An R-package for network based, integrative biomarker signature discovery

Yupeng Cun

December 3, 2013

Bonn-Aachen International Center for IT, Bonn, Germany

(version 1.2.1)

User's Guide for **netClass**

Contents

1 Packages Overview

The package netClass was implemet 5 network based methods for integrating prior knowledge to omics data classification. *netClass* allows for integrating miRNA and mRNA expression profiles together with protein-protein interactions network and known predicted miRNA target mRNA list[2]. Meanwhile, we implement Pathway activity classification [4], Classification based on differential expressed hub nodes (genes) [7]; Filtering of genes according to GeneRank algorithm [9, 1]. In this vignette, we demonstrate how to use the package.

The package can be loaded by typing:

```
> library(netClass)
```

2 Data for netClass

2.1 gene expression data

In this vignette, we take gene expression dataset published by [6] (MAINZ) as example:

```
> library(breastCancerMAINZ)
>library(Biobase)
> data(mainz)
```

We use following comends to get a expression data as matrix x with n patients

of p observations.

```
> x <- t(exprs(mainz)):
>dim(x)
[1] 200 22283
```

This data set contains 22283 probesets measured in 200 samples.

the analysis.

We then classify the patient with e.dmfs or not:

```
> y <- pData(mainz)$e.dmfs
> y[y==0] <- -1
> y <- factor(y)
```

A expression matrix x with labels y were ready for the latter analysis.

2.2 Network

A adjacency matrix that represents the interaction of the biological entity is required. A easy access binary protein-protein interactions (PPI) table could be download from this websites:

http://www.hprd.org/download .

After extracting the zip files, we can read the tab-delimited:

```
> library(pathClass)
> hprd.ppi <- read.hprd(BINARY_PROTEIN_PROTEIN_INTERACTIONS.txt)
```

We use pahClass to get the matched features between the gene expression data matrix x and the hprd adjacency matrix A. Therefore, we need a mapping containing the information which protein of hprd matches to which probe set in x.

As an alternative, the user can load a small mimic network using the command:

```
> data(ad.matrix)
```

For each microarrays platform, a annotation package is matched for further analysis. In this example, we dealing with expression data from chip ***hgu133a*** to MAINZ data to create a mapping from probe set ID to protein ID via pathClass:

```
> ann <- annotation(mainz)
> library(hug133a.db)
> mapped_probes <- mappedkeys(x)
> entrezIDs <- "ENTREZID"
> rs <- get(paste(ann, graphIDs, sep=))
> refEntrez <- mget(featureNames(mainz), rs)
> times <- sapply(refEntrez, length)
> mapping <- data.frame(probesetID=rep(names(refEntrez), times=times),
+ graphID=unlist(refEntrez), row.names=NULL, stringsAsFactors=FALSE)
> nas <- which(is.na(mapping$graphID))
> mapping <- mapping[-nas,]
> mapping <- unique(mapping)
> head (mapping)
probesetID graphID
1 1007_s_at 780
2 1053_at 5982
3 117_at 3310
4 121_at 7849
5 1255_g_at 2978
6 1294_at 7318
```

And then, we will use the function *matchMatrices()* of *pathClass* to match the data matrix x to the network:

```
> library(pathClass)
> matched <- matchMatrices(x=x, adjacency=ad.matrix, mapping=mapping)
```

The upper matched list contains copies of x, *ad.matrix* and *mapping* with matching dimensions.

3 Network-based classication methods in netClass

We implements these network-based gene selection methods in **netClass** package:

1. Random walk kernel based smoothing of t-statistics over a network structure [2].
2. Filtering of genes according to a modified Google PageRank algorithm [9, 1].
3. Classification based on differential expression of hub genes and correlated partners [7].
4. Pathway activity classification [4].
5. Average expression profile of pathways [3].

3.1 Network smooth t-statistic (stSVM)

We implement two methods for selecting features in stSVM [2]: one use permutation test, and another use spanbond to select optimal top ranked sets. We first illustrate permutation test case for the stSVM,

```
> r.stsvm <- cv.stsvm( x=x, x.mi=NULL, y=y, #data matrix with factor labels
+ folds=10, repeats=5, #cross-validation parameters.
+ Gsub=ad.matrix, # adjacency matrix of network.
+ op.method="pt", #use permutation test to optimal
+ op=0.9, # the 10% top ranked features
+ aa=1000, pt.pvalue=0.05,# with 1000 times, and the cutoff p value is 0.05
+ parallel=FALSE, cores=2, # parameter for parallel computing
+ a=1, p=2, # parameter for random walk kenel, p is step, a is integer constant
+ allF=TRUE, # use all features or only matched features
+ Cs=10^(-3:3), # C parameter for C-SVM
+ DEBUG=TRUE, seed=1234 )
```

or, to use the span bound case of the stSVM:

```
r.stsvm <- cv.stsvm( x=x, x.mi=NULL, y=y, #data matrix with factor labels
+ folds=10, repeats=5, #cross-validation parameters.
+ Gsub=ad.matrix, # adjacency matrix of network.
+ op.method="spb", #use span bound technique to optimal
+ op=20, aa=1000, # the top 100 to 20 ranked genes
+ parallel=FALSE, cores=2, # parameter for parallel computing
+ a=1, p=2, # parameter for random walk kenel, p is step, a is integer constant
+ allF=TRUE, # use all features or only matched features
+ Cs=10^(-3:3), # C parameter for C-SVM
+ DEBUG=TRUE, seed=1234 )
```

For more detail, please take a look into the help files or the paper ([2]) . The combined miRNA and mRNA case is similar.

3.2 Filtering of genes according to GeneRank (FrSVM)

The GeneRank algorithm gives a ranking of genes according to their centrality in a interaction network based on their differential gene expression values [5]. The top ranked usually have a strong association with the disease pathology. We use the the span rule as a bound on the leave-one-out error of Support Vector Machines (SVMs) to select a optimal subset of the top ranked genes as biomarker and then using these biomarker to construct a classier [1]. Meanwhile, another group also proposed similar methods that using GeneRank algorithm to select top ranked genes as biomarker [9]. Both paper extends their approach further by integrating information of disease associated Transcript Factors (TFs) will help increase the prediction accuracy. The top ranked genes were training in a linear Support Vector Machine (SVM) classier.

```
> r.frsvm <-cv.frsvm( x=x, x.mi=NULL, y=y, #data matrix with factor labels
+ folds=10, repeats=5, #cross-validation parameters.
+ Gsub=ad.matrix, # adjacency matrix of network.
+ top.uper=20,top.lower=100, #optimal the top 100 to 20 ranked genes
+ op=0.9, # the 10% top ranked features
+ aa=1000, pt.pvalue=0.05,# with 1000 times, and the cutoff p value is 0.05
+ parallel=FALSE, cores=2, # parameter for parallel computing
+ d=0.85, # the damping paramter for GeneRank
+ Cs=10^(-3:3), # C parameter for C-SVM
+ DEBUG=TRUE, seed=1234 )
```

3.3 Classification based on hub genes (HubC)

Recently, Taylor et al. [7] found that differentially expressed hub proteins in a protein-protein interaction network could be related to breast cancer disease outcome. We here applied their approach (called HubC).

```
> data(Gs2)
> r.hubC <- cv.hubc( x=x, x.mi=NULL, y=y, #data matrix with factor labels
+ folds=10, repeats=5, #cross-validation parameters.
+ Gsub=ad.matrix, # adjacency matrix of network.
+ op=0.9, # the 10% top ranked features
+ nperm=1000,# permutation test with 1000 times,
+ pt.pvalue=0.05, #and the cutoff p value is 0.05
+ parallel=FALSE, cores=2, # parameter for parallel computing
+ Gs=Gs2, # graph of adjacency matrix Gsub
+ node.ct=0.5, # the cutoff value for top quantile of nodes
+ Cs=10^(-3:3), # C parameter for C-SVM
+ DEBUG=TRUE )
```

3.4 Pathway activity classification (PAC)

Lee et al. [4] proposed PAC that first selects genes within each KEGG-pathway based on a t-test and then summarizes gene expression in each pathway to a pathway activity score. According to the original paper by Lee et al. [4] only the top 10% pathways with highest differences in their activity between sample groups were selected.

```
> library(KEGG.db)
> r.pac <- cv.pac( x=x, x.mi=NULL, y=y, #data matrix with factor labels
+ folds=10, repeats=5, #cross-validation parameters.
+ Gsub=ad.matrix, # adjacency matrix of network.
+ parallel=FALSE, cores=2, # parameter for parallel computing
+ DEBUG=TRUE, seed=1234 )
```

3.5 Average expression profile of pathways (AEP)

Guo et al. ([3]) proposed a methods that use whole KEGG-pathways were selected or not selected based on their average differential expression between patient groups. This was done based on a SAM-test ([8]) with FDR cutoff 5% .

```
> library(KEGG.db)
> r.aep <- cv.aep( x=x, x.mi=NULL, y=y, #data matrix with factor labels
+ folds=10, repeats=5, #cross-validation parameters.
+ Gsub=ad.matrix, # adjacency matrix of network.
+ parallel=FALSE, cores=2, # parameter for parallel computing
+ Cs=10^(-3:3), # C parameter for C-SVM
+ DEBUG=TRUE, seed=1234 )
```

References

[1] Yupeng Cun and Holger Fröhlich. Integrating prior knowledge into prognostic biomarker discovery based on network strucutre. *arXiv preprint arXiv:1212.3214*, 2012.

[2] Yupeng Cun and Holger Fröhlich. Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PloS one*, 8(e73074):9, 2013.

[3] Zheng Guo, Tianwen Zhang, Xia Li, Qi Wang, Jianzhen Xu, Hui Yu, Jing Zhu, Haiyun Wang, Chenguang Wang, Eric J Topol, Qing Wang, and Shaoqi Rao. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6:58, 2005.

[4] Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. Inferring pathway activity toward precise disease classification. *PLoS Comput Biol*, 4(11):e1000217, Nov 2008.

[5] Julie L Morrison, Rainer Breitling, Desmond J Higham, and David R Gilbert. Generank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, 6:233, 2005.

[6] Marcus Schmidt, Daniel Böhm, Christian von Törne, Eric Steiner, Alexander Puhl, Henryk Pilch, Hans-Anton Lehr, Jan G Hengstler, Heinz Kölbl, and Mathias Gehrmann. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res*, 68(13):5405–5413, Jul 2008.

[7] Ian W Taylor, Rune Linding, David Warde-Farley, Yongmei Liu, Catia Pesquita, Daniel Faria, Shelley Bull, Tony Pawson, Quaid Morris, and Jeffrey L Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol*, 27(2):199–204, Feb 2009.

[8] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121, Apr 2001.

[9] Christof Winter, Glen Kristiansen, Stephan Kersting, Janine Roy, Daniela Aust, Thomas Knösel, Petra Rümmele, Beatrix Jahnke, Vera Hentrich, Felix Rückert, et al. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Computational Biology*, 8(5):e1002511, 2012.