

Package ‘KEGGprofile’

December 11, 2014

Type Package

Title An annotation and visualization package for multi-types and multi-groups expression data in KEGG pathway

Version 1.8.0

Date 2014-09-15

Author Shilin Zhao, Yu Shyr

Maintainer Shilin Zhao <shilin.zhao@vanderbilt.edu>

Description KEGGprofile is an annotation and visualization tool which integrated the expression profiles and the function annotation in KEGG pathway maps. The multi-types and multi-groups expression data can be visualized in one pathway map. KEGGprofile facilitated more detailed analysis about the specific function changes inner pathway or temporal correlations in different genes and samples.

License GPL (>= 2)

LazyLoad yes

Imports AnnotationDbi,png,TeachingDemos,XML,KEGG.db,KEGGREST,biomaRt

biocViews Pathways

R topics documented:

| | |
|-----------------------------------|----|
| col_by_value | 2 |
| convertId | 3 |
| download_KEGGfile | 4 |
| download_latest_pathway | 4 |
| find_enriched_pathway | 5 |
| newIdMatrix | 6 |
| parse_XMLfile | 7 |
| pho_sites_count | 8 |
| plot_pathway | 8 |
| plot_pathway_cor | 9 |
| plot_profile | 10 |
| pro_pho_expr | 12 |

| | |
|--------------|-----------|
| Index | 13 |
|--------------|-----------|

| | |
|--------------|---------------------|
| col_by_value | <i>col_by_value</i> |
|--------------|---------------------|

Description

The function will transfer a numeric matrix into a matrix of colors, in which the colors represent the values of numeric matrix

Usage

```
col_by_value(x, col, range = NA, breaks = NA, showColorBar = T)
```

Arguments

| | |
|--------------|---|
| x | a numeric matrix |
| col | colors used to represent the values. (See also 'Details') |
| range | values out of the range will be modified to in the range. |
| breaks | a numeric vector of three or more cut points giving the number of intervals into which x is to be cut. See also 'Details' |
| showColorBar | Logical. Indicates display the colorbar or not. The default value is TRUE. |

Details

A colorbar would also be plotted. The returned colors of the function can be used in function plot_profile. if breaks not equal to NA, col must have the same length with breaks-1.

Value

a matrix equal to x, but the values were instead by colors.

Examples

```
data(pho_sites_count)
col<-col_by_value(pho_sites_count,col=colorRampPalette(c(white,khaki2))(4),breaks=c(0,1,4,10,Inf))
```

| | |
|-----------|------------------|
| convertId | <i>convertId</i> |
|-----------|------------------|

Description

A function to convert ID based on the biomaRt package.

Usage

```
convertId(x, dataset = "hsapiens_gene_ensembl",
  filters = "uniprot_swissprot_accession", attributes = c(filters,
  "entrezgene"), genesKept = c("foldchange", "first", "random", "var", "abs"),
  keepNoId = T, keepMultipleId = F, verbose = F)
```

Arguments

| | |
|----------------|---|
| keepMultipleId | Logical. Indicate keep the multiple target IDs related to one source ID or not. |
| keepNoId | Logical. Indicate keep the source IDs without target IDs or not. |
| verbose | Logical. Indicate report extra information on progress or not. |
| filters | Filters (one or more) that should be used in the query. A possible list of filters can be retrieved using the function listFilters. |
| attributes | Attributes you want to retrieve. A possible list of attributes can be retrieved using the function listAttributes. |
| dataset | Dataset you want to use. To see the different datasets available within a biomaRt you can e.g. do: mart = useMart('ensembl'), followed by listDatasets(mart). |
| x | the expression data matrix. |
| genesKept | The method to select target gene in more than one targets. "var"/"foldchange"/"abs" means selecting the gene with largest variation/fold change/absolute value. "first" means selecting the first target and "random" means randomly selection. |

Details

A function to convert ID based on the biomaRt package..

Examples

```
temp<-cbind(rnorm(10),rnorm(10))
row.names(temp)<-c("Q04837", "P0C0L4", "P0C0L5", "O75379", "Q13068", "A2MYD1", "P60709", "P30462", "P30475", "P30479")
colnames(temp)<-c("Exp1", "Exp2")
convertId(temp, filters="uniprot_swissprot_accession", keepMultipleId=TRUE)
## Not run:
temp<-cbind(rnorm(5000),rnorm(5000),rnorm(5000),rnorm(5000),rnorm(5000),rnorm(5000))
row.names(temp)<-1000:5999
colnames(temp)<-c("Control1", "Control2", "Control3", "Treatment1", "Treatment2", "Treatment3")
convertId(temp, filters="entrezgene", attributes =c("entrezgene", "uniprot_swissprot_accession"), keepNoId=FALSE)

## End(Not run)
```

download_KEGGfile *download_KEGGfile*

Description

The function download XML files and png files from KEGG website to local disk

Usage

```
download_KEGGfile(pathway_id = "00010", species = "hsa",
                  target_dir = getwd())
```

Arguments

| | |
|------------|---|
| pathway_id | the KEGG pathway id, such as '00010' |
| species | the species id in KEGG database, 'hsa' means human, 'mmu' means mouse, 'rno' means rat, etc |
| target_dir | the local directory where the downloaded files are saved |

Details

If pathway_id is set as 'all', all KEGG pathway ids in KEGG.db package will be used and downloaded from KEGG website

Examples

```
download_KEGGfile(pathway_id="00010",species=hsa)
```

download_latest_pathway
download_latest_pathway

Description

The function will download the latest pathway gene link from KEGG website.

Usage

```
download_latest_pathway(species)
```

Arguments

| | |
|---------|---|
| species | the species id in KEGG database, 'hsa' means human, 'mmu' means mouse, 'rno' means rat, etc |
|---------|---|

Details

The function will download the latest pathway gene link from KEGG website.

Value

a list with two parts

name keggpathway2gene

description a list with the genes for each pathway

name pathway2name

description a list with the names for each pathway

Examples

```
## Not run: download_latest_pathway(species="hsa")
```

```
find_enriched_pathway find_enriched_pathway
```

Description

The function will map the genes in KEGG pathway database, and then hypergeometric tests would be used to estimate the significance of enrichment for each pathway

Usage

```
find_enriched_pathway(gene, species = "hsa", returned_pvalue = 0.01,  
  returned_adjpvalue = 0.05, returned_genenumber = 5,  
  download_latest = FALSE)
```

Arguments

gene a numeric matrix

returned_pvalue

the minimum p value for enriched pathways

returned_adjpvalue

the minimum adjusted p value for enriched pathways

returned_genenumber

the minimum number of annotated genes for enriched pathways

download_latest

logical. Indicate if the function will download the latest pathway/gene link from KEGG website. As the KEGG.db package was not updated for a long time due to the KEGG policy change, we provided this parameter so that the users could get the latest KEGG database.

species

the species id in KEGG database, 'hsa' means human, 'mmu' means mouse, 'rno' means rat, etc

Details

Only the pathways with p value \leq returned_pvalue in hypergeometric tests and number of annotated genes \geq returned_genenumber would be taken as enriched and returned.

Value

a list with two parts

| | |
|--------------|--|
| name stastic | description a matrix containing the pathway IDs of enriched pathways, and their names, p values, number of annotated genes |
| name detail | description a list with the genes annotated for each pathway |

Examples

```
data(pho_sites_count)
#the 300 genes with most phosphorylation sites quantified
genes<-names(rev(sort(pho_sites_count[,1]))[1:300])
pho_KEGGresult<-find_enriched_pathway(genes,species=hsa)
```

newIdMatrix

newIdMatrix

Description

A function to convert ID.

Usage

```
newIdMatrix(x, convertIdTable, genesKept = c("var", "foldchange", "abs",
      "first", "random"))
```

Arguments

| | |
|----------------|---|
| x | the expression data matrix. |
| convertIdTable | A vector. The names should be the source IDs, and the values should be the target IDs. |
| genesKept | The method to select target gene in more than one targets. "var"/"foldchange"/"abs" means selecting the gene with largest variation/fold change/absolute value. "first" means selecting the first target and "random" means randomly selection. |

Details

A function to convert ID.

Examples

```
convertIdTable<-paste("New",c(1,2,2,2,1,3,4,4,5,5))
names(convertIdTable)<-paste("Old",1:length(convertIdTable))
temp<-matrix(rnorm(20),ncol=2)
row.names(temp)<-names(convertIdTable)
colnames(temp)<-c("Exp1","Exp2")
newIdMatrix(temp,genesKept="foldchange",convertIdTable)
```

parse_XMLfile

parse_XMLfile

Description

The function parses KEGG XML (KGML) files

Usage

```
parse_XMLfile(pathway_id, species, database_dir = getwd())
```

Arguments

| | |
|--------------|---|
| database_dir | the directory where the XML files and png files are located |
| pathway_id | the KEGG pathway id, such as '00010' |
| species | the species id in KEGG database, 'hsa' means human, 'mmu' means mouse, 'rno' means rat, etc |

Details

This function will parse the KEGG XML (KGML) file. Then a matrix with genes in this pathway and related infomations will be returned. This matrix can be used for plot the expression profiles on the pathway figure.

Value

a matrix containing genes in this pathway, and their names, locations etc, which could be used in the function plot_profile as param KEGG_database

Examples

```
XML2database<-parse_XMLfile(pathway_id="04110",species="hsa",database_dir=system.file("extdata",package="KEGGp
```

| | |
|-----------------|---|
| pho_sites_count | <i>number of phosphorylation sites quantified for each gene</i> |
|-----------------|---|

Description

This data set is a data.frame with number of phosphorylation sites quantified for each gene in the analysis.

Usage

```
pho_sites_count
```

Source

Olsen, J.V., et al. (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis, *Sci Signal*, 3, ra3.

| | |
|--------------|---------------------|
| plot_pathway | <i>plot_pathway</i> |
|--------------|---------------------|

Description

A wrapper for function `download_KEGGfile`, `parse_XMLfile` and `plot_profile`

Usage

```
plot_pathway(gene_expr, line_col, groups, pathway_id = "00010",
             species = "hsa", pathway_min = 5, database_dir = getwd(), ...)
```

Arguments

| | |
|--------------|---|
| ... | any other Arguments for function <code>plot_profile</code> |
| gene_expr | the matrix for gene expression, row.names should be NCBI gene ID, such as 67040, 93683 |
| line_col | line color for expression in different samples in the pathway map, valid when <code>type='lines'</code> |
| groups | a character used to indicate expression values from different types of samples |
| species | the species id in KEGG database, 'hsa' means human, 'mmu' means mouse, 'rno' means rat, etc |
| pathway_min | The pathways with number of annotated genes less than <code>pathway_min</code> would be ignored |
| database_dir | the directory where the XML files and png files are located |
| pathway_id | the KEGG pathway id, such as '00010' |

Details

This wrapper function is developed to make the visualization process more easier. Firstly the existence of XML file and png file would be checked, if not, the `download_KEGGfile` function would be used to download the files. Then the `parse_XMLfile` function would be used to parse the XML file. At last the `plot_profile` function would be used to generate the pathway map.

See Also

[download_KEGGfile](#), [parse_XMLfile](#), [plot_profile](#)

Examples

```
data(pro_pho_expr)
data(pho_sites_count)
#type=lines
col<-col_by_value(pho_sites_count,col=colorRampPalette(c(white,khaki2))(4),breaks=c(0,1,4,10,Inf))
temp<-plot_pathway(pro_pho_expr,bg_col=col,line_col=c("brown1","seagreen3"),groups=c(rep("Proteome",6),rep("P
#type=bg
pho_expr<-pro_pho_expr[,7:12]
temp<-apply(pho_expr,1,function(x) length(which(is.na(x))))
pho_expr<-pho_expr[which(temp==0),]
col<-col_by_value(pho_expr,col=colorRampPalette(c(green,black,red))(1024),range=c(-6,6))
temp<-plot_pathway(pho_expr,type="bg",bg_col=col,text_col="white",magnify=1.2,species=hsa,database_dir=system.
#Compound and gene data
set.seed(124)
testData1<-rbind(rnorm(6),rnorm(6),rnorm(6),rnorm(6),rnorm(6),rnorm(6),rnorm(6),rnorm(6))
row.names(testData1)<-c("4967","55753","1743","8802","47","50","cpd:C15972","cpd:C16255")
colnames(testData1)<-c("Control0","Control2","Control5","Sample0","Sample2","Sample5")
temp<-plot_pathway(testData1,type="lines",line_col=c("brown1","seagreen3"),groups=c(rep("Control",3),rep("Samp
testData2<-testData1[,4:6]-testData1[,1:3]
col<-col_by_value(testData2,col=colorRampPalette(c(green,black,red))(1024),range=c(-2,2))
temp<-plot_pathway(testData2,type="bg",bg_col=col,text_col="white",magnify=1.2,species=hsa,database_dir=system
```

plot_pathway_cor

plot_pathway_cor

Description

The function will plot the correlation distributions for each enriched pathway (result from `find_enriched_pathway` function), and then Wilcoxon tests would be used to estimate the significance of correlations distribution between genes in each pathway and all genes.

Usage

```
plot_pathway_cor(gene_expr, kegg_enriched_pathway, groups = NULL,
  side = c("both", "pos", "neg"), alternative = NULL)
```

Arguments

| | |
|-----------------------|---|
| kegg_enriched_pathway | The returned value from find_enriched_pathway function, the enriched pathways. |
| side | a character string specifying the correlation directions interested, must be one of "both" (default), "pos" or "neg". |
| alternative | a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter. |
| gene_expr | the matrix for gene expression, row.names should be NCBI gene ID, such as 67040, 93683 |
| groups | a character used to indicate expression values from different types of samples |

Value

p values for Wilcoxon tests in each pathway

Examples

```
data(pro_pho_expr)
data(pho_sites_count)
genes<-row.names(pho_sites_count)[which(pho_sites_count>=10)]
pho_KEGGresult<-find_enriched_pathway(genes, species=hsa)
result<-plot_pathway_cor(gene_expr=pro_pho_expr, kegg_enriched_pathway=pho_KEGGresult)
```

plot_profile

plot_profile

Description

The function plot gene expression profiles on KEGG pathway maps

Usage

```
plot_profile(gene_expr, pathway_name, result_name = paste(pathway_name,
  "_profile_", type, ".png", sep = ""), KEGG_database, groups,
  bg_col = "white", text_col = "black", line_col, border_col = "grey",
  text_cex = 0.25, magnify = 1, type = c("lines", "bg"),
  pathway_min = 5, genes_kept = c("foldchange", "first", "random", "var",
  "abs"), species = "hsa", database_dir = getwd(), max_dist, lwd = 1.2)
```

Arguments

| | |
|--------------|--|
| gene_expr | the matrix for gene expression, row.names should be NCBI gene ID, such as 67040, 93683 |
| pathway_name | the species id and KEGG pathway id, such as 'hsa00010' |

| | |
|---------------|--|
| KEGG_database | the matrix returned by function <code>parse_XMLfile</code> , which contains genes in this pathway, and their names, locations etc |
| result_name | the name of figure file generated by <code>KEGGprofile</code> . The default name is <code>pathway_name+'_profile_'+type+'.png'</code> , such as <code>'hsa04110_profile_lines.png'</code> |
| groups | a character used to indicate expression values from different types of samples |
| bg_col | background color for gene rectangles in the pathway map |
| line_col | line color for expression in different samples in the pathway map, valid when <code>type='lines'</code> |
| text_col | the colors for text in the pathway map. A color matrix generated by function <code>col_by_value</code> can be used here |
| border_col | border color for gene rectangles in the pathway map. A color matrix generated by function <code>col_by_value</code> can be used here |
| text_cex | cex for text in the pathway map. A color matrix generated by function <code>col_by_value</code> can be used here |
| magnify | the coefficient used to magnify the gene rectangles |
| type | the type of pathway map visualization, could be <code>'bg'</code> or <code>'lines'</code> . Default is <code>'bg'</code> . See also <code>'Details'</code> |
| pathway_min | The pathways with number of annotated genes less than <code>pathway_min</code> would be ignored |
| genes_kept | methods used for choosing genes when several genes corresponding to one location in pathway map. Default is <code>'foldchange'</code> , which kept the gene with largest fold changes. <code>'first'</code> kept the first gene. <code>'random'</code> chose gene random. <code>'var'</code> kept the gene with largest variation. <code>'abs'</code> kept the gene with largest absolute value |
| max_dist | The expression changes that represented by the distance from the bottom to the top of gene rectangle, valid when <code>type='lines'</code> . This param is used to ensure the dynamic changes of lines in different gene polygon represent equal variation. It would be calculated from the maximum changes of genes in this pathway by default. If <code>max_dist=NA</code> , then the lines would be plotted from top to bottom in each gene rectangle |
| lwd | The line width when <code>type='lines'</code> |
| species | the species id in KEGG database, <code>'hsa'</code> means human, <code>'mmu'</code> means mouse, <code>'rno'</code> means rat, etc |
| database_dir | the directory where the XML files and png files are located |

Details

There are two visualization methods to represent gene expression profiles: `'background'` and `'lines'`. The first one is applicable for analysis with only one sample or one type of data, which divides the gene polygon into several sub-polygons to represent different time points. And each sub-polygon has a specific background color to represent expression changes in that time point. The second method plots lines with different colors in the gene polygon to represent different samples or different types of data. The dynamic changes of lines mean the profiles of genes in different time points.

Value

a matrix containing genes mapped in this pathway, and their names, expressions

Examples

```
XML2database<-parse_XMLfile(pathway_id="04110",species="hsa",database_dir=system.file("extdata",package="KEGGp  
data(pro_pho_expr)  
temp<-plot_profile(pro_pho_expr,pathway_name="hsa04110",KEGG_database=XML2database,line_col=c("brown1","seagre
```

pro_pho_expr

expression profiles in proteome and phosphoproteome

Description

This data set is from a previously published data of proteome and phosphoproteome analysis in different cell phase. The column 1-6 are proteome data and column 7-12 are phosphoproteome data in this data.frame. The 6 time points are G1, G1/S, Early S, Late S, G2, Mitosis.

Usage

```
pro_pho_expr
```

Source

Olsen, J.V., et al. (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis, *Sci Signal*, 3, ra3.

Index

`col_by_value`, [2](#), [11](#)

`convertId`, [3](#)

`download_KEGGfile`, [4](#), [9](#)

`download_latest_pathway`, [4](#)

`find_enriched_pathway`, [5](#)

`newIdMatrix`, [6](#)

`parse_XMLfile`, [7](#), [9](#)

`pho_sites_count`, [8](#)

`plot_pathway`, [8](#)

`plot_pathway_cor`, [9](#)

`plot_profile`, [9](#), [10](#)

`pro_pho_expr`, [12](#)