R data Package: The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks.

Emmert-Streib F., de Matos Simoes R.,
Mullan P., Haibe-Kains B., Dehmer M.

February 5, 2020

Contents

Abstract

In this vignette, we show how to access and perform basic operations on the gene regulatory network and the processed gene expression data contained in the `BreastCancerGRN R` package. Further we show the preprocessing steps for gene regulatory network inference. The package is a supplementary of `BreastCancerGRN The gene regulatory network for breast cancer: Integrated regulatory landscape of cancer hallmarks`.

1 Introduction

The package contains a preprocessed were inferred using bc3net. The data of the BreastCancerGRN package is a supplementary of [**?**]. The BC3Net [6] algorithm is a bagging approach for C3Net [1, 2]. Briefly, BC3Net consists of 3 major steps. In the first step, a bootstrap ensemble of 100 data sets is generated. For each data set in the ensemble a gene regulatory network is inferred using C3Net. For the network inference, we use a Pearson estimator for mutual information. We apply a multiple testing correction on the inferred edges using Bonferroni. In step two, the resulting ensemble of networks is aggregated into a weighted network, where the weights describe the ensemble consensus rate for an edge. In step three, we apply a binomial test whether or not an edge should be included in the resulting network. We retain only edges for a significance level of $\alpha = 0.05$ that pass a Bonferroni multiple testing correction.

- Gene expression dataset using EntrezID:GeneSymbols identifiers

  ```
  data(data.BC)
  ```

- bc3net BreastCancer gene regulatory network (igraph object)

  ```
  data(net.BC)
  ```

  A detailed description is given below.

2 Data

2.1 Breast Cancer gene expression dataset from ExpO

The data set represents a data subset of the ExpO dataset comprising 351 breast cancer tissue samples that was procesed available matching to EntrezID|GeneSymbol identifiers (`https://www.ncbi.nlm.nih.gov/gene/`).

```
  The data matrix format is:
  str(data.BC)

   num [1:19738, 1:351] 9.84 6.55 5.87 4.99 4.38 ...
 - attr(*, "dimnames")=List of 2
  ..$ : Named chr [1:19738] "DDR1|780" "RFC2|5982"
  "HSPA6|3310" "PAX8|7849" ...
  .. ..- attr(*, "names")= chr [1:19738] "780" "5982" "3310" "7849" ...
  ..$ : chr [1:351] "GSM38051.CEL" "GSM38054.CEL"
  "GSM38057.CEL" "GSM38059.CEL" ...
```

The raw expression dataset in CEL format is available at the ncbi GEO database with accession GSE2109 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109`). An example for the retrieval and preprocessing of the gene expression dataset is shown in the last section of this document.

2.2 Gene regulatory network inferred by bc3net

The GRN object net.BC is an igraph object. *net.BC* is a weighted undirected gene regulatory network inferred from a large-scale gene expression dataset [**?**]. The network comprises a total of 180, 171 interactions for 19738 genes.

```
data(net.BC)

net.BC

IGRAPH UNW- 19738 180171 --
+ attr: name (v/c), weight (e/n)
+ edges (vertex names):
```

```
 [1] PROM2|150696--CDH1|999      PROM2|150696--KRT18|3875
 PROM2|150696--EPCAM|4072
 [4] PROM2|150696--CDS1|1040     PROM2|150696--SPINT2|10653
 PROM2|150696--RAB25|57111
 [7] PROM2|150696--SPINT1|6692   PROM2|150696--ABHD11|83451
 PROM2|150696--KRT7|3855
[10] PROM2|150696--DDR1|780      PROM2|150696--ZEB2|9839
PROM2|150696--RASGRP3|25780
[13] PROM2|150696--LGALS2|3957   PROM2|150696--SYT11|23208
PROM2|150696--PKP3|11187
[16] PROM2|150696--MAL2|114569   JUP|3728    --RAB25|57111
 JUP|3728    --DDR1|780
[19] JUP|3728    --PRRG4|79056   JUP|3728    --LYPD3|27076
JUP|3728    --RHOD|29984
[22] CDH1|999    --EPCAM|4072    CDH1|999    --SPINT2|10653
CDH1|999    --ESRP2|80004
+ ... omitted several edges
```

The GRN was inferred using bc3net from the dataset *data(data.BC)*. Note the following operation requires a memory and can run for a couple of hours. In case memory and time is limited the network inference can be performed on a subset of the data. For example genes with low variability can be excluded from the analysis.

```
# net.BC=bc3net(data.BC, verbatim=TRUE)
```

3 Basic usage and network operations on the GRN

```
library(igraph)
data(net.BC) # igraph bc3net GRN

# node names are defined by entrezID|genesymbol
# and unmapped probeset identifiers
# example first 10 identifiers
V(net.BC)$name[1:10]
```

```
# symmetric adjacency matrix
mat=as.matrix(get.adjacency(net.BC))

# symmetric weighted matrix
mat=as.matrix(get.adjacency(net.BC, attr="weight"))

# degree of top 10 hubgenes
sort(degree(net.BC), decreasing=TRUE)[1:10]

# data.frame of edges and weight vector
bc3.edges=get.edgelist(net.bc3)

# edge weght of the bc3net GRN
# ensemble consenus rate (ecr)
weight=E(net.BC)$weight

# threshold network example for consenus rate >0.1
net=subgraph.edges(net.BC, eids = which(E(net.BC)$weight>0.1))

# igraph to graphNEL format
BC.graphNEL=igraph.to.graphNEL(net.BC)
```

3.1 Retrieval of expression data from ncbi GEO

example1

3.2 Preprocessing of microarray data

example2

3.3 Retrieval of meta information

example3

References

[1] Altay, G. and Emmert-Streib, F., Inferring the conservative causal core of gene regulatory networks. BMC Systems Biology 4 132, 2010

[2] Altay, G. and Emmert-Streib, F., Structural Influence of gene networks on their inference: Analysis of C3NET. Biology Direct 6 31, 2011

[3] Basso, K., Margolin, A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. Reverse engineering of regulatory networks in human B cells. Nat Genet 37 382-90, 2005

[4] Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. R package version 2.9.0., 2013

[5] de Matos Simoes, R., Tripathi, S. and Emmert-Streib F. Organizational structure and the periphery of the gene regulatory network in B-cell lymphoma BMC Systems Biology 2012, 6:38, 2011

[6] de Matos Simoes and Emmert-Streib F. Bagging Statistical Network Inference from Large-Scale Gene Expression Data PLoS ONE 7(3): e33624

[7] de Matos Simoes, R., Dehmer, M. and Emmert-Streib, F. B-cell lymphoma gene regulatory networks:Biological consistency among inference methods. Front Genet. 2013 Dec 16;4:281, 2013

[8] Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al., ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1 S7, 2006

[9] The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks.