Package ‘rhcoclust’

October 14, 2022

Title Robust Hierarchical Co-Clustering to Identify Significant Co-Cluster

Version 1.1.0

Description Here we performs robust hierarchical co-clustering between row and column entities of a data matrix in absence and presence of outlying observations. It can be used to explore important co-clusters consisting of important samples and their regulatory significant features. Please see Hasan, Badsha and Mollah (2020) <doi:10.1101/2020.05.13.094946>.

License GPL (>= 2)

Depends R (>= 3.0.0)

Encoding UTF-8

LazyData true

NeedsCompilation no

Imports fields, grDevices, graphics, igraph, stats

RoxygenNote 7.1.0

Author Md. Bahadur Badsha [aut, cre],
Mohammad Nazmol Hasan [aut],
Md. Nurul Haque Mollah [aut]

Maintainer Md. Bahadur Badsha <mbbadshar@gmail.com>

Repository CRAN

Date/Publication 2021-02-26 10:10:05 UTC

R topics documented:

plot_rhcoclust .................................................. 2
reversestring .................................................... 3
rhcoclust ....................................................... 4
rhcoclust_internet ............................................ 7
rhcoclust_network ............................................. 8
Description

This function used for two plots from output of rhcoclust (i) plot results for gene (row) and compound (column) co-cluster graph, and (ii) plot graph of QCC for identification of biomarker co-cluster.

Usage

```r
plot_rhcoclust(CoClustObj, plot.cocluster = FALSE, plot.ccim = FALSE, cex.xaxis = 0.7, cex.yaxis = 0.5)
```

Arguments

- `CoClustObj`: Output objects from rhcoclust
- `plot.cocluster`: To set no plotting as the default for cocluster.
- `plot.ccim`: To set no plotting as the default for icc.
- `cex.xaxis`: A numerical value giving to control/annotation text size in x-axis. Default is 0.7.
- `cex.yaxis`: A numerical value giving to control/annotation text size in y-axis. Default is 0.5.

Value

Plots

Author(s)

Md. Bahadur Badsha <mbbadshar@gmail.com>

See Also

rhcoclust for generating a graph objects for clustering network
Examples

# Load necessary library
library(rhcoclust)
library(fields)

# Load real data
data("toxygates_data")

# Load predefined simulated data
data("simu_data")

# Data analysis using real data by rhcoclust. Note, for analysis of predefined
# simulated data use data <- simu_data
data <- toxygates_data

# Apply rhcoclust to identify significant co-cluster of samples and their regulatory features
CoClustObj <- rhcoclust(data, rk = 3, ck = 3, method.dist = "manhattan", method.hclust = "ward.D")

# Plot co-cluster
# Please use par(mar=c(6, 10, 3, 6)) or modify if needed for best fit of the graph
# mar order: bottom, left, top, and right
plot_rhcoclust (CoClustObj, plot.coclust = TRUE, plot.ccim = FALSE, cex.xaxis = 0.7, cex.yaxis = 0.5)

# Plot CCIM
# Please use dev.off() to avoid the figure margin from previous plot
plot_rhcoclust (CoClustObj, plot.coclust = FALSE, plot.ccim = TRUE)
# Please add legend with change or add any parameters if needed.
legend("topleft",
    legend = c("Upper-significant", "Insignificant","Down-significant"),
    col = c("red","black","blue"),
    bty = "n",
    pch = c(20,20,20),
    pt.cex = 2,
    cex = 1.2,
    x.intersp = 0.2,
    y.intersp = 0.4,
    text.col = "black",
    horiz = FALSE,
    inset = c(0.3, -0.08))

reversestring(string, n = 1)

Description

This is the function for reverse string

Usage

reversestring(string, n = 1)
Arguments

*string*  
Given string or seq

*n*  
By which n-plets we should reverse the given string

Details

This function is used to reverse given string or seq

Value

Reversed string or seq

Author(s)

Md. Bahadur Badsha <mbbadshar@gmail.com>

Examples

```r
x1 <- c("R1C1", "R2C2", "R3C3")
reversestring(x1, 2)
```

---

**Description**

Toxicogenomic studies require co-clustering to identify biomarker genes for the assessment of chemical toxicity from gene expression levels. It is also essential in the drug discovery experiments. However, gene expression datasets are often contaminated by outliers due to several steps involve in the data generating process. This package performs robust hierarchical co-clustering between row and column entities of a data matrix by reducing the influence of outlying observations. It can be used to explore biomarker genes those are divided into upregulatory and downregulatory groups by the influence of different chemical compounds groups more accurately. It can also provide the statistical significance of the identified co-clusters.

**Usage**

```r
rhcoclust(data, rk, ck, method.dist = "manhattan", method.hclust= "ward.D")
```

**Arguments**

*data*  
A data matrix containing data having the characteristics of interval and ratio level of measurement or continuous data

*rk*  
Number of clusters in the row entities of the data matrix

*ck*  
Number of clusters in the column entities of the data matrix
method.dist  The distance measure to be used. The default is "manhattan". The other options are "euclidean", "maximum", "canberra", "binary" or "minkowski". Any unambiguous substring can be given.

method.hclust  The agglomeration method to be used. The default is "ward.D". The other options are "ward.D2", "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) or "centroid" (= UPGMC).

Value

A list of object that containing the following:

Coclust_MeanMat : A data frame containing combination of row and column cluster number in the first column and their ranked co-cluster mean in the second column. In the first column first number indicates row cluster index and second number indicates column cluster index, respectively.

CoClsDtMat : The reorganized transformed data matrix to generate co-cluster graph.

NG_Cocls : The index of gene/row names.

NC_Cocls : The index of column names.

rowclust : The gene/row entity clusters.

colclust : The column entity clusters.

colorsG: Colors of genes/row entity clusters to generate co-cluster graph.

colorsC: Colors of DCCs/column entity clusters to generate co-cluster graph.

CentralLine: Central Line of individual control chart to generate graph of control chart and to identify significant co-clusters.

UpContLimit: Upper Control Limit to generate graph of control chart and to identify significant co-clusters.

LowrContLimit: Lower Control Limit to generate graph of control chart and to identify significant co-clusters.

color: Colors to generate individual control chart.

pchmark: Shape of points to generate individual control chart.

Author(s)

Md. Bahadur Badsha <mbbadshar@gmail.com>

Examples

# Load necessary library
library(rhcoclust)
library(fields)

# Load real data
data("toxygates_data")

# Load predefined simulated data
data("simu_data")

# Data analysis using real data by rhcoclust. Note, for analysis of predefined
# simulated data use data <- simu_data
data <- toxigates_data

# Apply rhcoclust to identify significant co-cluster of samples and their regulatory features
CoClustObj <- rhcoclust(data, rk=3, ck=3, method.dist = "manhattan", method.hclust = "ward.D")

# A data frame containing combination of row and column cluster number in the first
column and their ranked co-cluster mean in the second cluster.
GC_cls_MeanMat <- CoClustObj$NGC_cls_MeanMat

# The reorganized transformed data matrix to generate co-cluster graph.
CoClsDtMat <- CoClustObj$CoClsDtMat

# The gene/row entity clusters.
rowclust <- CoClustObj$rowclust

# The column entity clusters.
colclust <- CoClustObj$colclust

# Colors of genes/row entity clusters to generate co-cluster graph
colorsG <- CoClustObj$colorsG

# Colors of DCCs/column entity clusters to generate co-cluster graph
colorsC <- CoClustObj$colorsC

# Central Line of individual control chart to generate graph of control chart and to
# identify significant co-clusters.
CntrLine_QC <- CoClustObj$CentralLine

# Upper Control Limit to generate graph of control chart and to identify significant
# co-clusters.
UCL_QC <- CoClustObj$UpContLimit

# Lower Control Limit to generate graph of control chart and to identify significant
# co-clusters.
LCL_QC <- CoClustObj$LowrContLimit

# Colors to generate individual control chart.
ColorQC <- CoClustObj$color

# Shape of points to generate individual control chart.
PcmQC <- CoClustObj$pchmark

# Plot co-cluster
# par(mar=c(6,10,3,6)) # Modify if needed
# mar order: bottom, left, top, and right
plot_rhcoclust (CoClustObj, plot.coclust = TRUE, plot.ccim = FALSE)

# Plot CCIM
# use dev.off() to avoid the figure margin from previous plot
plot_rhcoclust (CoClustObj, plot.coclust = FALSE, plot.ccim = TRUE)
rhcoclust_internet  Interaction network (internet) of the 'rhcoclust' objects

Description

This function is used for visualization of clustering interaction network plot for the objects that are generated by rhcoclust and list of up-regulated and down-regulated variables list. There are three layers in the network plot, (i) genes (rows) are shown in the first layer, (ii) co-cluster are in the second layer and (iii) columns (compounds) are in the last layer. Red and blue color indicates up and down regulated respectively.

Usage

rhcoclust_internet(data, CoClustObj, plot.internet = FALSE, cex.nodes = 0.7, edge.width = 1)

Arguments

data  A data matrix containing data having the characteristics of interval and ratio level of measurement or continuous data

CoClustObj  A list of output objects from rhcoclust

plot.internet  To set no plotting as the default for interaction network (internet)

cex.nodes  A numerical value giving to control/annotation node size in the network. Default is 0.7.

edge.width  A numerical value giving to control/annotation edge width in the network. Default is 1.

Value

Plot A list up-regulated and down-regulated variables.

Author(s)

Md. Bahadur Badsha <mbbadshar@gmail.com>

See Also

rhcoclust for generating a graph objects for clustering network

Examples

# Load necessary library
library(rhcoclust)
library(fields)

# Load real data
data("toxygates_data")
# Load predefined simulated data
data("simu_data")

# Data analysis using real data by rhcoclust. Note, for analysis of predefined
# simulated data
data <- simu_data
# Real data use: data <- toxygates_data

# Apply rhcoclust to identify significant co-cluster of samples and their regulatory features
CoClustObj <- rhcoclust(data, rk=4, ck=3, method.dist = "manhattan", method.hclust = "ward.D")

# Plot interaction network (internet)
# Please use dev.off() to avoid the figure margin from previous plot
# mar order: bottom, left, top, and right
# please use par(mar=c(5,2,5,2)) or modify when necessary to best fit for the plot
Netrhcoclust <- rhcoclust_internet(data, CoClustObj = CoClustObj, plot.internet = TRUE, cex.nodes = 0.7, edge.width = 1)
# Please change or add any parameter if needed.
text(x = -1, y = 1.1, "Row Cluster", cex = 0.7)
# Please change or add any parameter if needed.
text(x = 0, y = 1.1, "Co-Cluster", cex = 0.7)
# Please change or add any parameter if needed.
text(x = 1, y = 1.1, "Column Cluster", cex = 0.7)

**rhcoclust_network**

*Visualization of clustering network plot*

**Description**

This function is used for visualization of clustering network plot, the plot objects are generated by rhcoclust.

**Usage**

```
rhcoclust_network(CoClustObj, scale.threshold = 10)
```

**Arguments**

- `CoClustObj` Output objects from rhcoclust
- `scale.threshold` Edge width scaling threshold, defaults is 10

**Value**

Plot

**Author(s)**

Md. Bahadur Badsha <mbbadshar@gmail.com>
**simulate_data**

*Simulate data for robust hierarchical clustering to identify significant co-cluster*

**Description**

We generate fold change gene expression (FCGE) data according to the characteristics of toxicogenomic data.

**Usage**

```r
simulate_data(no.gene, no.dcc)
```

**Arguments**

- `no.gene` Number of genes in the simulated data.
- `no.dcc` Number of doses of chemical compounds (dcc) in the simulated data.

**See Also**

`rhcoclust` for generating a graph objects for clustering network

**Examples**

```r
# Load necessary library
library(rhcoclust)
library(fields)
library(igraph)

# Load real data
data("toxygates_data")

# Load predefined simulated data
data("simu_data")

# Data analysis using real data by rhcoclust. Note, for analysis of predefined
# simulated data use data <- simu_data
data <- toxygates_data

# Apply rhcoclust to identify significant co-cluster of samples and their regulatory features
CoClustObj <- rhcoclust(data, rk = 3, ck = 3, method.dist = "manhattan", method.hclust = "ward.D")

# Visualization of clustering network plot
rhcoclust_network(CoClustObj, scale.threshold = 10)
```
Details

There are four gene groups and three DCCs groups in the simulated dataset. The gene group 1, 2, 3 and 4 consists the genes G1-G10, G11-G20, G21-G30 and G31-G50 respectively and DCCs group 1, 2 and 3 consists the DCCs C1_High-C5_High-C1_Middle-C5_Middle, C6_High-C10_High-C6_Middle-C10_Middle and C1_Low-C12_Low-C11_Middle-C12_Middle-C11_High-C12_High respectively. Where, G stands for gene and C stands for chemical compound arranged in the row and column of the simulated data matrix respectively. The error term N(0,0.35) from normal distribution with mean 0 and variance 0.35 is added to each element of the simulated dataset. In the simulated dataset the gene group-1 is up-regulated by the DCCs group-1, gene group-2 is up and down-regulated by the DCCs group-2 and 1 respectively. The gene group-3 is down-regulated by the DCs group-2. The gene group-4 is not regulated by any of the DCCs groups and DCCs group-3 does not influence any of the genes in the dataset.

Value

A list of object that containing the following:

SimData: A simulated data matrix as generated.
SimDataRnd: Randomly distributed row and column entity of simulated data.
GCmat: Transformed simulated data.

Author(s)

Md. Bahadur Badsha <mbbadshar@gmail.com>

Examples

# Load library
library(rhcoclust)

# Number of genes in the simulated data.
no.gene <- 50

# Number of doses of chemical compounds (dcc) in the simulated data.
no.dcc <- 12

SimulteData <- simulate_data(no.gene,no.dcc)

# A simulated data matrix as generated.
SimulteData$SimData

# A randomly distributed row and column entity of simulated data.
SimulteData$SimDataRnd

# A transformed simulated data.
SimulteData$GCmat
simu_data

A predefined simulated data for 'rhcoclust' package

Description

This is the predefined simulated data matrix that used as an example in 'rhcoclust' package.

Details

The column of the simulated data matrix represents doses of chemical compounds (DCCs) and row represents genes. Each element in the dataset represents fold change gene expression data.

Value

Matrix

Author(s)

Md Bahadur Badsha <mbbadshar@gmail.com>

See Also

simulate_data

Examples

library(rhcoclust)

# Load data
data("simu_data")

toxygates_data

A real dataset for 'rhcoclust' package

Description

This is the real data matrix that used as an example in 'rhcoclust' package.

Details

The real data matrix collected from Nyström-Persson et al., 2013 (https://toxygates.nibiohn.go.jp/toxygates/#columns). The column of the data matrix represents doses of chemical compounds (DCCs) and row represents genes. Each element in the dataset represents fold change gene expression data.

Value

Matrix
Author(s)

Md. Bahadur Badsha <mbbadshar@gmail.com>

References


Examples

```r
# Load library
library(rhcoclust)

# Load real data
data("toxygates_data")
```
Index

list, 5, 7, 10
plot_rhcoclust, 2
reversestring, 3
rhcoclust, 2, 4, 7, 9
rhcoclust_internet, 7
rhcoclust_network, 8
simu_data, 11
simulate_data, 9, 11
toxygates_data, 11