\mathbf{\Lambda}}(t) = \int_0^t \mathbf{V}(s)\mathbf{1}'\, \mathrm{d}\mathbf{N}(s)$, where $\mathbf{V} = C\mathbf{Y}(\mathbf{1}'C\mathbf{Y})^{-1}$. Wherever possible we will drop the dependence on time for convenience of notation.

Note that the compensator of $\hat{\mathbf{\Lambda}}$ is $\mathbf{\Lambda}$, which is also the compensator of $\mathbf{N}$. Write $\widehat{\mathbf{M}} = \mathbf{N} - \hat{\mathbf{\Lambda}}$. Then $\mathbf{d} - \hat{\mathbf{u}} = \widehat{\mathbf{M}}(\infty)$ and $\widehat{\mathbf{M}}(t) = \int_0^t (I - \mathbf{V}\mathbf{1}')\, \mathrm{d}\mathbf{N}$ is a martingale vector. Subtracting the intensities and writing $\mathbf{M} = \mathbf{N} - \mathbf{\Lambda}$,

$$\widehat{\mathbf{M}}(t) = \int_0^t (I_n - \mathbf{Y}\mathbf{1}')\, \mathrm{d}\mathbf{M}.$$

The statistic T is $T(\infty)$, with

$$T(t) = \operatorname{trace}[R\widehat{\mathbf{M}}\widehat{\mathbf{M}}' - R\operatorname{diag}(\hat{\mathbf{\Lambda}})].$$

From the integration by parts formula (Fleming and Harrington, 1991, theorem A.1.2) it follows that, almost surely,

$$\begin{aligned}\widehat{\mathbf{M}}\widehat{\mathbf{M}}' &= \int_0^t \widehat{\mathbf{M}}^- \,\mathrm{d}\widehat{\mathbf{M}}' + \int_0^t \mathrm{d}\widehat{\mathbf{M}}\,(\widehat{\mathbf{M}}^-)' \\ &\quad + \int_0^t (I - \mathbf{1}\mathbf{V}')\operatorname{diag}(\mathrm{d}\mathbf{N})(I - \mathbf{V}\mathbf{1}') \qquad (3.8)\end{aligned}$$

where $\widehat{\mathbf{M}}^-(s) = \widehat{\mathbf{M}}(s-)$ is a predictable process. Using (3.8) and some linear algebra we can say that, almost surely,

$$T(t) = \int_0^t (\operatorname{diag}(R)' + 2(\widehat{\mathbf{M}}^-)'R - \mathbf{V}'R)(I - \mathbf{V}\mathbf{1}')\,\mathrm{d}\mathbf{N} - \int_0^t \mathbf{V}'R\,\mathrm{d}\mathbf{N}.$$

Because $\int_0^t (I - \mathbf{V}\mathbf{1}')\,\mathrm{d}\mathbf{N}$ is a martingale and $\operatorname{diag}(R)' + 2(\widehat{\mathbf{M}}^-)'R - \mathbf{V}'R$ is predictable, the compensator of the process T is $-\int_0^t \mathbf{V}'R\,\mathrm{d}\mathbf{\Lambda}$, which we can estimate by

$$\hat{\mathrm{E}}T = -\int_0^t \mathbf{V}'R\,\mathrm{d}\hat{\mathbf{\Lambda}} = -\int_0^t \mathbf{V}'R\mathbf{V}\mathbf{1}'\,\mathrm{d}\mathbf{N}$$

The process $S = T - \hat{\mathrm{E}}T$ is a martingale. It can be written in the following way

$$S = \int_0^t (\operatorname{diag}(R)' + 2(\widehat{\mathbf{M}}^- - \mathbf{V})'R)(I - \mathbf{V}\mathbf{1}')\,\mathrm{d}\mathbf{M} \qquad (3.9)$$

as the integral of the predictable process vector

$$\mathbf{K} = (\operatorname{diag}(R)' + 2(\widehat{\mathbf{M}}^- - \mathbf{V})'R)(I - \mathbf{V}\mathbf{1}')$$

over the martingale vector $\mathbf{M}$. The predictable variation process of S is therefore $\langle S\rangle = \int_0^t \operatorname{diag}(\mathbf{K}\mathbf{K}')'\,\mathrm{d}\mathbf{\Lambda}$, which we can estimate by

$$\widehat{\operatorname{Var}}(T) = \int_0^t \operatorname{diag}(\mathbf{K}\mathbf{K}')'\,\mathrm{d}\hat{\mathbf{\Lambda}} = \int_0^t \operatorname{diag}(\mathbf{K}\mathbf{K}')'\mathbf{V}\mathbf{1}'\,\mathrm{d}\mathbf{N}$$

To evaluate $\hat{\mathrm{E}}T$ and $\widehat{\operatorname{Var}}(T)$ we use

$$p_{ij} = \int_0^\infty \frac{e^{c_i}Y_i}{Y}\,\mathrm{d}N_j = \mathbf{1}_{\{t_i \geq t_j\}} \frac{e^{c_i}d_j}{\sum_{t_k \geq t_j} e^{c_k}}.$$

and

$$m_{ij} = \int_0^\infty \widehat{M}_i^-\,\mathrm{d}N_j = \mathbf{1}_{\{t_i < t_j\}} d_i - \sum_{k=1}^n \mathbf{1}_{\{t_k < t_j\}} p_{ik}$$

Writing P for the $n \times n$ matrix with elements p_{ij} and M for the $n \times n$ matrix with elements m_{ij}, the results (3.5) and (3.6) follow.

3.4 Interpretation

When testing a specific pathway for a specific sample of patients, it is usually not satisfactory to only report the resulting p-value. In this section we will discuss some issues related to interpretation of the test result. We show how to calculate and visualize the influence of individual genes on the test result. We also propose an diagnostic which can be used when many genes are associated with survival, to assess whether a gene group is exceptional. We only give the theory here; for an example see section 3.5.

Similarity

The test of this paper is derived from the Cox model in the same way as the Global Test in Goeman et al. (2004) was derived from the generalized linear model. The functional form of the test statistic is therefore quite similar, the martingale residuals taking the place of the residuals from the generalized linear model in that paper. Much of the interpretation of the test statistic is also quite similar.

Central to all interpretation of the test outcome is the matrix $R = XX'$ which figures prominently in the formula for the test statistic. It is an $n \times n$ matrix which can be seen as describing the similarities in expression profile between the samples. The entry R_{ij} is relatively large if samples i and j have a relatively similar expression profile over the pathway of interest.

To show the role of the matrix R, we can rewrite the unstandardized test statistic T_0 as

$$T_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij}(d_i - \hat{u}_i)(d_j - \hat{u}_j),$$

which is the sum over the term-by-term product of the entries of R and the entries of the matrix $(\mathbf{d} - \hat{\mathbf{u}})(\mathbf{d} - \hat{\mathbf{u}})'$. The i, j-th entry of the latter matrix is large whenever samples i and j have similar martingale residuals. The test statistic T_0 is, therefore, relatively large whenever the entries of the matrices R and $(\mathbf{d} - \hat{\mathbf{u}})(\mathbf{d} - \hat{\mathbf{u}})'$ are correlated, which happens when similarity in gene expressions tends to coincide with similarity in the martingale residual. Hence, the test statistic is large if individuals who die sooner than expected tend to be relatively similar in their gene expression profile and, similarly, the individuals who live longer than expected also tend to be similar in their gene expression profile.

Gene plot

To investigate the influence of individual genes on the test outcome we can rewrite $R = \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i'$, where $\mathbf{x}_i$ is the i-th column of X ($i = 1, \ldots, m$), containing the measurements for the i-th gene. The unstandardized test statistic then becomes

$$T_0 = \sum_{i=1}^{m} T_i$$

where $T_i = (\mathbf{d} - \hat{\mathbf{u}})'\mathbf{x}_i\mathbf{x}_i'(\mathbf{d} - \hat{\mathbf{u}})$ is exactly the unstandardized 'global' test statistic for testing whether the 'pathway' containing only gene i is associated with survival. The test statistic of a pathway is therefore a weighted average of the test statistics for the m genes in the pathway.

In a plot we can visualize the influence of the individual genes by showing the values $T_i - \hat{\mathrm{E}}T_i$, with their standard deviation under the null hypothesis (calculated using the methods of section 3.3). An example of such a 'gene plot' is given in figure 3.1. In this plot, large positive values indicate genes with a large (positive or negative) association with survival and hence genes that make the pathway more significant. As $T_i \propto \|\mathbf{x}_i\|^2$, genes with more expression variance tend to carry more weight in the pathway.

Note that the visualized values of the gene influences T_i in the gene plot are essentially univariate: they only depend on the gene i itself. The multivariate nature of the test statistic Q is therefore not visible in the gene plot. It comes in because, although T_0 is the sum of the T_i and $\hat{\mathrm{E}}T_0$ is the sum of the $\hat{\mathrm{E}}T_i$, the variance of T_0 is generally not the sum of the variances of the T_i.

The comparative p

The global test tests the null hypothesis that the pathway is not associated with survival. This null hypothesis only depends on the observed survival and on the genes in the pathway itself: the result is absolute, not relative to the other pathways.

However, there are situations in which one would be more interested in a relative result. If the global test on the set of all genes is very significant, we can usually expect a sizeable proportion of the genes on the array to be associated with survival. In that case we can expect many pathways to show association with survival as well. This will hold especially for the larger pathways, which will often include some of the genes which are associated with survival.

In such situations we propose a diagnostic called "comparative p", which can help interpret the p-value that comes out of the test. The comparative p for a pathway of size m with p-value $\bar{p}$ is defined as the proportion of randomly selected sets of genes of the size m that have an global test p-value smaller than or equal to $\bar{p}$. To calculate this comparative p we draw 1,000 or 10,000 random gene sets from the array without replacement.

The comparative p fulfills a role different from the p-value and should only be used alongside it. It is a diagnostic, not a p-value in the statistical sense.

It tells whether the p-value of a group of genes is much lower than could be expected from a gene group of its size in this data set.

3.5 Application: osteosarcoma data

We applied the above methodology to a data set of 17 osteosarcoma patients from the Leiden University Medical Center.

Data

A genome wide screen of gene expression in osteosarcoma was done using Hu133a gene expression chips (Affymetrix, Santa Clara, CA). This chip contains 22,283 genes. A successful hybridization was obtained for 17 osteosarcoma biopsies. Three of the samples were amplified, labelled and hybridized in duplicate, one sample in triplicate. These technical replicates were averaged after gene expression measures were obtained, which was done using *gcrma* (Wu et al., 2004). No pre-selection of genes was made.

The 17 patients were followed up to 10 years. Median survival time was 40 months. Available covariates included the presence of metastasis at diagnosis, histology and response to neo-adjuvant chemotherapy. However, as treatment was not uniform over all patients, these covariates were not prognostic and we did not consider them.

Pathway information was obtained from the Gene Ontology (GO Ashburner et al., 2000) database, using the BioConductor (Gentleman et al., 2004) GO package (Zhang, 2004). Pathways that were considered of specific interest were cell cycle (GO: 7049), DNA repair (GO: 6281), Angiogenesis (GO: 1525), Skeletal development (GO: 1501) and Apoptosis (GO: 6915).

Analysis

When testing pathways of interest, it is advisable to also test the 'pathway' of all genes on the chip for association with survival. This shows whether the overall gene expression profile is associated with survival. The results for the pathway of all genes and for the five pathways of primary interest are given in table 3.1. We calculated the p-value using both the asymptotic theory method and the permutation method (using 100,000 permutations).

The permutation p-values tend to be somewhat more conservative than the asymptotic p-values, reflecting both the slight loss of power for the permutation test and a deviation from asymptotic normality due to the small number of samples.

In this data set the expression profile over the set of all genes on the chip is significantly associated with survival. Note that this does *not* mean that every

TABLE 3.1: *Global Test results for the Osteosarcoma data and the pathways of primary interest. The p-values were calculated using the permutation and asymptotic method. The final column gives the comparative p (see section 3.4) .*

pathway	genes	Q	perm. p	asym. p	comp. p
All genes	22283	2.446	0.0120	0.0072	—
Cell cycle	1115	2.957	0.0042	0.0016	0.006
DNA rep.	271	3.123	0.0006	0.0009	0.011
Angiogen.	66	0.917	0.1429	0.1795	0.774
Skel. dev.	185	0.002	0.4133	0.4992	0.998
Apoptosis	656	2.533	0.0093	0.0057	0.210

gene on the chip is associated with survival. It means that the patients who die early are relatively similar to each other in terms of their overall expression profile, while patients who live long are likewise relatively similar. It also means that there is some potential for prediction of survival based on gene expression, even before any pre-selection of genes. The cell cycle, DNA repair and apoptosis pathways are clearly associated with survival, while there is no evidence for this association in angiogenesis and skeletal development.

Because the test for all genes was significant, we expect a sizeable proportion of genes to be associated with survival, so that many pathways will be associated with survival. The comparative p gives a measure whether the p-value found for the pathway is unusually low given that it is a pathway of its size from this data set (see section 3.4). For the results in table 3.1 10,000 gene sets were sampled for each pathway. We used the asymptotic p-values for the comparative p calculations.

We conclude that cell cycle and DNA repair are more clearly associated than could be expected from a gene set of its size in this data set: only around 60 out of 10,000 random gene sets of size 1,115 have a lower p-value than the cell cycle pathway. The expression profile of the apoptosis pathway is clearly associated with survival, as can be seen from the p-values; however it is not exceptional in that: more than 20% of random gene sets have a lower p-value than apoptosis. The Skeletal development pathway is interesting in its own way: it is clearly not associated with survival ($p = 0.5$) and this is quite exceptional for a pathway of this size in this data set: only around 20 in 10,000 random gene sets had a higher p-value. The skeletal development pathway seems to include uncommonly few genes which are associated with survival.

It can occur in some data sets that the set of all genes is not significant, while some pathways (eg. DNA repair) are significant. This occurs in table 3.1 for example if we use FDR-adjusted p-values with a threshold of 0.01 (Benjamini and Yekutieli, 2001). The result for all genes can be seen as a false negative test result. However, another valid interpretation is that prediction of survival without biological pre-selection of genes is uncertain, but if it is known a priori that the genes in the DNA repair pathway are likely to be informative, some prediction of survival is possible.

Mining the GO database

If it is not a priori known which pathways are of specific interest, one can also use a data-mining approach, trying to find those pathways which are most significantly associated with survival.

For the osteosarcoma data we explored the Gene Ontology database. Of all GO terms, 4,032 matched at least one gene on the hu133a chip. We excluded all terms which matched only one gene, because the interesting single genes pathways would already have been found in single gene testing. This left 3,080 pathways, which we all tested for association with survival. We used the asymptotic p-value, because due to the randomness in the permutation p-value it does not give a unique list. Table 3.2 gives the ten GO-terms with the smallest p-values.

To adjust for multiple testing, one can use the Benjamini and Hochberg FDR (Benjamini and Yekutieli, 2001). All 10 pathways in table 3.2 are significant on an FDR of 0.05. The p-values of the pathways tend to have positive correlations because of pathway overlap and pathways being subsets of other pathways. A FDR-controlling procedure that would make use of these dependencies would potentially gain much power in this situation.

TABLE 3.2: *Global Test results for the Osteosarcoma data on 3,080 Gene Ontology pathways, showing the top 10 FDR-adjusted p-values.*

pathway	# genes	Q	FDR-adjusted p
GO:0015630	21	4.306	0.016
GO:0019932	8	4.176	0.016
GO:0045192	2	4.148	0.016
GO:0045595	17	4.060	0.016
GO:0042518	7	4.054	0.017
GO:0000158	8	3.993	0.018
GO:0040008	9	3.944	0.018
GO:0010033	10	3.844	0.023
GO:0006479	13	3.791	0.026
GO:0030111	9	3.766	0.026

The literature confirmed the importance of many of these GO-terms in tumorigenesis. For example, both microtubule cytoskeleton (GO:0015630) and phosphorylation of Stat3 protein (GO:0042518) are known to be involved in growth and differentiation signaling, processes which are often disturbed in tumors. Second-messenger mediated signaling (GO:0019932) is a superset of the Stat3 pathway. Protein amino acid methylation (GO:0006479) is involved in protein degradation. Alterations in the stability of proteins is often a hallmark of tumors and may affect the aggressiveness of a tumor and thereby the patient's survival.

A diagnostic plot

To learn more about the outcome of the Global Test than just the p-value one can use the diagnostic plot described in section 3.4. We illustrate the use of this plot on the microtubule cytoskeleton pathway, which emerged on top of table 3.2.

The gene plot for the 21 genes in this pathway is given in figure 3.1. Each bar gives the global test statistic for testing whether the gene set containing only that single gene is associated with survival. The test statistic for the whole pathway is a weighted average of the bars of the genes (see section 3.4). The colour of the bars distinguishes between positive and negative association with survival.

Figure 3.1 shows that only four out of 21 genes in the microtubule cytoskeleton pathway show a significant association with survival on their own. Further, the pathway is a mix of genes which are positively and negatively associated with survival. Looking more closely at the gene plot can be a basis for investigating more deeply into the structure of the pathway, perhaps to formulate hypotheses on interesting subpathways.

3.6 Discussion

It has often been remarked that the key to successful microarray data analysis lies in an intelligent integration of advanced statistical methods with the vast domain of biological knowledge that is already available. The global test for survival presented in this paper is a step forward in this direction, combining known biological pathway information with the statistical sophistication of the Cox proportional hazards model.

Due to its complexity the Cox model has been slow to find its way to microarray methodology. Most methods require survival to be reduced to a two-valued variable, using an arbitrary cut-off, resulting in unnecessary loss of information. By using the Cox model for survival, gene expression analysis can improve performance and also become better connected to traditional medical

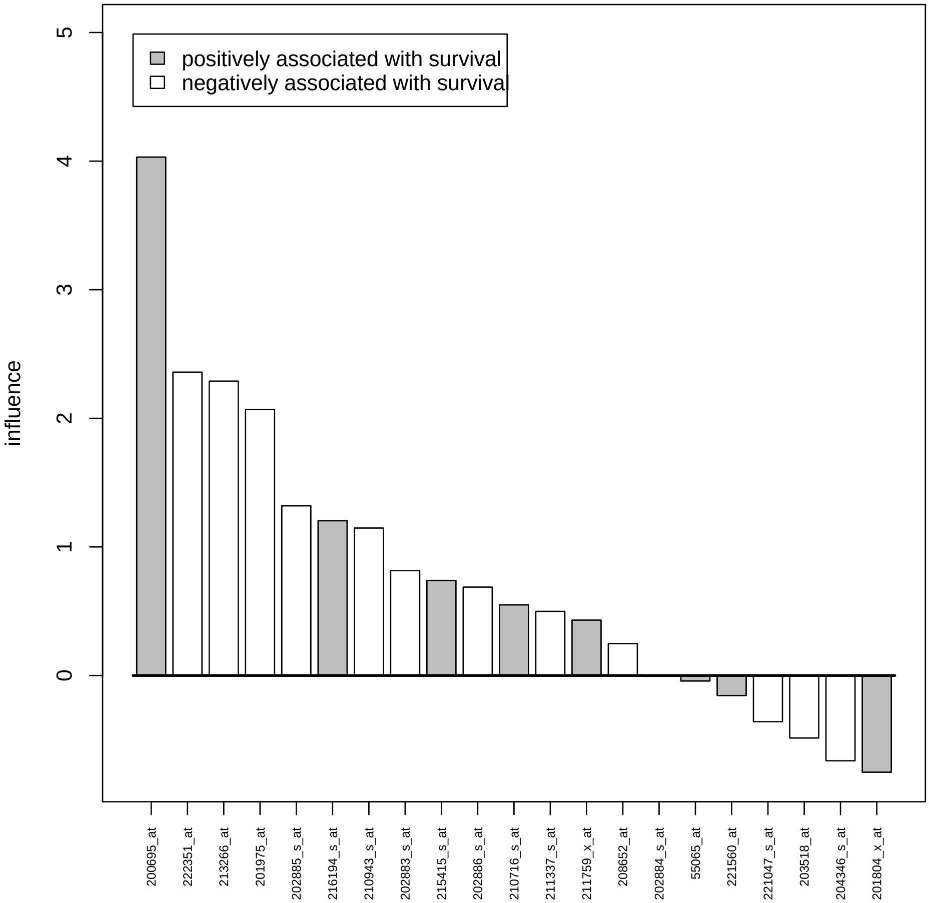

FIGURE 3.1: *Gene plot of microtubule cytoskeleton pathway, showing the sorted global test statistics for testing the 21 single gene pathways which make up the pathway.*

statistics.

Pathway information is available from many databases and is essential for the understanding of the outcomes of a microarray experiment. The Global Test methodology allows researchers to look directly for important pathways, without first having to go through single gene testing. This may lead to a better use of pathway information and more directly interpretable results.