Journal of Statistical Software

January 2009, Volume 29, Issue 1. *http://www.jstatsoft.org/*

Robust Likelihood-Based Survival Modeling with Microarray Data

HyungJun Cho
Korea University

Ami Yu
Korea University

Sukwoo Kim
Korea University

Jaewoo Kang
Korea University

Seung-Mo Hong
Johns Hopkins Medical Institutions

Abstract

Gene expression data can be associated with various clinical outcomes. In particular, these data can be of importance in discovering survival-associated genes for medical applications. As alternatives to traditional statistical methods, sophisticated methods and software programs have been developed to overcome the high-dimensional difficulty of microarray data. Nevertheless, new algorithms and software programs are needed to include practical functions such as the discovery of multiple sets of survival-associated genes and the incorporation of risk factors, and to use in the R environment which many statisticians are familiar with. For survival modeling with microarray data, we have developed a software program (called **rbsurv**) which can be used conveniently and interactively in the R environment. This program selects survival-associated genes based on the partial likelihood of the Cox model and separates training and validation sets of samples for robustness. It can discover multiple sets of genes by iterative forward selection rather than one large set of genes. It can also allow adjustment for risk factors in microarray survival modeling. This software package, the **rbsurv** package, can be used to discover survival-associated genes with microarray data conveniently.

Keywords: microarray data, survival data, likelihood, robustness, R.

1. Introduction

Gene expression can be associated with clinical outcomes such as survival. Genes associated with clinical outcomes can play a role as biomarkers for medical uses. Efforts to discover such biomarker genes have been made by many investigators. For these purposes, the development

of microrray technology presents a challenge to quantitative researchers as well as biological researchers.

To discover survival-associated genes, statistical methods for survival analysis such as the Cox model, the log-rank test, and the Wald test have been applied to various disease studies with microarray experiments (Rosenwald *et al.* 2002; Beer *et al.* 2002; Wigle *et al.* 2002; Jenssen *et al.* 2002; Freije *et al.* 2004; Sanchez-Carbayo *et al.* 2006; Mandruzzato *et al.* 2006; Matsui 2006). Score or Mantel tests were also employed (Shannon *et al.* 2002; Goeman *et al.* 2005; Jung *et al.* 2005). The L_1 or L_2 penalized estimation for the Cox model or the transformed model were applied to improve performance (Bair and Tibshirani 2004; Gui and Li 2005; Xu *et al.* 2005) and partial least squares and LASSO were utilized for data reduction (Nguyen and Rocke 2002; Park *et al.* 2002). Bayesian approaches were also applied (Tadesse *et al.* 2005; Sha *et al.* 2006).

Though there exist such various algorithms, we wanted to develop a new software program that we could use conveniently and interactively in the R environment (R Development Core Team 2008) because R is a widely used statistical package and practical functions for microarray survival analysis need to be included. We utilized the partial likelihood of the Cox model which has been the basis for many of the aforementioned methods. Our algorithm is simple and straight-forward, but its functions such as the generation of multiple gene models and the incorporation of risk factors are practical. For robustness, it also selects survival-associated genes by separating training and validation sets of samples. This is because such a cross-validation technique is essential in predictive modeling for data with large variability. The program employs forward selection to generate a series of gene models and later select an optimal model. Furthermore, iterative runs after putting aside the previously selected genes can discover the masked genes that may be missed by the forward selection. This software package, the **rbsurv** package, is available from the **Bioconductor** website at `http://www.bioconductor.org/` (Gentleman *et al.* 2004), which provides many bioinformatics packages used in the R environment (R Development Core Team 2008). A programming example can be found in the package's vignette.

2. Implementation

2.1. Robust likelihood-based survival modeling

Suppose the data consist of G genes and N samples, and each sample has its observed (survival or censoring) time and censoring status. Thus, it consists of the triple $(Y_j, \delta_j, \mathbf{X}_j)$, $j = 1, \ldots, N$, where Y_j and δ_j are observed time and censoring status (usually, 1=died, 0=censored) for the j-th sample respectively, and $\mathbf{X}_j = (X_{1j}, X_{2j}, \ldots, X_{Kj})$ is the j-th vector of the expression values for K genes ($K < N$ and $K \subset G$). Let $Y_{(1)} < Y_{(2)} < \ldots < Y_{(D)}$ denote the ordered times with D distinct values and $X_{(i)k}$ be the k-th gene associated with the sample corresponding to $Y_{(i)}$. The Cox proportional hazards model (Cox 1972) is $h(y|X_1, X_2, \ldots, X_K) = h_0(y) \exp(\sum_{k=1}^{K} \beta_k X_k)$, where $h(y|X_1, X_2, \ldots, X_K)$ is the hazard rate at time y for a sample with risk vector $(X_1, X_2, \ldots, X_K)$, $h_0(y)$ is an arbitrary baseline hazard rate, and β_k is the coefficient for the k-th gene. The partial likelihood for the Cox

model is

$$\sum_{i=1}^{D}\sum_{k=1}^{K}\beta_k X_{(i)k} - \sum_{i=1}^{D}\log\left(\sum_{j\in R(Y_{(i)})}\exp(\sum_{k=1}^{K}\beta_k X_{jk})\right), \tag{1}$$

where $R(Y_{(i)})$ is the set of all samples that are still under study at a time just prior to $Y_{(i)}$. Maximizing the likelihood provides the maximum likelihood estimates (MLE) of the coefficients, so denote the MLEs by $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$. Then, as a goodness-of-fit, we can use the fitted partial likelihood:

$$\text{loglik} = \sum_{i=1}^{D}\sum_{k=1}^{K}\hat{\beta}_k X_{(i)k} - \sum_{i=1}^{D}\log\left(\sum_{j\in R(Y_{(i)})}\exp(\sum_{k=1}^{K}\hat{\beta}_k X_{jk})\right). \tag{2}$$

The negative log-likelihood (-loglik) is greater than zero, so the smaller -loglik the model better. For robustness, however, the model should be evaluated by independent validation samples rather than the training samples used for fitting the model such as

$$\text{loglik}^* = \sum_{i=1}^{D^*}\sum_{k=1}^{K}\hat{\beta}^0_k X^*_{(i)k} - \sum_{i=1}^{D^*}\log\left(\sum_{j\in R(Y^*_{(i)})}\exp(\sum_{k=1}^{K}\hat{\beta}^0_k X^*_{jk})\right), \tag{3}$$

where $*$ indicate the use of the validation samples and the estimates $\hat{\beta}^0_1, \hat{\beta}^0_2, \ldots, \hat{\beta}^0_k$ are obtained by the training samples. For robust gene selection, we thus use training samples for model fitting and validation samples for model validation. This cross-validation is repeated many times independently. In other words, we fit the Cox model with a gene (or genes) and select a gene (or genes) maximizing mean loglik* (i.e., minimizing the mean negative loglik*).

2.2. Robust gene and model selections

Independent validation samples are usually unavailable. Therefore, the given samples are randomly divided into training and validation sets of samples. To avoid a bias arising from a partition of samples, such partitions are repeated many times and the measurements are averaged. By robust likelihood-based modeling with repeated random partitions, we select the best predictive gene $g_{(1)}$, and then the next best predictive gene $g_{(2)}$ retaining the first selected gene $g_{(1)}$. This forward gene selection generates a series of gene predictive models: $g_{(1)}$, $g_{(1)} + g_{(2)}$, $g_{(1)} + g_{(2)} + g_{(3)}$, ... This can be continued until fitting is impossible because of the lack of samples. It is also important to select an optimal model among a series of gene predictive models. In microarray studies, sufficient samples for double-partitioning as well as independent samples are often unavailable because of the limited supplies of biological samples or the high cost of the experiments. Double-partitioning means that the samples are partitioned into training and validation sets for selecting predictive genes and a test set for selecting an optimal model. Thus, we may have to reuse samples used for gene selection. For this, it is essential to employ an error measure that can prevent over-fitting. For instance, if we test the series of gene models with the samples with loglik, we always select the largest model. To avoid this over-fitting, we employ the Akaike information criterion (AIC), $-2\text{loglik} + ak$, where k is the number of parameters in the model and a is some pre-specified constant (usually, $a = 2$). We select a model with the smallest AIC. Then we can select an optimal robust model, which may not be the largest model selected by using loglik. We discuss the demonstration in the results and discussion section.

2.3. Multiple optimal sets of genes

As described above, an optimal model consists of several survival-associated genes which can be selected and then utilized. However, other genes may also be associated with survival, but not the members of the model because of the masking effect. For instance, suppose there exist two genes similarly associated with survival. If one gene is selected, the other gene may not help improve the model so as not to be selected. Nevertheless, the other survival-associated gene may play an important role. Thus, it is more meaningful to select and provide multiple optimal sets of genes rather than a single optimal set of genes, as indicated in Ein-Dor et al. (Ein-Dor *et al.* 2005). Thus, we put aside the genes in the first model. We then construct another optimal model. Again, we can construct a third optimal model after putting aside the genes in the first and second optimal models. In this manner, we can make multiple optimal models. This iterative procedure can also mitigate the limitation of forward gene selection, which highly depends on the previously selected genes. The number of optimal models can be determined in the practical point of view. The cost and time of confirmation experiments in each lab can be factors to determine it. Among the selected models, the first optimal one might be the best statistically, but it does not mean that it is the best biologically.

2.4. Adjusting for risk factors

Survival may be associated with some risk factors, such as age and a disease stage rather than certain genes. Survival modeling without an adjustment of risk factors may result in finding the genes associated with other risk factors rather than survival. Therefore, we can improve the ability to discover truly survival-associated genes by modeling genes after adjusting for certain risk factors. Thus, we allow adjustment for risk factors in the above robust likelihood-based survival modeling.

2.5. Reducing computing time

Forward gene selection by repeatedly swapping training and validation samples substantially inflates computing time for modeling high-throughput data such as microarray gene expression data with tens of thousands of genes. Therefore, it is crucial to reduce computing time without loss of meaningful information in a practical view. It is a fundamental step to filter out meaningless probe sets meaning in the microarray experiments. For instance, probe sets are kept if a coefficient of variation is greater than 0.2 and at least 10% of samples have an expression intensity greater than 500 (Freije *et al.* 2004). However, a lot of probe sets are often left after initial filtering, so it is wasting computing time to evaluate all of them because many of them may not be associated with survival. Thus, univariate survival modeling and evaluating with whole samples can reduce the number of candidate genes without losing important genes. This univariate survival pre-selection is adopted into our software program.

We can also consider a biological approach if we can obtain other data sets from related studies. For instance, suppose we have a microarray data set with drug A-treated or untreated samples, in addition to a microarray data set for cancer patients with survival. Examining the differential expression under two conditions of the experiments provides Drug A-induced genes. This handful of Drug A-induced genes can be used for robust survival modeling. Finally the selected genes are Drug A-induced as well as associated with survival.

2.6. Algorithm for modeling microarray data with survivals

We now summarize the algorithm described above. Suppose the data consists of expression values X for G genes and N samples. Each sample has its observed (survival or censoring) time Y and censoring status δ. We assume that expression values are already normalized and transformed in appropriate ways. Prior gene selection such as univariate survival modeling and/or biological pre-selection can also be performed if necessary. Univariate survival modeling can be performed in our software program. Our algorithm is summarized as follows.

1. Randomly divide the samples into the training set with $N(1-p)$ samples and the validation set with Np samples (e.g., $p = 1/3$). Fit a gene to the training set of samples and obtain the parameter estimate $\hat{\beta}_i^0$ for the gene. Then evaluate loglik* with the parameter estimate, $\hat{\beta}_i^0$, and the validation set of samples, $(Y_i^*, \delta_i^*, X_i^*)$. Perform this evaluation for each gene.

2. Repeat the above procedure B times (e.g., $B = 100$), thus obtaining B loglik*s for each gene. Then select the best gene with the smallest mean negative loglik* (or the largest mean loglik*). The best gene is the most survival-associated one that is selected by the robust likelihood-based approach.

3. Let $g_{(1)}$ be the selected best gene in the previous step. Adjusting for $g_{(1)}$, find the next best gene by repeating the previous two steps. In other words, evaluate $g_{(1)} + g_j$ for every j and select an optimal two-gene model, $g_{(1)} + g_{(2)}$.

4. Continue this forward gene selection procedure until fitting is impossible because of the lack of samples, resulting in a series of K models $\mathcal{M}_1 = g_{(1)}$, $\mathcal{M}_2 = g_{(1)} + g_{(2)}$, ..., $\mathcal{M}_{K-1} = g_{(1)} + g_{(2)} + \ldots + g_{(K-1)}$, $\mathcal{M}_K = g_{(1)} + g_{(2)} + \ldots + g_{(K)}$.

5. Compute AICs for all the K candidate models, $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K$, and select an optimal model with the smallest AIC.

6. Put aside the genes in the optimal model in the previous step. Then repeat steps 2-6. This can be repeated several times sequentially (e.g, 3 times), generating multiple optimal models.

In addition, suppose that p risk factors, $Z_1, Z_2, \ldots, Z_p$, are available for each sample. Then risk factors can be included in every modeling of the previous algorithm.

2.7. Software

The above algorithm was implemented into a R package (called **rbsurv**), employing two other R packages **Biobase** (Gentleman *et al.* 2004) and **survival** (Therneau and Lumley 2008). This **rbsurv** package can be used conveniently and interactively in the R environment (R Development Core Team 2008). For instance, the package can be run as follows.

```
R> library("rbsurv")
R> fit <- rbsurv(time = time, status = status, x = x, z = z, alpha = 0.05,
+    gene.ID = NULL, method = "efron", max.n.genes = 100, n.iter = 100,
+    n.fold = 3, n.seq = 3, seed = 1234)
```

Argument	Description
`time`	a vector for survival times
`status`	a vector for survival status, 0=censored, 1=event
`x`	a matrix for expression values (genes in rows, samples in columns)
`z`	a matrix for risk factors
`alpha`	a significance level for evaluating risk factors
`gene.ID`	a vector for gene IDs; if `NULL`, row numbers are assigned.
`method`	a character string specifying the method for tie handling.
`n.iter`	the number of iterations for gene selection
`n.fold`	the number of partitions of samples
`n.seq`	the number of sequential runs or multiple models
`seed`	a seed for sample partitioning
`max.n.genes`	the maximum number of genes considered

Table 1: Argument description.

The required arguments `time` and `status` are vectors for survival times and survival status (0=censored, 1=event) and `x` is a matrix for expression values (genes in rows, samples in columns). The optional argument `z` is a matrix for risk factors. To include only the significant risk factors, a significance level less than 1 is given to `alpha`, e.g., `alpha = 0.05`. In addition, there are several controlled arguments. `gene.ID` is a vector for gene IDs; if `NULL`, row numbers are assigned. `method` is a character string specifying the method for tie handling. One of `efron`, `breslow`, `exact` can be chosen. If there are no tied death times all the methods are equivalent. In the algorithm of Section 2.6, `n.fold` is the number of partitions of samples in step 1, `n.iter` is the number of iterations for gene selection in step 2, and `n.seq` is the number of sequential runs (or multiple models) in step 6. `seed` is a seed for sample partitioning. `max.n.genes` is the maximum number of genes considered. As described in Section 2.5, if the number of the input genes is greater than the given maximum number, it is reduced by fitting individual Cox models and selecting the genes with the smallest p-values. The input arguments of **rbsurv** are summarized in Table 1. The major output `fit$model` contains survival-associated gene models with gene IDs, nlogliks, and AICs. The genes in the optimal model with the smallest AIC are marked with asterisks (*).

The open-source R statistical package is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/` and our developed program **rbsurv** is available at the **Bioconductor** website (`http://www.bioconductor.org/`). A programming example can be found in the accompanying vignette.

3. Examples

We now describe a demonstration of our developed algorithm with a microarray data set for patients with gliomas. This real data set consists of gene expression from 85 patients with gliomas (Freije *et al.* 2004). For this study, Affymetrix U133A and U133B chips were used and **dCHIP** was used to convert data files (.CEL) into expression values with median intensity normalization. As suggested in the paper, we first selected about 8,000 genes with a coefficient of variation greater than 0.2 and at least 10% of the samples having an expression

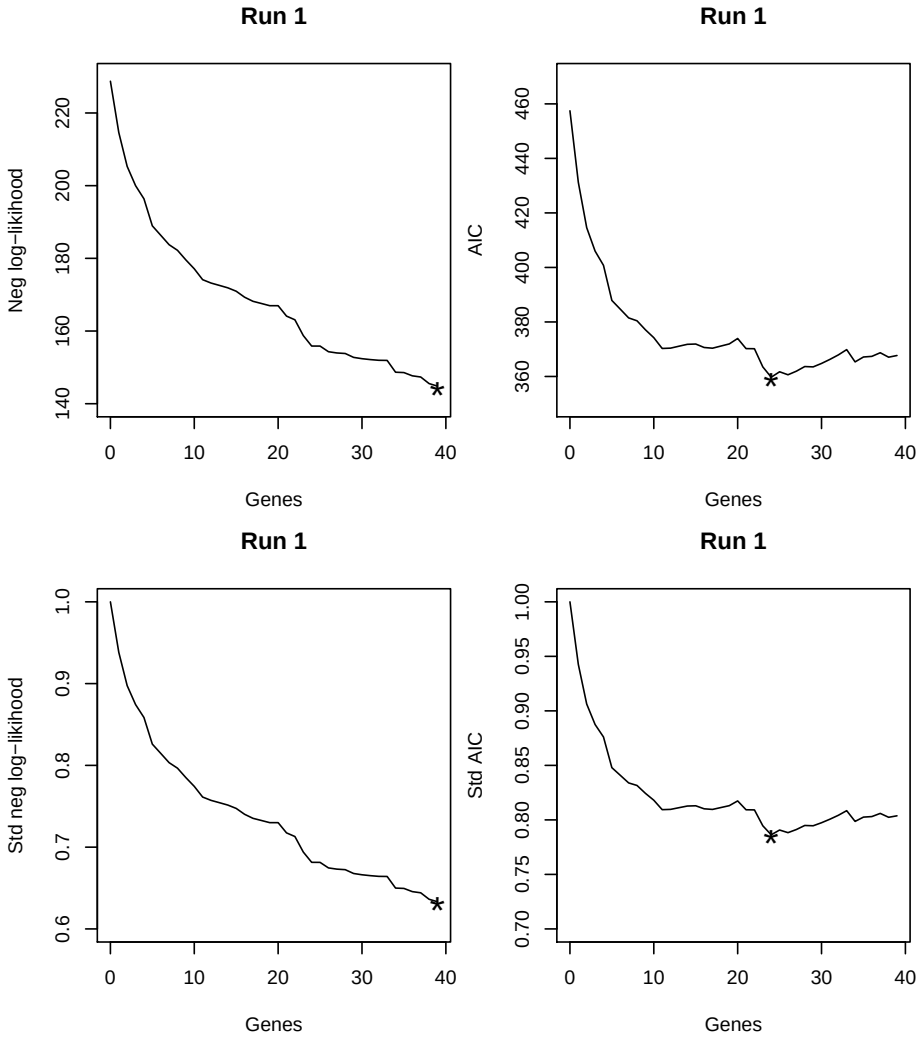

Figure 1: Negative log-likelihood and AIC (1st run). These plots show negative log-likelihoods and AICs against genes. Two plots at the bottom utilized standardized negative log-likelihoods and standardized AICs, which were divided by those with no gene, respectively. The asterisk (*) indicates the smallest value.

intensity greater than 500. We ran our developed software program (called **rbsurv**) to discover survival-associated genes with microarray data for the 85 patients with gliomas.

3.1. Negative log-likelihoods and AICs for two iterative runs

Figure 1 shows that the negative log-likelihoods (nloglik) always decrease as the number of genes increases. Thus, the largest model is always selected because it has the smallest nloglik. However, it could be over-fitted, i.e, consisting of too many genes. In contrast, AICs tend to decrease for a while and then increase with the number of genes. This implies

Probe set ID	Gene title	Gene symbol
200909_s_at	ribosomal protein, large, P2	RPLP2
201105_at	lectin, galactoside−binding, soluble, 1 (galectin 1)	LGALS1
201186_at	low density lipoprotein receptor−related protein	LRPAP1
201318_s_at	myosin regulatory light chain MRCL3	MRLC2
202345_s_at	fatty acid binding protein 5 (psoriasis−associated)	LOC653327
203026_at	zinc finger and BTB domain containing 5	ZBTB5
203303_at	dynein, light chain, Tctex−type 3	DYNLT3
203554_x_at	pituitary tumor−transforming 1	PTTG1
209191_at	tubulin, beta 6	TUBB6
211935_at	ADP−ribosylation factor−like 6 interacting protein	ARL6IP
211937_at	eukaryotic translation initiation factor 4B	EIF4B
212473_s_at	microtubule associated monoxygenase	MICAL2
213447_at	imprinted in Prader−Willi syndrome	IPW
215947_s_at	hypothetical protein FLJ14668	FLJ14668
217733_s_at	thymosin, beta 10	TMSB10
217969_at	chromosome 11 open reading frame2	C11orf2
221249_s_at	family with sequence similarity 117, member A	FAM117A
221623_at	brevican	BCAN
222586_s_at	oxysterol binding protein−like 11	OSBPL11
222820_at	trinucleotide repeat containing 6C	TNRC6C
225864_at	family with sequence similarity 84, member B	FAM84B
226623_at	phytanoyl−CoA 2−hydroxylase interacting protein−like	PHYHIPL
226981_at	Myeloid/lymphoid or mixed−lineage leukemia	MLL
227506_at	solute carrier family 16, member 9	SLC16A9

Table 2: Gene model 1.

Probe set ID	Gene title	Gene symbol
201141_at	glycoprotein (transmembrane) nmb	GPNMB
202182_at	GCN5 general control of amino-acid synthesis 5-like 2	GCN5L2
202409_at	insulin-like growth factor 2 (somatomedin A)	IGF2
207721_x_at	histidine triad nucleotide binding protein 1	HINT1
209180_at	Rab geranylgeranyltransferase, beta subunit	RABGGTB
209395_at	chitinase 3-like 1 (cartilage glycoprotein−39)	CHI3L1
211938_at	eukaryotic translation initiation factor 4B	EIF4B
213479_at	neuronal pentraxin II	NPTX2
215998_at	Sidekick homolog 1 (chicken)	SDK1
218009_s_at	protein regulator of cytokinesis 1	PRC1
220136_s_at	crystallin, beta A2	CRYBA2
220152_at	chromosome 10 open reading frame 95	C10orf95
227082_at	MRNA; cDNA	DKFZp586K1922
229982_at	glutamine and serine rich 1	QSER1

Table 3: Gene model 2.

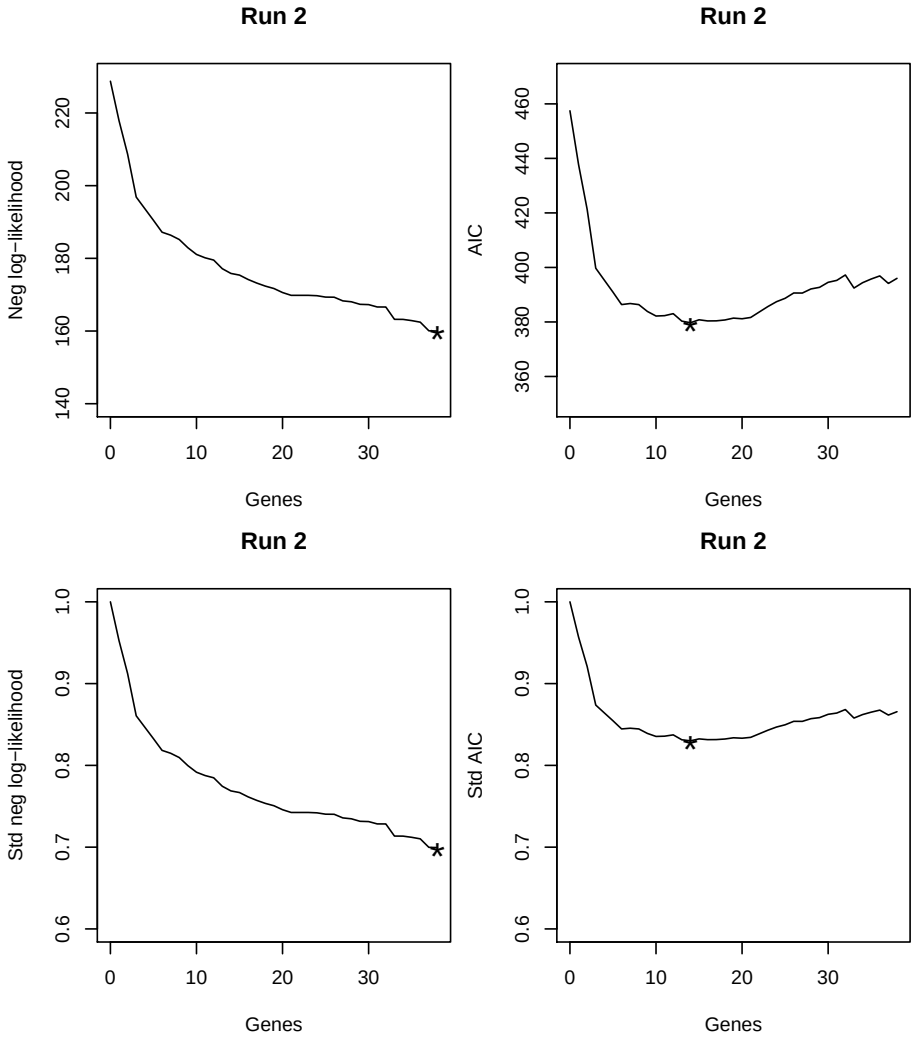

Figure 2: Negative log-likelihood and AIC (2nd run). These plots show negative log-likelihoods and AICs against genes. Two plots at the bottom utilized standardized negative log-likelihoods and standardized AICs, which were divided by those with no gene, respectively. The asterisk (*) indicates the smallest value.

that the optimal gene model is not necessarily very large. In this run, the 24-gene model was selected (Table 2). Among the genes in the model, BCAN(Brevican)/ BEHAB(Brain enriched hyaluronan binding) is one of the members of the lectican family protein. It comprises extracellular matrix of the brain with hyaluronan and tenascin-R (Gary *et al.* 2000; Yamaguchi 2000). Although its role is not completely understood, BCAN/ BEHAB is considered to play an important role in structural maintenance of the brain's extracellular matrix (Nakada *et al.* 2005). In normal adult brain, the expression level of BEHAB/brevican is very low, however, its expression is increased in glial origin tumors, including glioblastoma (Jaworski *et al.* 1996; Held-Feindt *et al.* 2006). In rat model studies, over-expression of BCAN/ BEHAB

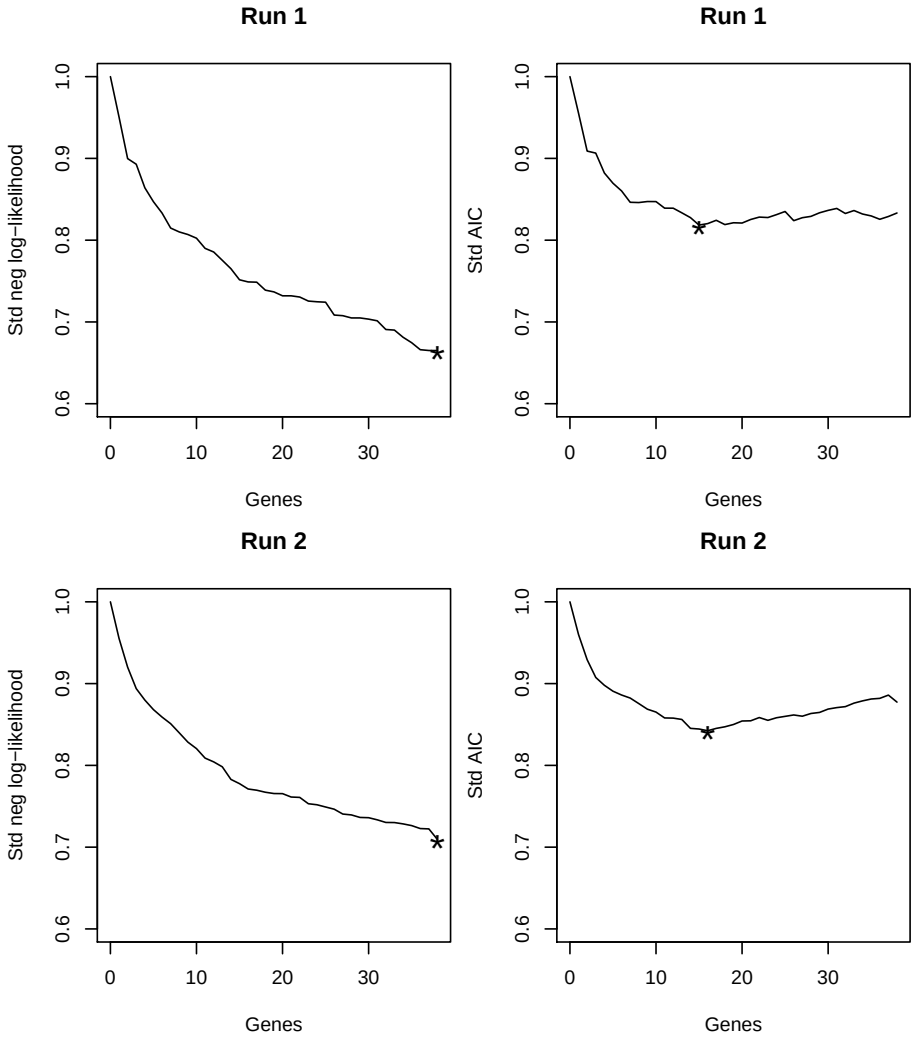

Figure 3: Negative log-likelihood and AIC (with age). These plots show standardized negative log-likelihoods and standardized AICs with age, which were divided by those with no gene, respectively. The asterisk (*) indicates the smallest value.

was reported to be associated with tumor invasion. Rats having intracranial grafts of BCAN/ BEHAB-transfected glioma cell line demonstrated worse prognosis (Jaworski *et al.* 1996; Nutt *et al.* 2001). Taken these together, over-expression of BCAN/ BEHAB may be associated with aggressiveness and worse survival rate in patients with glioblastoma. We re-ran **rbsurv** after putting to one side the genes in the selected model at the first run. Note that these iterative runs can be conducted automatically according to a user's choice in **rbsurv**. In a similar way, we obtained another model containing 14 survival-associated genes (Figure 2 and Table 3). The nlogliks and AICs in Figure 2 had the trends similar to those in Figure 1.

Probe set ID	Gene title	Gene symbol
202182_at	GCN5 general control of amino-acid synthesis 5-like 2	GCN5L2
202442_at	adaptor-related protein complex 3, sigma 1 subunit	AP3S1
205139_s_at	uronyl-2-sulfotransferase	UST
208691_at	transferrin receptor (p90, CD71)	TFRC
209507_at	replication protein A3, 14kDa	RPA3
212468_at	sperm associated antigen 9	SPAG9
213447_at	imprinted in Prader-Willi syndrome	IPW
213479_at	neuronal pentraxin II	NPTX2
217733_s_at	thymosin, beta 10	TMSB10
218009_s_at	protein regulator of cytokinesis 1	PRC1
220136_s_at	crystallin, beta A2	CRYBA2
221249_s_at	family with sequence similarity 117, member A	FAM117A
227082_at	cDNA DKFZp586K1922	—
227506_at	solute carrier family 16, member 9	SLC16A9
229982_at	glutamine and serine rich 1	QSER1

Table 4: Gene model 3

Probe set ID	Gene title	Gene symbol
201186_at	low density lipoprotein receptor-related protein associated protein 1	LRPAP1
202409_at	insulin-like growth factor 2 (somatomedin A)	INS-IGF2
203026_at	zinc finger and BTB domain containing 5	ZBTB5
203303_at	dynein, light chain, Tctex-type 3	DYNLT3
203554_x_at	pituitary tumor-transforming 1	PTTG1
204900_x_at	Sin3A-associated protein, 30kDa	SAP30
205480_s_at	UDP-glucose pyrophosphorylase 2	UGP2
211762_s_at	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)	LOC643995
211938_at	eukaryotic translation initiation factor 4B	EIF4B
215998_at	Sidekick homolog 1 (chicken)	SDK1
218407_x_at	neuron derived neurotrophic factor	NENF
221623_at	brevican	BCAN
224850_at	ATPase family, AAA domain containing 1	ATAD1
225864_at	family with sequence similarity 84, member B	FAM84B
232125_at	CDNA FLJ34585 fis, clone KIDNE2008758	—
242134_at	Transcribed locus	—

Table 5: Gene model 4

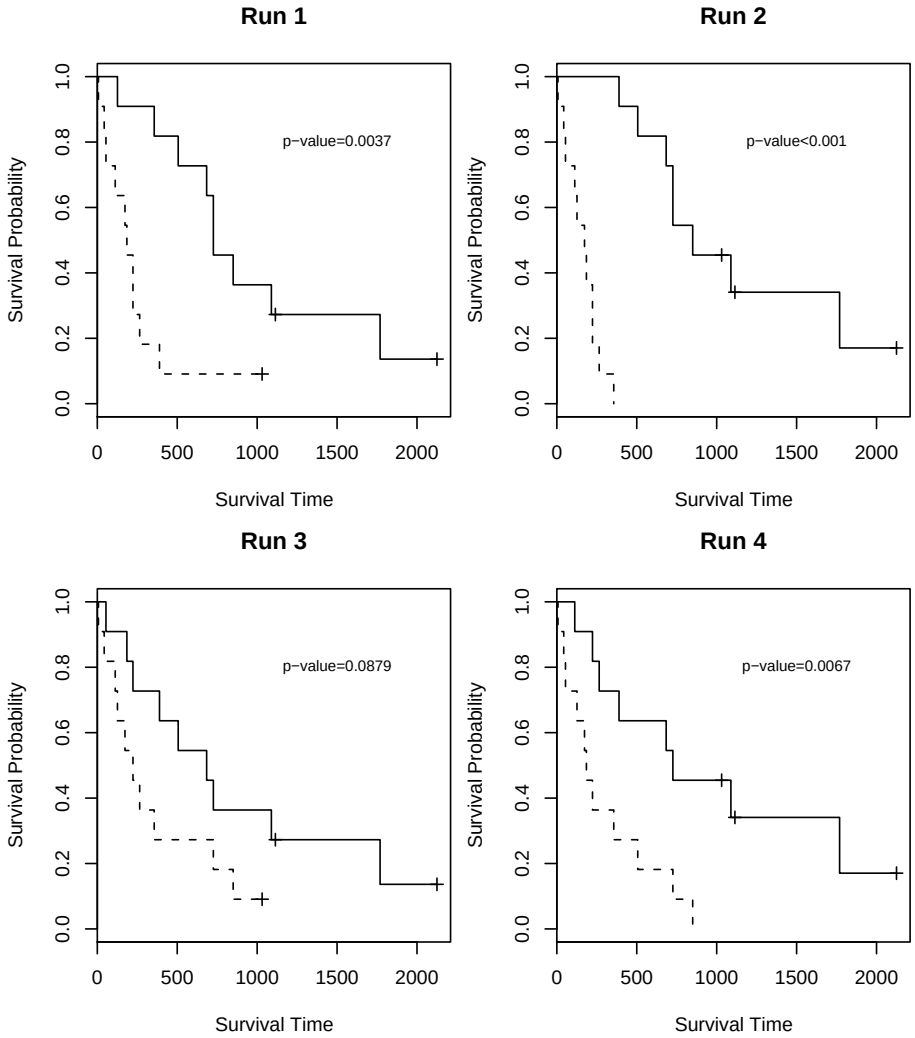

Figure 4: Kaplan-Meier survival curves for high- and low-risk groups. The plots for four sequential runs were drawn by the test samples for the high- and low-risk groups from each optimal gene model.

3.2. Negative log-likelihoods and AICs with adjusting for the covariates

Next, we accounted for potential risk factors of age and gender in this survival modeling. All the risk factors can be incorporated into modeling. Gender was not significant at a 5% significance level. Thus, we let **rbsurv** include age only and generate a series of gene models with nlogliks decreasing and AICs decreasing after increasing as the number of genes increases, as shown in Figure 3. The trends are the same as those in Figures 1 and 2. The selected genes in the two iterative runs are displayed in Tables 4 and 5. In this case, BCAN was included in the second run.

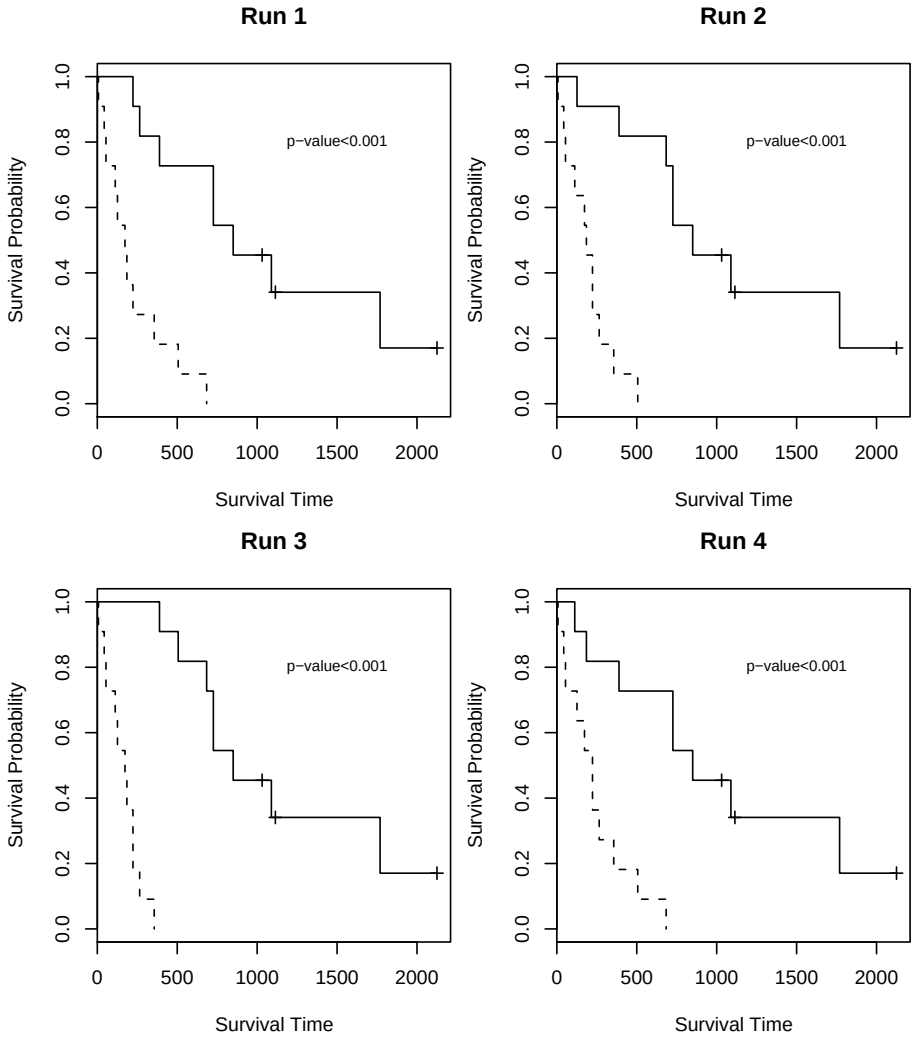

Figure 5: Kaplan-Meier survival curves for high- and low-risk groups with covariate age. The plots for four sequential runs were drawn by the test samples with the high- and low-risk groups from each optimal gene model adjusted for age.

3.3. Evaluation of prediction accuracy with test samples

To see prediction accuracy, we randomly divided 85 samples into two sets (75% for training and validating and 25% for testing). Reserving the test set (22 samples), we ran **rbsurv** with the training and validation set (63 samples) resulting in an optimal gene model. We computed risk scores with the 22 test samples for the selected genes. The risk scores were divided into high-risk and low-risk groups by the median. We did this four times iteratively. Figure 4 shows that the two groups for each run differ significantly. Only in the third run, their difference was not large. Including the significant risk factors, we ran **rbsurv** in the same way and found that the high-risk and low-risk groups differed significantly in all runs. Thus, inclusion of the significant risk factors helps to improve prediction accuracy (Figure 5).

4. Conclusions

For survival modeling with microarray gene expression data, we have developed a robust likelihood-based algorithm and software program called **rbsurv**. This can be used conveniently and interactively in the R environment. Using **rbsurv**, we can discover multiple sets of survival-associated genes while adjusting for risk factors.

Acknowledgments

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-331-C00065).

References

Bair E, Tibshirani R (2004). "Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data." *PLoS Biology*, **2**, 0511–0521.

Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S (2002). "Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma." *Nature Medicine*, **8**, 816–824.

Cox DR (1972). "Regression Models and Life Tables." *Journal of the Royal Statistical Society B*, **34**, 187–220.

Ein-Dor L, Kela I, Getz G, Givol D, E D (2005). "Outcome Signature Genes in Breast Cancer: Is There a Unique Set?" *Bioinformatics*, **21**, 171–178.

Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, Liau LM, Mischel PS, Nelson SF (2004). "Gene Expression Profiling of Gliomas Strongly Predicts Survival." *Cancer Research*, **64**, 6503–6510.

Gary SC, Zerillo CA, Chiang VL, Gaw JU, Gray G, Hockfield S (2000). "cDNA Cloning, Chromosomal Localization, and Expression Analysis of Human BEHAB/Brevican, a Brain Specific Proteoglycan Regulated during Cortical Development and in Glioma." *Gene*, **256**, 139–147.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J (2004). "**Bioconductor**: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology*, **5**, R80. URL http://genomebiology.com/2004/5/10/R80.

Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC (2005). "Testing Association of a Pathway with Survival Using Gene Expression Data." *Bioinformatics*, **21**, 1950–1957.

Gui J, 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**, 3001–3008.

Held-Feindt J, Paredes EB, Blomer U, Seidenbecher C, Stark AM, Mehdorn HM, R M (2006). "Matrix-Degrading Proteases ADAMTS4 and ADAMTS5 (Disintegrins and Metalloproteinases with Thrombospondin Motifs 4 and 5) Are Expressed in Human Glioblastomas." *International Journal of Cancer*, **118**, 55–61.

Jaworski DM, Kelly GM, Piepmeier JM, S H (1996). "BEHAB (Brain Enriched Hyaluronan Binding) Is Expressed in Surgical Samples of Glioma and in Intracranial Grafts of Invasive Glioma Cell Lines." *Cancer Research*, **56**, 2293–2298.

Jenssen TK, Kuo WP, Stokke T, E H (2002). "Associations between Gene Expressions in Breast Cancer and Patient Survival." *Human Genetics*, **111**, 411–420.

Jung SH, Owzar K, George SL (2005). "A Multiple Testing Procedure to Associate Gene Expression Levels with Survival." *Statistics in Medicine*, **24**, 3077–2088.

Mandruzzato S, Callegaro A, Turcatel G, Francescato S, Montesco MC, Chiarion-Sileni V, Mocellin S, Rossi CR, Bicciato S, Wang E, Marincola FM, Zanovello P (2006). "A Gene Expression Signature Associated with Survival in Metastatic Melanoma." *Journal of Translational Medicine*, **4**, 50.

Matsui S (2006). "Predicting Survival Outcomes Using Subsets of Significant Genes in Prognostic Marker Studies with Microarrays." *BMC Bioinformatics*, **7**, 156.

Nakada M, Miyamori H, Kita D, Takahashi T, Yamashita J, Sato H, Miura R, Yamaguchi Y, Okada Y (2005). "Human Glioblastomas Overexpress ADAMTS-5 that Degrades Brevican." *Acta Neuropathologica*, **110**, 239–246.

Nguyen DV, Rocke DM (2002). "Partial Least Squares Proportional Hazard Regression for Application to DNA Microarray Survival Data." *Bioinformatics*, **18**, 1625–32.

Nutt CL, Zerillo CA, Kelly GM, Hockfield S (2001). "Brain Enriched Hyaluronan Binding (BEHAB)/Brevican Increases Aggressiveness of CNS-1 Gliomas in Lewis Rats." *Cancer Research*, **61**, 7056–7059.

Park PJ, Tian L, Kohane IS (2002). "Linking Gene Expression Data with Patient Survival Times Using Partial Least Squares." *Bioinformatics*, **18 Suppl 1**, S120–S127.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, Lopez-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM (2002). "The Use of Molecular Profiling to Predict Survival

after Chemotherapy for Diffuse Large-B-Cell Lymphoma." *The New England Journal of Medicine*, **346**, 1937–1947.

Sanchez-Carbayo M, Socci ND, Lozano J, Saint F, Cordon-Cardo C (2006). "Defining Molecular Profiles of Poor Outcome in Patients with Invasive Bladder Cancer Using Oligonucleotide Microarrays." *Journal of Clinical Oncology*, **24**, 778–789.

Sha N, Tadesse MG, Vannucci M (2006). "Bayesian Variable Selection for the Analysis of Microarray Data with Censored Outcomes." *Bioinformatics*, **22**, 2262–2268.

Shannon WD, Watson MA, Perry A, Rich K (2002). "Mantel Statistics to Correlate Gene Expression Levels from Microarrays with Clinical Covariates." *Genetic Epidemiology*, **23**, 87–96.

Tadesse MG, Ibrahim JG, Gentleman R, Chiaretti S, Ritz J, Foa R (2005). "Bayesian Error-in-Variable Survival Model for the Analysis of GeneChip Arrays." *Biometrics*, **61**, 488–497.

Therneau T, Lumley T (2008). ***survival**: Survival Analysis Including Penalised Likelihood*. R package version 2.34-1, URL http://CRAN.R-project.org/package=survival.

Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, Lu C, Woodgett J, Seiden I, Johnston M, Keshavjee S, Darling G, Winton T, Breitkreutz BJ, Jorgenson P, Tyers M, Shepherd FA, Tsao MS (2002). "Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-Free Survival." *Cancer Research*, **62**, 3005–3008.

Xu J, Yang Y, Ott J (2005). "Survival Analysis of Microarray Expression Data by Transformation Models." *Computational Biology and Chemistry*, **29**, 91–94.

Yamaguchi Y (2000). "Lecticans: Organizers of the Brain Extracellular Matrix." *Cellular and Molecular Life Sciences*, **57**, 276–289.

Affiliation:

HyungJun Cho
Departments of Statistics and Biostatistics
Korea University
Anam-dong, Seongbuk-gu
Seoul, 136-701, Korea
E-mail: hj4cho@korea.ac.kr
URL: http://www.korea.ac.kr/~stat2242/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 29, Issue 1 *Submitted:* 2007-10-16
January 2009 *Accepted:* 2008-09-23