Package 'PAGI'

February 19, 2015

Version 1.0

Title The package can identify the dysregulated KEGG pathways based on global influence from the internal effect of pathways and crosstalk between pathways.

Author Junwei Han,Yanjun Xu,Haixiu Yang,Chunquan Li and Xia Li

Maintainer Junwei Han <hanjunwei1981@163.com>

Description The package can identify the dysregulated KEGG pathways based on global influence from the internal effect of pathways and crosstalk between pathways. (1) The PAGI package can prioritize the pathways associated with two biological states by statistical significance or FDR. (2) The PAGI package can evaluated the global influence factor (GIF) score in the global gene-gene network constructed based on the relationships of genes extracted from each pathway in KEGG database and the overlapped genes between pathways.

Depends R (>= 2.15.1),igraph

Suggests Matrix

Collate PAGI.1.0.R getExample.R

LazyData Yes

License GPL (>= 2)

biocViews Statistics, crosstalk, Pathways, Graphs, Networks

Repository CRAN

Date/Publication 2013-11-01 08:31:57

NeedsCompilation no

R topics documented:

`CalGIF` *Calculate the global influence factor (GIF)*

Description

CalGIF is an attempt to calculate the GIF score which is used to distinguish the non-equivalence of gene influenced by both internal effect of pathways and crosstalk between pathways. The random walk with restart (RWR) algorithm was used to evaluate the GIF by integrating the global network topology and the correlation of gene with phenotype.

Usage

```
CalGIF(dataset,class.labels)
```

Arguments

`dataset`	A dataframe of gene expression data whose first column are genes symbols and whose names are samples.
`class.labels`	A vector of binary labels. The vector is used to distinguish the class of phenotype.

Details

When users input interesting gene expression data and the vector of binary labels (class labels), the function can calculate the GIF values for all genes in the global gene-gene network constructed based on the relationships of genes extracted from pathway database.

The argument `dataset` is gene expression data set stored in a dataframe. The first column of the dataframe are gene symbols and the names of the dataframe are samples names.

Value

A vector.

Each element is the GIF score and whose name correspond to gene symbol in the gene expression data.

Author(s)

Junwei Han <hanjunwei1981@163.com> Yanjun Xu <tonghua605@163.com> Haixiu Yang <yanghaixiu@ems.hrbmu.edu.cn> Chunquan Li <lcqbio@yahoo.com.cn> and Xia Li <lixia@hrbmu.edu.cn>

Examples

```
## Not run:

###calculate the global influence factor (GIF) by using the random walk with restart (RWR) algorithm###
#example 1
#get example data
dataset<-getdataset()
class.labels<-getclass.labels()

#calculate the global influence factor (GIF)
GIFscore<-CalGIF(dataset,class.labels)
#print the top ten results to screen
GIFscore[rev(order(GIFscore))][1:10]

#Each element is the GIF score and whose name correspond to gene symbol in the gene expression data.
#If the genes in gene expression data are not included in the global gene-gene network, their GIF
#scores will be zero.

#example 2
#get example data
dataset<-read.table(paste(system.file(package="PAGI"),"/localdata/dataset.txt",sep=""),
      header=T,sep="\t","\"")
class.labels<-as.character(read.table(paste(system.file(package="PAGI"),
      "/localdata/class.labels.txt",sep=""),quote="\"", stringsAsFactors=FALSE)[1,])

#calculate the global influence factor (GIF)
GIFscore<-CalGIF(dataset,class.labels)
#print the top ten results to screen
GIFscore[rev(order(GIFscore))][1:10]

#Each element is the GIF score and whose name correspond to gene symbol in the gene expression data.
#If the genes in gene expression data are not included in the global gene-gene network, their
# GIF scores will be zero.


## End(Not run)
```

ExampleData *The variables in the environment variable ExampleData of the system*

Description

The variables in the environment variable ExampleData of the system.

Format

An environment variable

Details

The environment variable includes the variable `dataset`, `class.labels` etc.

Author(s)

Junwei Han <hanjunwei1981@163.com> Yanjun Xu <tonghua605@163.com> Haixiu Yang <yanghaixiu@ems.hrbmu.edu.cn> Chunquan Li <lcqbio@yahoo.com.cn> and Xia Li <lixia@hrbmu.edu.cn>

`getclass.labels` *Get the labels of example dataset*

Description

Get the labels of example dataset.

Usage

```
getclass.labels()
```

Details

The labels of the example data are obtained from the environment variable `ExampleData`.

Value

A character vector of class labels.

Author(s)

Junwei Han <hanjunwei1981@163.com> Yanjun Xu <tonghua605@163.com> Haixiu Yang <yanghaixiu@ems.hrbmu.edu.cn> Chunquan Li <lcqbio@yahoo.com.cn> and Xia Li <lixia@hrbmu.edu.cn>

See Also

`getdataset`

Examples

```
## Not run:

#obtain the labels of the example dataset
class.labels<-getclass.labels()


## End(Not run)
```

getdataset *Get the example dataset*

Description

Get the example dataset.

Usage

```
getdataset()
```

Details

The example data are obtained from the environment variable ExampleData.

Value

A dataframe of p53 status in NCl-60 cell lines.

Author(s)

Junwei Han <hanjunwei1981@163.com> Yanjun Xu <tonghua605@163.com> Haixiu Yang <yang-haixiu@ems.hrbmu.edu.cn> Chunquan Li <lcqbio@yahoo.com.cn> and Xia Li <lixia@hrbmu.edu.cn>

References

Stratton MR. (1992) The p53 gene in human cancer, Eur J Cancer. 1992;28(1):293-5.

See Also

getclass.labels

Examples

```
## Not run:

#obtain the example data
dataset<-getdataset()
head(dataset)


## End(Not run)
```

PAGI.Main *A novel pathway identification approach based on global influence from both the internal effect of pathways and crosstalk between pathways*

Description

PAGI.Main is an attempt to identify dysregulated pathways, which are influenced by both the internal effect of pathways and crosstalk between pathways, integrating pathway topological information and differences between two biological phenotypes.

Usage

```
PAGI.Main(dataset,class.labels,nperm = 100, p.val.threshold = -1, FDR.threshold = 0.01,
 gs.size.threshold.min= 25, gs.size.threshold.max = 500 )
```

Arguments

`dataset`	A dataframe of gene expression data whose first column are genes symbols and whose names are samples.
`class.labels`	A vector of binary labels. The vector is used to distinguish the class of phenotype.
`nperm`	An integer. The number of random permutations. The default value is 100.
`p.val.threshold`	A value. The significance threshold of NOM p-value for pathways whose detail results of pathways to be presented. The default value is -1, which means no threshold.
`FDR.threshold`	A value. The significance threshold of FDR q-value for pathways whose detail results of pathways to be presented. The default value is 0.01.
`gs.size.threshold.min`	An integer. The minimum size (in genes) for pathways to be considered. The default value is 25.
`gs.size.threshold.max`	An integer. The maximum size (in genes) for pathways to be considered. The default value is 500.

Details

When users input interesting gene expression data and the vector of binary labels (class labels), the function can identify dysregulated pathways mainly through: (1) Mapping genes with the absolute t-score more than 0 to the global graph reconstructed based on the relationships of genes extracted from each pathway in KEGG database and the overlapped genes between pathways; (2) We defined a global influence factor (GIF) to distinguish the non-equivalence of gene influenced by both internal effect of pathways and crosstalk between pathways in the global network. The random walk with restart (RWR) algorithm was used to evaluate the GIF by integrating the global network topology and the correlation of gene with phenotype; (3) We used cumulative distribution functions

(CDFs) to prioritize the dysregulated pathways. The permutation is used to identify the stasistical significance of pathways (normal p-values) and the FDR is used to to account for false positives.

The argument `dataset` is gene expression data set stored in a dataframe. The first column of the dataframe are gene symbols and the names of the dataframe are samples names.

Value

A list. It includes two elements: `SummaryResult` and `PathwayList`.

`SummaryResult` is a dataframe. It is the summary of the result of pathways. Each rows of the dataframe represents a pathway. Its columns include "Pathway Name", "SIZE", "PathwayID", "Pathway Score", "NOM p-val", "FDR q-val", "Tag percentage" (Percent of gene set before running enrichment peak), "Gene percentage" (Percent of gene list before running enrichment peak), "Signal strength" (enrichment signal strength).

`PathwayList` is list of pathways which present the detail results of pathways with NOM p-val< `p.val.threshold` or FDR< `FDR.threshold`. Each element of the list is a dataframe. Each rows of the dataframe represents a gene. Its columns include "Gene number in the (sorted) pathway", "gene symbol from the gene express data", "location of the gene in the sorted gene list", "the T-score of gene between two biological states", "global influence impactor", "if the gene contribute to the score of pathway".

Author(s)

Junwei Han <hanjunwei1981@163.com> Yanjun Xu <tonghua605@163.com> Haixiu Yang <yanghaixiu@ems.hrbmu.edu.cn> Chunquan Li <lcqbio@yahoo.com.cn> and Xia Li <lixia@hrbmu.edu.cn>

References

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102, 15545-15550.

Li, C., Li, X., Miao, Y., Wang, Q., Jiang, W., Xu, C., Li, J., Han, J., Zhang, F., Gong, B. et al. (2009) SubpathwayMiner: a software package for flexible identification of pathways. Nucleic Acids Res, 37, e131.

Examples

```
## Not run:

##########identify dysregulated pathways by using the function PAGI.Main############
#example 1
#get example data
dataset<-getdataset()
class.labels<-getclass.labels()

#identify dysregulated pathways
result<-PAGI.Main(dataset,class.labels,nperm = 100,p.val.threshold = -1,FDR.threshold = 0.01,
gs.size.threshold.min = 25, gs.size.threshold.max = 500 )
```

```
#print the summary results of pathways to screen
result[[1]][1:10,]

#The result is a dataframe. The rows of the dataframe are ranked by the values of False
#discovery rate (FDR). Each row of the result (dataframe) is a pathway. It columns include
#"Pathway Name", "SIZE", "PathwayID", "Pathway Score",  "NOM p-val", "FDR q-val", "Tag
#percentage", "Gene percentage", "Signal strength". They correspond to pathway names,
#the number of genes which were mapped to the pathway from gene expression profiles, pathway ID,
#the scores of pathway, the nominal p-values of the pathways, the FDR values, the percent of
#gene set before running enrichment peak, the percent of gene list before running enrichment peak,
#enrichment signal strength.

#print the detail results of pathways to screen
result[[2]][1:5]

#The result is a list. Each element of the list is a dataframe whcih present the detail results of
#genes in the pathway with FDR.threshold< 0.01. Each rows of the dataframe represents a gene.
#Its columns include "Gene number in the (sorted) pathway", "gene symbol from the gene express data",
#"location of the gene in the sorted gene list", "the T-score of gene between two biological states",
#"global influence impactor", "if the gene contribute to the score of pathway".

#write the summary results of pathways to tab delimited file.
write.table(result[[1]], file = "SUMMARY RESULTS.txt", quote=F, row.names=F, sep = "\t")

#write the detail results of genes for each pathway with FDR.threshold< 0.01 to tab delimited file.
for(i in 1:length(result[[2]])){
gene.report<-result[[2]][[i]]
filename <- paste(names(result[[2]][i]),".txt", sep="", collapse="")
write.table(gene.report, file = filename, quote=F, row.names=F, sep = "\t")
}

#example 2
#get example data
dataset<-read.table(paste(system.file(package="PAGI"),"/localdata/dataset.txt",sep=""),
header=T,sep="\t","\"")
class.labels<-as.character(read.table(paste(system.file(package="PAGI"),
"/localdata/class.labels.txt",sep=""),quote="\"", stringsAsFactors=FALSE)[1,])

#identify dysregulated pathways
result<-PAGI.Main(dataset,class.labels,nperm = 100,p.val.threshold = -1,FDR.threshold = 0.01,
gs.size.threshold.min = 25, gs.size.threshold.max = 500 )

#print the summary results of pathways to screen
result[[1]][1:10,]

#The result is a dataframe. The rows of the dataframe are ranked by the values of False
#discovery rate (FDR). Each row of the result (dataframe) is a pathway. It columns include
#"Pathway Name", "SIZE", "PathwayID", "Pathway Score",  "NOM p-val", "FDR q-val", "Tag
#percentage", "Gene percentage", "Signal strength". They correspond to pathway names,
#the number of genes which were mapped to the pathway from gene expression profiles, pathway ID,
#the scores of pathway, the nominal p-values of the pathways, the FDR values, the percent of
#gene set before running enrichment peak, the percent of gene list before running enrichment peak,
#enrichment signal strength.
```

```
#print the detail results of pathways to screen
result[[2]][1:5]

#The result is a list. Each element of the list is a dataframe whcih present the detail results of
#genes in the pathway with FDR.threshold< 0.01. Each rows of the dataframe represents a gene.
#Its columns include "Gene number in the (sorted) pathway", "gene symbol from the gene express data",
#"location of the gene in the sorted gene list", "the T-score of gene between two biological states",
#"global influence impactor", "if the gene contribute to the score of pathway".

#write the summary results of pathways to tab delimited file.
write.table(result[[1]], file = "SUMMARY RESULTS.txt", quote=F, row.names=F, sep = "\t")

#write the detail results of genes for each pathway with FDR.threshold< 0.01 to tab delimited file.
for(i in 1:length(result[[2]])){
gene.report<-result[[2]][[i]]
filename <- paste(names(result[[2]][i]),".txt", sep="", collapse="")
write.table(gene.report, file = filename, quote=F, row.names=F, sep = "\t")
}


## End(Not run)
```

PAGIData *The variables in the environment variable PAGIData of the system*

Description

The variables in the environment variable PAGIData of the system.

Format

An environment variable

Details

The environment variable includes the variable `netWorkdata`, `pathway.db` etc.

Author(s)

Junwei Han <hanjunwei1981@163.com> Yanjun Xu <tonghua605@163.com> Haixiu Yang <yang-haixiu@ems.hrbmu.edu.cn> Chunquan Li <lcqbio@yahoo.com.cn> and Xia Li <lixia@hrbmu.edu.cn>

Index