Description of the sigsquared package

UnJin Lee

May 3, 2016

Contents

1 Introduction

sigsquared is a package designed to parameterize custom models as described in Yun, et al. (2011) via an optimization and cross-validation strategy as described in Lee and Frankenberger, et al. (2013). The customs models described are in the form of linear signaling pathways that are implicated in having a differential survival outcome (using e.g. metastasis-free survival data). The sigsquared package attempts to detect the presence of alternate signaling states of the input pathway that significantly predict differential survival outcome within a mixed cohort of patients. The main goal of this package is to generate gene signatures given known signaling pathways that are predictive of differential survival outcome. The two main functions used to accomplish this goal are first, the ability to train model parameters for a given linear network model, and second, the ability to apply the model and trained parameters to transcript data.

The simplest example of a linear pathway consists of two genes, A and B, with the only known interaction to be that of A strongly inhibiting B. Assuming little connectivity, basal levels of transcript production, and gene decay, high expression of A implies low expression of B, or low expression of A imples high expression of B. This is represented internally as vectors with elements as either -1 or 1. For example, the state with low expression of A and high expression of B would be represented as {-1, 1}, while the state with high expression of A and low expression of B would be represented as {1, -1}. The main goal of this package is to determine if such alternate signaling states, like the de-repression of B via repression of A, can predict differential survival outcome.

Unlike typical network inference models, such as Bayesian methods, the models used for prediction in this package are very simple and based solely upon a single threshold for each network node. While network inference models generally require N^2 model parameters for each pairwise node interaction, where N is the number of nodes, the models used here require only N parameters. However, the model used within require both a linear pathway as well as a priori information on the general regulatory direction

of each interaction (i.e. up-/down-regulation). As such, the features of this package are not likely to be of interest outside of the context of generating a molecular gene signature for a known signaling pathway.

To train the model parameters, this package utilizes Nelder-Mead optimization via R's 'optim' function. Potential parameters are first trained, then cross-validated in an separate cross-validation set. During cross-validation, solutions below a given significance threshold are discarded.

2 Getting started

To install the package:

```
R CMD INSTALL sigsquared_x.y.z.tar.gz
```

sigsquared imports several functions from other packages. Make sure to have the following installed:
Biobase, and *survival*.

3 First Steps

For demonstration purposes we use a test data set provided by this package. To do this, we must first import the library, then import the data set.

```
> library(sigsquared)
> data(BrCa443)
```

We first want to establish the signaling environment for which we are optimizing. In this demonstration, we will use the network described in Yun, et al. (2011).

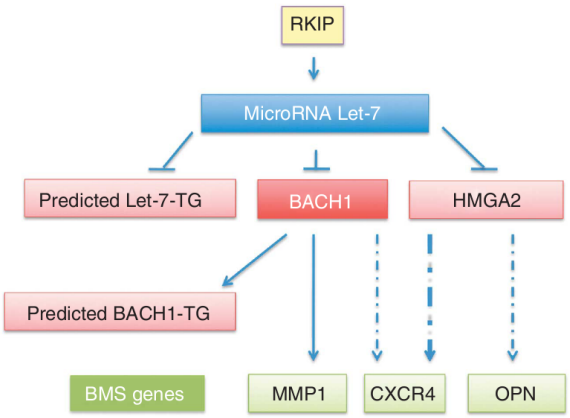


Figure 1: Network, Yun, et al. 2011

In this network, when RKIP is inhibited, Let-7 is also inhibited. As Let-7 is inhibited, its downstream targets, labeled in the data set through its surrogates metaLET7, are derepressed. Similarly, the levels of the other downstream targets BACH1, its surrogates metaBACH1, as well as HMGA2, MMP1, CXCR4, and OPN are predicted to have increased expression levels. As per the parameterizable model described in Yun, et al. (2011), abberant behavior of the entire RKIP signaling environment can be described as patients with RKIP levels below a certain threshold as well as metaLET7, metaBACH1, HMGA2, MMP1, CXCR4, and OPN above their respective thresholds. We encode this directionality as a vector with elements selected from the set {-1, 1}, with -1 describing RKIP thresholding, and 1 describing the other genes (i.e. c(-1, 1, 1, 1, 1, 1, 1), where 'genes' has been defined as c("RKIP", "HMGA2", "SPP1", "CXCR4", "MMP1", "MetaLET7", "MetaBACH1")). At this point, we have enough information to begin generating our new geneSignature object.

```
> gs <- new("geneSignature")
> genes <- c("RKIP", "HMGA2", "SPP1", "CXCR4", "MMP1", "MetaLET7", "MetaBACH1")
> gs <- setGeneSignature(gs, direct=c(-1, 1, 1, 1, 1, 1, 1), genes=genes)
```

4 Setting Thresholds

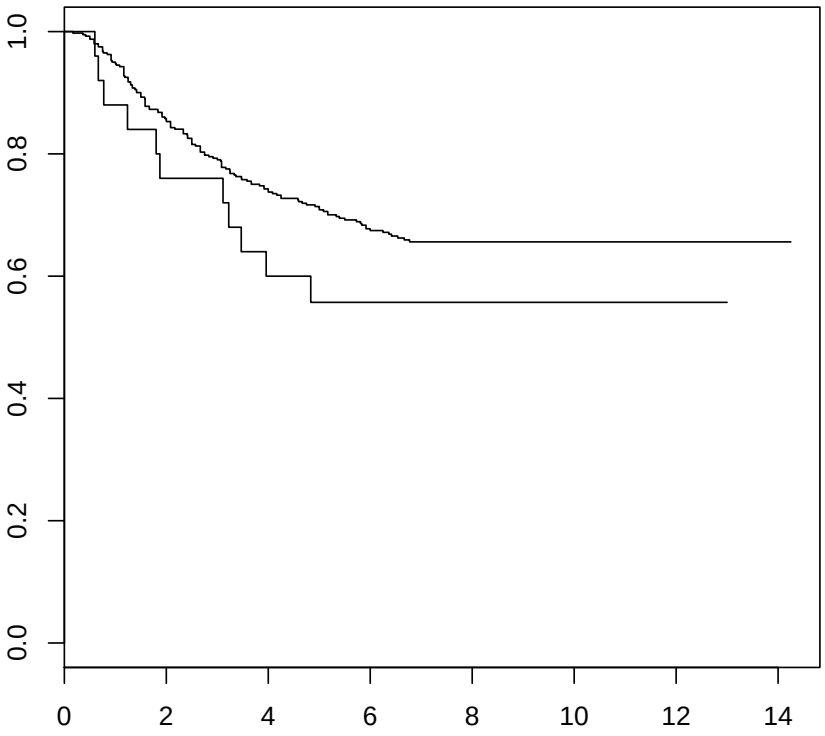
Now that we have established the signaling environment for our geneSignature object, we may now begin optimization. The package will use k-fold cross-validation, with iterPerK optimizations per each k. In this example, we will use a total of 100 optimizations, though in practice, at minimum 2,500 optimizations are suggested. For detailed information on the optimization process, see Lee and Frankenberger, et al. (2013)

```
> gs <- analysisPipeline(dataSet=BrCa443, geneSig=gs, iterPerK=50, k=2, rand=FALSE)
```

5 Applying New Thresholds

With these new thresholds, we want to see if the thresholds yield significance in our data set. To do this, we apply our geneSignature object to our data set and generate a Kaplan-Meier plot.

```
> s <- ensembleAdjustable(dataSet=BrCa443, geneSig=gs)
> plot(survfit(Surv(MFS, met) ~ s, data=pData(BrCa443)))
```


1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14

REFERENCES

Lee U, Frankenberger C, Yun J, Bevilacqua E, Caldas C, et al. (2013) A Prognostic Gene Signature for Metastasis-Free Survival of Triple Negative Breast Cancer Patients. PLoS ONE 8(12): e82125. doi:10.1371/journal.pone.0082125

Yun, J., Frankenberger, C.A., Kuo, W.L., Boelens, M.C., Eves, E. M., Cheng, N., Liang, H., Li, W.H., Ishwaran, H., Minn, A.J. and Rosner, M.R. (2011), Signalling pathway for RKIP and Let-7 regulates and predicts metastatic breast cancer. The EMBO Journal, 30: 4500-4514