Package 'ESEA'

February 19, 2015

Version 1.0

Title ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis

Author Junwei Han, Xinrui Shi, Chunquan Li

Maintainer Xinrui Shi <xinrui103@163.com>

Description

The package can identify the dysregulated canonical pathways by investigating the changes of biological relationships of pathways in the context of gene expression data. (1) The ESEA package constructs a background set of edges by extracting pathway structure (e.g. interaction, regulation, modification, and binding etc.) from the seven public databases (KEGG; Reactome; Biocarta; NCI; SPIKE; HumanCyc; Panther) and the edge sets of pathways for each of the above databases. (2) The ESEA package can can quantify the change of correlation between genes for each edge based on gene expression data with cases and controls. (3) The ESEA package uses the weighted Kolmogorov-Smirnov statistic to calculate an edge enrichment score (EES), which reflects the degree to which a given pathway is associated the specific phenotype. (4) The ESEA package can provide the visualization of the results.

Depends R (>= 2.10),igraph,XML,parmigene

Suggests Matrix,graph

Collate calEdgeCorScore.R ESEA.Main.R PlotGlobEdgeCorProfile.R PlotPathwayGraph.R PlotRunEnrichment.R SavePathway2File.R getEnvironmentData.R GetExampleData.R GetEdgesBackgrandData.R GetPathwayEdgeData.R

LazyData Yes

License GPL (>= 2)

biocViews Statistics, Pathways, edge, enrichment analysis

NeedsCompilation no

Repository CRAN

Date/Publication 2015-01-22 15:58:44

R topics documented:

`calEdgeCorScore` *Calculate the differential correlation score for edges*

Description

CalEdgeCorScore attempts to calculate the differential correlation scores of two genes in the edge between the expression data of all samples and control samples.

Usage

```
calEdgeCorScore(dataset, class.labels, controlcharactor, edgesbackgrand)
```

Arguments

`dataset`	A marix of gene expression data whose row names are genes symbols and whose column names are samples.
`class.labels`	A vector of binary labels. The vector is used to distinguish the class of phenotype.
`controlcharactor`	A character string of control sample label.
`edgesbackgrand`	A marix which deposits the data of background set of edges.

Details

For each edge, we estimated the mutual information (MI) between two genes in the expression data of all samples and control samples respectively. The difference of MI between all samples and control samples is used as the differential correlation score of the edge.

Value

A vector. Each element is the differential correlation score of an edge and whose name correspond to edge ID in the background set of edges.

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

References

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics, 7 Suppl 1, S7.

Mani, K.M., Lefebvre, C., Wang, K., Lim, W.K., Basso, K., Dalla-Favera, R. and Califano, A. (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Molecular systems biology, 4, 169.

Examples

```
## Not run:

#get example data
dataset<-GetExampleData("dataset")
class.labels<-GetExampleData("class.labels")
controlcharactor<-GetExampleData("controlcharactor")

#get the data for background set of edges
edgesbackgrand<-GetEdgesBackgrandData()

#Calculate the differential correlation score for edges
EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor, edgesbackgrand)

#print the top ten results to screen
EdgeCorScore[1:10]

#Each element is the differential correlation score of an edge and whose name correspond to
# the edge in the background set of edges.


## End(Not run)
```

EdgesBackgrandData *The data for the background set of edges*

Description

The data for the background set of edges in the environment variable of the system.

Format

An environment variable

Details

The data for the background set of edges are extracted from seven pathway database (KEGG; Biocarta; Reactome; NCI; SPIKE; HumanCyc; Panther).

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

`envData` *The variables in the environment variable* `envData` *of the system*

Description

The variables in the environment variable `envData` of the system.

Format

An environment variable

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

`ESEA.Main` *Identify dysregulated pathways based on edge set enrichment analysis*

Description

A edge-centric method to identify dysregulated pathways by investigating the changes of biological relationships of pathways in the context of gene expression data.

Usage

```
ESEA.Main(EdgeCorScore, pathwayEdge.db, weighted.score.type = 1, pathway = "kegg",
gs.size.threshold.min = 15, gs.size.threshold.max = 1000,
reshuffling.type = "edge.labels", nperm = 100, p.val.threshold = -1,
FDR.threshold = 0.05, topgs = 1)
```

Arguments

EdgeCorScore	A numeric vector. Each element is the differential correlation score of an edge.
pathwayEdge.db	A character vector, the length of it is the number of pathways.
weighted.score.type	A value. Edge enrichment correlation-based weighting: 0=no weight, 1=standard weigth, 2 = over-weigth. The default value is 1
pathway	A character string of pathway database. Should be one of "kegg","reactome", "nci","huamncyc","biocarta","spike" and "panther". The default value is "kegg"
gs.size.threshold.min	An integer. The minimum size (in edges) for pathways to be considered. The default value is 15.
gs.size.threshold.max	An integer. The maximum size (in edges) for pathways to be considered. The default value is 1000.
reshuffling.type	A character string. The type of permutation reshuffling: "edge.labels" or "gene.labels". The default value is "edge.labels".
nperm	An integer. The number of permutation reshuffling. 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 0.05.
topgs	An integer. The number of top scoring gene sets used for detailed reports. The default value is 1.

Details

ESEA integrates pathway structure (e.g. interaction, regulation, modification, and binding etc.) and differential correlation among genes in identifying dysregulated pathways. The biological pathways were collected from the seven public databases (KEGG; Reactome; Biocarta; NCI/Nature Pathway Interaction Database; SPIKE; HumanCyc; Panther). We constructed a background set of edges by extracting pathway structure from each pathway in the seven databases. We then applied an information-theoretic measure, mutual information(MI), to quantify the change of correlation between genes for each edge based on gene expression data with cases and controls. An edge list was formed by ranking the edges according to their changes of correlation. Finally, we used the weighted Kolmogorov-Smirnov statistic to evaluate each pathway by mapping the edges in the pathway to the edge list. The permutation is used to identify the statistical significance of pathways (normal p-values) and the FDR is used to to account for false positives.

Value

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

SummaryResult is a list. It is the summary of the result of pathways which include two elements: the results of Gain-of-correlation and Loss-of-correlation. Each element of the lists is a dataframe.

Each rows of the dataframe represents a pathway. Its columns include "Pathway Name", "Pathway source" "ES", "NES", "NOM p-val", "FDR q-val", "Tag percentage" (Percent of edge set before running enrichment peak), "edge percentage" (Percent of edge 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 or topgs<=topgs.threshold. Each element of the list is a dataframe. Each rows of the dataframe represents an edge. Its columns include "Edge number in the (sorted) pathway", "Edge ID", "location of the edge in the sorted edge list", "EdgeCorScore", "Running enrichment score", "Property of contribution".

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

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 knowledgebased approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102, 15545-15550.

Examples

```
## Not run:

#get example data
dataset<-GetExampleData("dataset")
class.labels<-GetExampleData("class.labels")
controlcharactor<-GetExampleData("controlcharactor")

#get the data for background set of edges
edgesbackgrand<-GetEdgesBackgrandData()

#get the edge sets of pathways
pathwayEdge.db<-GetPathwayEdgeData()

#calculate the differential correlation score for edges
EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor, edgesbackgrand)

#identify dysregulated pathways by using the function ESEA.Main
Results<-ESEA.Main(
EdgeCorScore,
pathwayEdge.db,
weighted.score.type = 1,
pathway = "kegg",
gs.size.threshold.min = 15,
gs.size.threshold.max = 1000,
reshuffling.type = "edge.labels",
nperm =10,
p.val.threshold=-1,
FDR.threshold = 0.05,
```

```
topgs =1
)

#print the summary results of pathways to screen
Results[[1]][[1]][1:5,]

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

#write the summary results of pathways to tab delimited file.
write.table(Results[[1]][[1]], file = "kegg-SUMMARY RESULTS Gain-of-correlation.txt", quote=F,
 row.names=F, sep = "\t")
write.table(Results[[1]][[2]], file = "kegg-SUMMARY RESULTS Loss-of-correlation.txt", quote=F,
 row.names=F, sep = "\t")

#write the detail results of genes for each pathway with FDR.threshold< 0.05 to tab delimited file.
for(i in 1:length(Results[[2]])){
PathwayList<-Results[[2]][[i]]
filename <- paste(names(Results[[2]][i]),".txt", sep="", collapse="")
write.table(PathwayList, file = filename, quote=F, row.names=F, sep = "\t")
}


## End(Not run)
```

ExampleData *The example data in the environment variable of the system*

Description

The example data in the environment variable of the system.

Format

An environment variable

Details

The example data includes the variable `dataset`, `class.labels`,`controlcharactor`,`Results`.

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

GetEdgesBackgrandData *Get the data for background set of edges*

Description

Get the data for background set of edges.

Usage

```
GetEdgesBackgrandData()
```

Details

The data for background set of edges are obtained from the environment variable envData.

Value

A marix which deposits the data of background set of edges. Each row includes a pair of genes which correspond to an edge.

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

Examples

```
## Not run:

#obtain the data for background set of edges.
edgesbackgrand<-GetEdgesBackgrandData()


## End(Not run)
```

GetExampleData *Get the example data*

Description

Get the example data.

Usage

```
GetExampleData(exampleData)
```

Arguments

exampleData	A character string, must be one of "dataset", "class.labels" and "controlcharactor".

Details

The function GetExampleData(exampleData="dataset") obtains gene expression dataset from the environment variable envData.

The function GetExampleData(exampleData="class.labels") obtains class labels from the environment variable envData.

The function GetExampleData(exampleData="controlcharactor") obtains control sample label from the environment variable envData.

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

Examples

```
## Not run:

#obtain the gene expression dataset.
dataset<-GetExampleData(exampleData="dataset")
dataset[1:10,]

#obtain the class labels.
class.labels<-GetExampleData(exampleData="class.labels")

#obtain the control sample label.
controlcharactor<-GetExampleData(exampleData="controlcharactor")


## End(Not run)
```

GetPathwayEdgeData *Get the edge sets of pathways*

Description

Get the edge sets of pathways for the seven pathway database (KEGG; Biocarta; Reactome; NCI; SPIKE; HumanCyc; Panther).

Usage

```
GetPathwayEdgeData()
```

Details

The edge sets of pathways for the seven pathway database (KEGG; Biocarta; Reactome; NCI; SPIKE; HumanCyc; Panther) are obtained from the environment variable envData.

Value

A character vector, the length of this vector is the number of pathways.

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

Examples

```
## Not run:

#obtain the edge sets of pathways.
pathwayEdge.db<-GetPathwayEdgeData()


## End(Not run)
```

PathwayEdgeData *The edge sets of pathways*

Description

The edge sets of pathways for the seven pathway database (KEGG; Biocarta; Reactome; NCI; SPIKE; HumanCyc; Panther) in the environment variable of the system.

Format

An environment variable

Details

For each pathway in the seven pathway database (KEGG; Biocarta; Reactome; NCI; SPIKE; HumanCyc; Panther), the edges are extracted respectively. The edge sets of pathways are then created.

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

PlotGlobEdgeCorProfile

Plot global edge correlation profile

Description

plot global edge correlation profile for differential correlation scores of edges

Usage

```
PlotGlobEdgeCorProfile(EdgeCorScore)
```

Arguments

EdgeCorScore	A numeric vector. Each element is the differential correlation score of an edge.

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

Examples

```
## Not run:

#get example data
dataset<-GetExampleData("dataset")
class.labels<-GetExampleData("class.labels")
controlcharactor<-GetExampleData("controlcharactor")

#get the data for background set of edges
edgesbackgrand<-GetEdgesBackgrandData()

#calculate the differential correlation score for edges
EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor, edgesbackgrand)

#plot global edge correlation profile
PlotGlobEdgeCorProfile(EdgeCorScore)


## End(Not run)
```

`PlotPathwayGraph` *Plot the pathway-result network diagram*

Description

Plot the pathway-result network diagram, the edges which contribute to the pathway enrichment score are marked with red.

Usage

```
PlotPathwayGraph(graph, margin = 0, vertex.label.cex = 0.6, vertex.label.font = 1,
vertex.size = 8, vertex.size2 = 6, vertex.shape = "rectangle", layout = layout.random,
vertex.label.color = "black", edge.color = "dimgray", vertex.color = "#C1FFC1",
vertex.frame.color = "dimgray", axes = FALSE, xlab = "", ylab = "", sub = NULL,
main = NULL, ...)
```

Arguments

`graph`	A dataframe of pathway result obtained from the ESEA.main function
`margin`	A numeric. The value is usually between -0.5 and 0.5, which is able to zoom in or out a pathway graph. The default is 0.
`vertex.label.cex`	A numeric vector of node label size.
`vertex.label.font`	A numeric vector of label font.
`vertex.size`	A numeric vector of Node size. See `plot.igraph`
`vertex.size2`	A numeric vector of Node size.
`vertex.shape`	A vector of node shape. The default is graphics_type.
`layout`	A matrix of x-y coordinates with two dims. Determine the placement of the nodes for drawing a graph.The default is layout.random.
`vertex.label.color`	A vector of node label colors. The default is black.
`vertex.color`	A vector of node colors. The default is the KEGG node color.
`vertex.frame.color`	A vector of node frame color. The default is dimgray.
`edge.color`	A vector of edge color. The default is dimgray.
`axes`	A logical. whether to plot axes. The default is FALSE.
`xlab`	A character string. The label of the horizontal axis. The default is the empty string.
`ylab`	A character string. The label of the vertical axis. The default is the empty string.
`sub`	A character string of subtitle.
`main`	A character string of main title.
`...`	The arguments passed to or from methods. See `plot.igraph` and see `plot`.

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

Examples

```
## Not run:

#get example data
dataset<-GetExampleData("dataset")
class.labels<-GetExampleData("class.labels")
controlcharactor<-GetExampleData("controlcharactor")

#get the data for background set of edges
edgesbackgrand<-GetEdgesBackgrandData()

#get the edge sets of pathways
pathwayEdge.db<-GetPathwayEdgeData()

#calculate the differential correlation score for edges
EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor,edgesbackgrand)

#identify dysregulated pathways by using the function ESEA.Main
#Results<-ESEA.Main(EdgeCorScore,pathwayEdge.db)
Results<-GetExampleData("PathwayResult")

#obtain the detail results of genes for a significant pathway
PathwayNetwork<-Results[[2]][[1]]

#Plot the pathway-result network diagram, the edges which contribute to the ES are labeled with red.
PlotPathwayGraph(PathwayNetwork,layout=layout.random)


## End(Not run)
```

PlotRunEnrichment *Plot running Edge enrichment score*

Description

Plot running edge enrichment score for the pathway result

Usage

```
PlotRunEnrichment(EdgeCorScore, PathwayResult, weighted.score.type = 1)
```

Arguments

EdgeCorScore	A numeric vector. Each element is the differential correlation score of an edge.
PathwayResult	A list of pathway result obtained from the ESEA.main function
weighted.score.type	A value. Edge enrichment correlation-based weighting: 0=no weight, 1=standard weigth, 2 = over-weigth. The default value is 1

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

Examples

```
## Not run:

#get example data
dataset<-GetExampleData("dataset")
class.labels<-GetExampleData("class.labels")
controlcharactor<-GetExampleData("controlcharactor")

#get the data for background set of edges
edgesbackgrand<-GetEdgesBackgrandData()

#get the edge sets of pathways
pathwayEdge.db<-GetPathwayEdgeData()

#calculate the differential correlation score for edges
EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor,edgesbackgrand)

#identify dysregulated pathways by using the function ESEA.Main
#Results<-ESEA.Main(EdgeCorScore,pathwayEdge.db)
Results<-GetExampleData("PathwayResult")

#obtain the detail results of genes for a significant pathway
PathwayResult<-Results[[2]][1]

#Plot running edge enrichment score for the pathway result
PlotRunEnrichment(EdgeCorScore,PathwayResult,weighted.score.type = 1)


## End(Not run)
```

SavePathway2File — *Save a pathway-result network to a file which can be input to the Cytoscape software*

Description

Save a pathway-result network to a file which can be input to the Cytoscape software.

Usage

```
SavePathway2File(network, layout = layout.random, name = "network", file = "Graph")
```

Arguments

network	A dataframe of pathway-result network obtained from the ESEA.main function
layout	A matrix of x-y coordinates with two dims. Determine the placement of the nodes for drawing a graph.The default is "layout.random".
name	The required variables for XGMML.destription. The default is "network"
file	A character string for file name. The default is "Graph"

Author(s)

Junwei Han <hanjunwei1981@163.com>, Xinrui Shi<xinrui103@163.com> and Chunquan Li <lcqbio@163.com>

Examples

```
## Not run:

#get example data
dataset<-GetExampleData("dataset")
class.labels<-GetExampleData("class.labels")
controlcharactor<-GetExampleData("controlcharactor")

#get the data for background set of edges
edgesbackgrand<-GetEdgesBackgrandData()

#get the edge sets of pathways
pathwayEdge.db<-GetPathwayEdgeData()

#calculate the differential correlation score for edges
EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor,edgesbackgrand)

#identify dysregulated pathways by using the function ESEA.Main
#Results<-ESEA.Main(EdgeCorScore,pathwayEdge.db)
Results<-GetExampleData("PathwayResult")

#obtain the detail results of genes for a significant pathway
PathwayNetwork<-Results[[2]][[1]]

#save the pathway-result network to a file which can be input to the Cytoscape software.
SavePathway2File(PathwayNetwork,layout=layout.circle,file="Graph")


## End(Not run)
```

Index