How to use the rbsurv Package

HyungJun Cho, Sukwoo Kim, Jaewoo Kang, Jae Won Lee

October 3, 2007

Contents

1 Introduction

The *rbsurv* package is designed to select survival-associated genes, based on a likelihood function. It utilizes the partial likelihood of the Cox model which has been the basis for many of the existing methods. Our algorithm is simple and straight-forward, but its functions such as the generation of multiple gene models and the incorporation of significant risk factors are practical. For robustness, this package also selects survival-associated genes by separating train and validation sets of samples because such a cross-validation technique is essential in predictive modeling for data with large variability. It employs forward selection, the limitation of which is mitigated by generating a series of gene models and selecting an optimal model. Furthermore, iterative runs after putting aside the previously selected genes can discover the masked genes that may be missed by forward selection (see Cho et al. for details). The *rbsurv* package employs libraries *survival* and *Biobase*.

2 Example: Glioma Data

We demonstrate the use of the package with a microarray data set for patients with gliomas. This real data set consists of gene expression value, survival time, and censoring status of each of 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. This data set originally consists of more than 40,000 probe sets, but only a sub-dataset made up of 100 probe sets was stored into the *rbsurv* package for demonstration.

To run *rbsurv*, the data can be prepared as follows.

```
> library(rbsurv)
> data(glioma)
> ls()

[1] "glioma.cov" "glioma.x"   "glioma.y"

> dim(glioma.x)

[1] 100  85

> glioma.y[1:5, ]

      Time Status
Chip1 1114      0
Chip2  877      0
Chip3  858      0
Chip4 2516      0
Chip5  442      1

> glioma.cov[1:5, ]

      Age Gender
Chip1  51   MALE
Chip2  30 FEMALE
Chip3  31 FEMALE
Chip4  41 FEMALE
Chip5  34 FEMALE

> time <- glioma.y$Time
> status <- glioma.y$Status
> z <- glioma.cov
> x <- log2(glioma.x)
```

We here took log2-transformation wihtout any other normalizations. An appropriate normalization can be taken if needed. If the data is ready, *rbsurv* can be run as follows.

```
> fit <- rbsurv(time = time, status = status, x = x, method = "efron",
+     max.n.genes = 20)

Please wait... Done.
```

This sequentially selects genes one gene at a time to build an optimal gene model. Once a large gene model is constructed, an optimal gene model is determined by AICs. If there exist ties in survival times, Efron's method is used (*method=”efron”*). Note that it is computationally expensive and the data is high-throughput. Therefore, you should be patient to obtain the output. To save time, we can reduce the number of genes considered up to 50 genes among 200 initial genes (*max.n.genes=50*). The 50 genes are selected by fitting univariate Cox models. The above command generates the following output.

```
> par(mfrow = c(2, 2))
> plot(fit$model$Order, fit$model$nloglik, type = "l")
> plot(fit$model$Order, fit$model$AIC, type = "l")
> fit$model
```

```
    Seq Order Gene nloglik    AIC Selected
0     1     0    0  228.74 457.47
110   1     1   46  218.53 439.05 *
2     1     2   57  202.21 408.42 *
3     1     3   99  199.85 405.69 *
4     1     4   36  197.76 403.51 *
5     1     5   43  191.25 392.50 *
6     1     6   34  189.81 391.63 *
7     1     7   15  189.24 392.48 *
8     1     8   29  187.79 391.58 *
9     1     9   19  187.64 393.29
10    1    10   86  186.39 392.79
11    1    11   56  186.22 394.44
12    1    12   68  185.83 395.66
13    1    13   28  185.23 396.47
14    1    14   40  184.47 396.93
15    1    15   75  184.46 398.92
16    1    16   39  184.22 400.44
17    1    17   67  183.58 401.17
18    1    18   96  183.19 402.38
19    1    19   98  182.12 402.24
```

This large gene model contains survival-associated genes which were selected one at a time by forward selection. Note that the first row has no gene ID because it was fitted with no expression profile. The size of the large gene model was determined by the numbers of samples and genes considered. The AICs tend to decrease and then increase, while negative log-likelihood (nloglik) always decrease (see Figure). Thus, we select an optimal model with the smallest AIC. The selected parsimonious model consists of the survival-associated genes.

Potential covariates can be included in modeling and it can be run iteratively. For example, use $rbsurv(time = time, status = status, x = x, z = z, alpha = 0.05, n.seq = 2)$ for significant covariates with level 0.05 and 2 iterative runs.

3 Conclusion

This package allows ones to build multiple gene models sequentially rather than a single gene model. Furthermore, other covariates such as age and gender can also be incorporated into modeling with gene expression profiles. Each model contains survival-associated genes that are selected robustly by separating train and test sets many times.

References

Cho H et al. Robust likelihood-based survival modeling for microarray data, *submitted*.

Freije, W.A., Castro-Vargas, F.E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L.M., Mischel, P.S. and Nelson, S.F. (2004). Gene expression profiling of gliomas strongly predicts survival, Cancer Research, 64:6503-6510.