Package 'pathfindR'

August 18, 2018

Type Package

Title Pathway Enrichment Analysis Utilizing Active Subnetworks

Version 1.2.1

Author Ege Ulgen, Ozan Ozisik

Maintainer Ege Ulgen <egeulgen@gmail.com>

Description Pathway enrichment analysis enables researchers to uncover mechanisms underlying the phenotype. pathfindR is a tool for pathway enrichment analysis utilizing active subnetworks. It identifies active subnetworks in a protein-protein interaction network using user-provided a list of genes. It performs pathway enrichment analyses on the identified subnetworks. pathfindR also offers functionalities to cluster enriched pathways and identify representative pathways and to score the pathways per sample. The method is described in detail in Ulgen E, Ozisik O, Sezerman OU. 2018. pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks. bioRxiv. <doi:10.1101/272450>.

License MIT + file LICENSE

URL https://github.com/egeulgen/pathfindR

BugReports https://github.com/egeulgen/pathfindR/issues

Encoding UTF-8

LazyData true

SystemRequirements Java JVM 1.8

Imports AnnotationDbi, DBI, doParallel, foreach, rmarkdown, org.Hs.eg.db, ggplot2, fpc, pathview

Depends R (>= 3.4), knitr, shiny

RoxygenNote 6.1.0

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2018-08-17 22:10:09 UTC

R topics documented:

`biocarta_genes` *BioCarta Gene Sets*

Description

A list containing the genes involved in each human BioCarta pathway. Each element is a vector of gene symbols located in the given pathway. This data was retrieved on May 13, 2018.

Usage

```
biocarta_genes
```

Format

list containing 217 vectors of gene symbols. Each vector corresponds to a gene set.

---

biocarta_pathways *BioCarta Pathway Descriptions*

Description

A list containing the descriptions for each human Reactome pathway. This data was retrieved on May 13, 2018.

Usage

```
biocarta_pathways
```

Format

list containing 217 character values, the descriptions for the given pathways.

---

calculate_pwd *Calculate Pairwise Distances Between Given Pathways*

Description

Calculate Pairwise Distances Between Given Pathways

Usage

```
calculate_pwd(pathway_ids, agg_method = "average",
  plot_heatmap = FALSE)
```

Arguments

pathway_ids	Vector of IDs of pathways selected to be clustered.
agg_method	the agglomeration method to be used if plotting heatmap (see next argument, default: average).
plot_heatmap	boolean value indicating whether or not to plot the heat map of pathway clustering (default: FALSE).

Details

See "Chen, Y. A. et al. Integrated pathway clusters with coherent biological themes for target prioritisation. PLoS One 9, e99030, doi:10.1371/journal.pone.0099030 (2014)." for details on the method of pathway clustering.

Value

Pairwise distance matrix. Optionally plots a heatmap of pathway clustering.

See Also

`hclust` for hierarchical clustering, `heatmap` for drawing a heatmap.

Examples

```
calculate_pwd(RA_output$ID)
```

---

calculate_pw_scores *Calculate Pathway Scores for Each Subject*

Description

Calculate Pathway Scores for Each Subject

Usage

```
calculate_pw_scores(pw_table, exp_mat, cases = NULL, plot_hmap = TRUE)
```

Arguments

`pw_table`	a data frame that must contain the 3 columns below: **Pathway** Description of the enriched pathway **Up_regulated** the up-regulated genes in the input involved in the given pathway, comma-separated **Down_regulated** the down-regulated genes in the input involved in the given pathway, comma-separated
`exp_mat`	the gene expression/methylation matrix. Columns are samples and rows are genes. Column names must contain sample names and row names must contain the gene symbols.
`cases`	(Optional) A vector of sample names that are cases in the case/control experiment.
`plot_hmap`	Boolean value to indicate whether or not to draw the heatmap plot of the scores. (default = TRUE)

Value

Matrix of pathway scores per sample. Columns are samples, rows are pathways. Optionally, displays a heatmap of this matrix.

Examples

```
score_matrix <- calculate_pw_scores(RA_output, RA_exp_mat, plot_hmap = FALSE)
```

choose_clusters *Cluster Pathways and Partition the Dendrogram*

Description

This function first calculates the pairwise distances between the pathways in the result_df data frame. Next, using this distance matrix, the pathways are clustered via hierarchical clustering. By default, the average silhouette width for each possible number of clusters is calculated. The optimal number of clusters is selected as the one with the highest average silhouette width. The dendrogram is cut into this optimal number of clusters, and the pathways with the lowest p value within each cluster are chosen as representative pathways. If 'auto == FALSE", the user can manually select at which height to cut the dendrogram via a shiny application. See "Chen, Y. A. et al. Integrated pathway clusters with coherent biological themes for target prioritisation. PLoS One 9, e99030, doi:10.1371/journal.pone.0099030 (2014)." for details on the method of pathway clustering.

Usage

```
choose_clusters(result_df, auto = TRUE, agg_method = "average",
  plot_heatmap = FALSE, plot_dend = FALSE)
```

Arguments

result_df	data frame of enriched pathways. Must-have columns are: 1. IDKEGG ID of the enriched pathway 2. lowest_pthe lowest adjusted-p value of the given pathway over all iterations 3. highest_pthe highest adjusted-p value of the given pathway over all iterations
auto	boolean value indicating whether to select the optimal number of clusters automatically. If FALSE, a shiny application is displayed, where the user can manually partition the clustering dendrogram (default: TRUE).
agg_method	the agglomeration method to be used if plotting heatmap. Must be one of "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median" or "centroid" (default: average).
plot_heatmap	boolean value indicating whether or not to plot the heat map of pathway clustering (default: FALSE).
plot_dend	boolean value indicating whether or not to plot the dendrogram partitioned into the optimal number of clusters, shown by red rectangles (default: FALSE)

Value

If 'auto' is FALSE, manual partitioning can be performed. Via a shiny HTML document, the hierarchical clustering dendrogram is visualized. In this HTML document, the user can select the agglomeration method and the distance value at which to cut the tree. The resulting cluster assignments of the pathways along with annotation of representative pathways (chosen by smallest lowest p value) are presented as a table and this table can be saved as a csv file. If 'auto' is TRUE,

automatic partitioning of clusters is performed. The function adds 2 additional columns to the input data frame and returns it:

Cluster the cluster to which the pathway is assigned

Status whether the pathway is the "Representative" pathway in its cluster or only a "Member"

See Also

See `calculate_pwd` for calculation of pairwise distances between enriched pathways. See `hclust` for more information on hierarchical clustering. See `run_pathfindR` for the wrapper function of the pathfindR enrichment workflow.

Examples

```
choose_clusters(RA_output)
```

---

custom_result *Custom Gene Set Enrichment Results*

Description

A data frame consisting of pathfindR enrichment results on the example TF target data.

Usage

```
custom_result
```

Format

data frame containing 2 rows and 8 columns. Each row is a gene set (the TF target gene sets).

---

enrichment *Perform Enrichment Analysis*

Description

Perform Enrichment Analysis

Usage

```
enrichment(genes_by_pathway, genes_of_interest, pathways_list,
  adj_method = "bonferroni", enrichment_threshold, pin_path, DEG_vec)
```

Arguments

genes_by_pathway	List that contains genes for each pathway. Names of this list are KEGG IDs.
genes_of_interest	The set of gene symbols to be used for enrichment analysis. In the scope of this package, these are genes that were identified for an active subnetwork.
pathways_list	List that contains pathway descriptions for KEGG pathway IDs. Names of this list are KEGG IDs.
adj_method	correction method to be used for adjusting p-values.
enrichment_threshold	adjusted-p value threshold used when filtering pathway enrichment results
pin_path	path to the Protein-Protein Interaction Network (PIN) file used in the analysis
DEG_vec	vector of differentially-expressed gene symbols

Value

A data frame that contains enrichment results.

See Also

p.adjust for adjustment of p values. See run_pathfindR for the wrapper function of the pathfindR workflow.

Examples

```
pin_path <- return_pin_path("KEGG")
enrichment(kegg_genes, c("PER1", "PER2", "CRY1", "CREB1"), kegg_pathways,
           "bonferroni", 0.05, pin_path, c("PER1"))
```

---

enrichment_chart *Plot the Bubble Chart of Enrichment Results*

Description

This function is used to plot a bubble chart displaying the enrichment results.

Usage

```
enrichment_chart(result_df, plot_by_cluster = FALSE)
```

Arguments

`result_df` a data frame that must contain the following columns:

Pathway Description of the enriched pathway

Fold_Enrichment Fold enrichment value for the enriched pathway

lowest_p the lowest adjusted-p value of the given pathway over all iterations

Up_regulated the up-regulated genes in the input involved in the given pathway, comma-separated

Down_regulated the down-regulated genes in the input involved in the given pathway, comma-separated

Cluster(OPTIONAL) the cluster to which the pathway is assigned

`plot_by_cluster`

boolean value indicating whether or not to group the pathways by cluster (works if "Cluster" is a column of ‘result_df‘).

Value

a ‘ggplot2‘ object containing the bubble chart. The x-axis corresponds to fold enrichment values while the y-axis indicates the enriched pathways. Size of the bubble indicates the number of DEGs in the given pathway. Color indicates the -log10(lowest-p) value. The closer the color is to red, the more significant the enrichment is. Optionally, if "Cluster" is a column of ‘result_df‘ and plot_by_cluster == TRUE, the pathways are grouped by clusters.

Examples

```
g <- enrichment_chart(RA_output)
```

`go_bp_genes` *Gene Ontology - Biological Process Ontology Gene Sets*

Description

A list containing the genes involved in each GO Biological Process. Each element is a vector of gene symbols located in the given gene set. This data was retrieved on May 13, 2018.

Usage

```
go_bp_genes
```

Format

list containing 3941 vectors of gene symbols. Each vector corresponds to a gene set.

go_bp_pathways *Gene Ontology - Biological Process Ontology Descriptions*

Description

A list containing the descriptions for each human GO Biological Process. This data was retrieved on May 13, 2018.

Usage

```
go_bp_pathways
```

Format

list containing 3941 character values, the descriptions for the given pathways.

go_cc_genes *Gene Ontology - Cellular Component Ontology Gene Sets*

Description

A list containing the genes involved in each GO Cellular Component. Each element is a vector of gene symbols located in the given gene set. This data was retrieved on May 13, 2018.

Usage

```
go_cc_genes
```

Format

list containing 470 vectors of gene symbols. Each vector corresponds to a gene set.

go_cc_pathways *Gene Ontology - Cellular Component Ontology Descriptions*

Description

A list containing the descriptions for each human GO Cellular Component. This data was retrieved on May 13, 2018.

Usage

```
go_cc_pathways
```

Format

list containing 470 character values, the descriptions for the given pathways.

go_mf_genes *Gene Ontology - Molecular Function Ontology Gene Sets*

Description

A list containing the genes involved in each GO Molecular Function. Each element is a vector of gene symbols located in the given gene set. This data was retrieved on May 13, 2018.

Usage

```
go_mf_genes
```

Format

list containing 713 vectors of gene symbols. Each vector corresponds to a gene set.

go_mf_pathways *Gene Ontology - Molecular Function Ontology Descriptions*

Description

A list containing the descriptions for each human GO Molecular Function. This data was retrieved on May 13, 2018.

Usage

```
go_mf_pathways
```

Format

list containing 713 character values, the descriptions for the given pathways.

`input_processing` *Process Input*

Description

Process Input

Usage

```
input_processing(input, p_val_threshold, pin_path)
```

Arguments

`input`	the input data that pathfindR uses. The input must be a data frame with three columns: 1. Gene Symbol (HGNC Gene Symbol) 2. Change value, e.g. log(fold change) 3. adjusted p value associated with test, e.g. differential expression/methylation
`p_val_threshold`	the adjusted-p value threshold to use when filtering the input data frame
`pin_path`	path to the Protein Interaction Network (PIN) file used in the analysis

Value

This function first filters the input so that all p values are less than or equal to the threshold. Next, gene symbols that are not found in the PIN are identified. If aliases of these gene symbols are found in the PIN, the symbols are converted to the corresponding aliases. The resulting data frame containing the original gene symbols, the updated symbols, change values and p values is then returned.

See Also

See `run_pathfindR` for the wrapper function of the pathfindR workflow

Examples

```
## Not run:
input_processing(RA_input, 0.05, return_pin_path("KEGG"))

## End(Not run)
```

input_testing *Input Testing*

Description

Input Testing

Usage

```
input_testing(input, p_val_threshold)
```

Arguments

input	the input data that pathfindR uses. The input must be a data frame with three columns: 1. Gene Symbol (HGNC Gene Symbol) 2. Change value, e.g. log(fold change) 3. adjusted p value associated with test, e.g. differential expression/methylation
p_val_threshold	the adjusted-p value threshold to use when filtering the input data frame. Must a numeric value between 0 and 1.

Value

Only checks if the input and the threshold follows the required specifications.

See Also

See run_pathfindR for the wrapper function of the pathfindR workflow

Examples

```
input_testing(RA_input, 0.05)
```

kegg_genes *KEGG Gene Sets*

Description

A list containing the genes involved in each human KEGG pathway. Each element is a vector of gene symbols located in the given pathway. Names correspond to the KEGG ID of the pathway. Pathways that did not contain any genes were discarded. This data was retrieved on May 13, 2018.

Usage

```
kegg_genes
```

Format

list containing 319 vectors of gene symbols. Each vector corresponds to a pathway.

---

kegg_pathways *KEGG Pathway Descriptions*

Description

A list containing the descriptions for each human KEGG pathway. Names of the list correspond to the KEGG ID of the pathway. Pathways that did not contain any genes were discarded. This data was retrieved on May 13, 2018.

Usage

```
kegg_pathways
```

Format

list containing 319 character values, the descriptions for the given pathways.

---

parseActiveSnwSearch *Parse Active Subnetwork Search Output File*

Description

Parse Active Subnetwork Search Output File

Usage

```
parseActiveSnwSearch(active_snw_path, signif_genes, score_thr = 3,
  sig_gene_thr = 2)
```

Arguments

active_snw_path	path to the output of an Active Subnetwork Search.
signif_genes	the vector of significant genes.
score_thr	active subnetwork score threshold (Default = 3)
sig_gene_thr	threshold for minimum number of significant genes (Default = 2)

Value

A list of genes in every active subnetwork that has a score > 3 and that has at least 2 significant genes.

See Also

See run_pathfindR for the wrapper function of the pathfindR workflow

Examples

```
## Not run:
parseActiveSnwSearch("path/to/output", significant_genes)

## End(Not run)
```

---

pathfindR *pathfindR: A package for Pathway Enrichment Analysis Utilizing Active Subnetworks*

Description

The pathfindR package provides two important functions: run_pathfindR and choose_clusters.

run_pathfindR

This function is the wrapper function for the pathfindR workflow. It takes in a data frame consisting of Gene Symbol, log-fold-change and adjusted-p values. After input testing, any gene symbols that are not in the PIN are converted to alias symbols if the alias is in the PIN. Next, active subnetwork search is performed. Pathway enrichment analysis is performed using the genes in each of the active subnetworks. Pathways with adjusted-p values lower than enrichment_threshold are discarded. The lowest adjusted-p value (over all subnetworks) for each pathway is kept. This process of active subnetwork search and enrichment is repeated for a selected number of iterations, which is executed in parallel. Over all iterations, the lowest and the highest adjusted-p values, as well as number of occurrences are reported for each enriched pathway.

choose_clusters

This function first calculates the pairwise distances between the pathways in the result_df data frame. Via a shiny HTML document, the hierarchical clustering dendrogram is visualized. In this HTML document, the user can select the value at which to cut the tree and the resulting representative pathways (chosen by smallest lowest p value) are presented as a table and pathways with cluster assignments can be saved as a csv file.

See Also

See run_pathfindR and choose_clusters for more details.

pathmap *Annotate Involved Genes In Pathways and Visualize Pathways*

Description

Annotate Involved Genes In Pathways and Visualize Pathways

Usage

```
pathmap(pw_table, gene_data)
```

Arguments

pw_table	Data frame of enrichment results. Must-have columns are: "ID" and "Pathway".
gene_data	Single column data frame containing change values (e.g. log(fold change) values) for significant genes. Row names are gene symbols.

Value

The function returns the input data frame with genes involved in each pathway. Added columns are: "Up_regulated" and "Down_regulated", the up- and down-regulated genes, respectively. The function also creates visualizations of the pathways with the package `pathview` and saves them in the folder "pathway_maps" under the current working directory.

See Also

`pathview` for pathway-based data integration and visualization. See `run_pathfindR` for the wrapper function of the pathfindR workflow

Examples

```
## Not run:
pathmap(pathway_table, change_data)

## End(Not run)
```

plot_scores *Plot the Heatmap of Pathway Scores*

Description

Plot the Heatmap of Pathway Scores

Usage

```
plot_scores(score_matrix, cases = NULL)
```

Arguments

score_matrix	Matrix of pathway scores per sample. Columns are samples, rows are pathways.
cases	(Optional) A vector of sample names that are cases in the case/control experiment.

Value

A ‘ggplot2‘ object containing the heatmap plot. x-axis indicates the samples. y-axis indicates the pathways. "Pathway Score" indicates the pathway score of a sample. If ‘cases‘ are provided, the plot is divided into 2 facets: "Case" and "Control".

Examples

```
score_mat <- calculate_pw_scores(RA_output, RA_exp_mat, plot_hmap = FALSE)
hmap <- plot_scores(score_mat)
```

RA_clustered — *Example Output for the pathfindR Clustering Workflow - Rheumatoid Arthritis*

Description

A dataset containing the results of pathfindR's pathway clustering and partitioning workflow performed on the rheumatoid arthritis enrichment results `RA_output`. The number of clusters were detected automatically as 11 and the agglomeration method was "average".

Usage

```
RA_clustered
```

Format

A data frame with 33 rows and 8 columns:

ID KEGG ID of the enriched pathway

Pathway Description of the enriched pathway

Fold_Enrichment Fold enrichment value for the enriched pathway

occurrence the number of iterations that the given pathway was found to enriched over all iterations

lowest_p the lowest adjusted-p value of the given pathway over all iterations

highest_p the highest adjusted-p value of the given pathway over all iterations

Up_regulated the up-regulated genes in the input involved in the given pathway, comma-separated

Down_regulated the down-regulated genes in the input involved in the given pathway, comma-separated

Cluster the cluster to which the pathway is assigned

Status whether the pathway is the "Representative" pathway in its cluster or only a "Member"

See Also

RA_input for example input of the enrichment workflow. RA_output for example output of the enrichment workflow.

RA_exp_mat *Example Input for pathfindR - pathway z-scores*

Description

A matrix containing the log2-normalized expression values of the differentially-expressed genes for 18 rheumatoid arthritis (RA) patients and 15 healthy subjects. Expression values of differentially-expressed genes with adj.P.Val <= 0.05 are presented in this dataset.

Usage

```
RA_exp_mat
```

Format

A matrix with 572 rows and 33 columns.

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15573

RA_input *Example Input for the pathfindR Enrichment Workflow - Rheumatoid Arthritis*

Description

A dataset containing the differentially-expressed genes along with the associated log2-fold-change values and adjusted p-values for the GEO dataset GSE15573. The microarray dataset aimed to characterize gene expression profiles in the peripheral blood mononuclear cells of 18 rheumatoid arthritis (RA) patients versus 15 healthy subjects. Differentially-expressed genes with adj.P.Val < 0.05 are presented in this dataset.

Usage

```
RA_input
```

Format

A data frame with 572 rows and 3 variables:

Gene.symbol HGNC gene symbols of the differentially-expressed genes

logFC log2-fold-change values

adj.P.Val adjusted p values, via the Benjamini & Hochberg (1995) method

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15573

See Also

RA_output for example output of the enrichment workflow. RA_clustered for example output of the clustering workflow.

RA_output *Example Output for the pathfindR Enrichment Workflow - Rheumatoid Arthritis*

Description

A dataset containing the results of pathfindR's active-subnetwork-oriented pathway enrichment workflow performed on the rheumatoid arthritis differential-expression dataset RA_input. Active subnetwork search was performed with Greedy Algorithm using the Biogrid PIN.

Usage

```
RA_output
```

Format

A data frame with 33 rows and 8 columns:

ID KEGG ID of the enriched pathway

Pathway Description of the enriched pathway

Fold_Enrichment Fold enrichment value for the enriched pathway

occurrence the number of iterations that the given pathway was found to enriched over all iterations

lowest_p the lowest adjusted-p value of the given pathway over all iterations

highest_p the highest adjusted-p value of the given pathway over all iterations

Up_regulated the up-regulated genes in the input involved in the given pathway, comma-separated

Down_regulated the down-regulated genes in the input involved in the given pathway, comma-separated

See Also

RA_input for example input of the enrichment workflow. RA_clustered for example output of the clustering workflow.

`reactome_genes` *Reactome Gene Sets*

Description

A list containing the genes involved in each human Reactome pathway. Each element is a vector of gene symbols located in the given pathway. Names correspond to the Reactome ID of the pathway. This data was retrieved on May 13, 2018.

Usage

```
reactome_genes
```

Format

list containing 2022 vectors of gene symbols. Each vector corresponds to a pathway.

`reactome_pathways` *Reactome Pathway Descriptions*

Description

A list containing the descriptions for each human Reactome pathway. Names of the list correspond to the Reactome ID of the pathway. This data was retrieved on May 13, 2018.

Usage

```
reactome_pathways
```

Format

list containing 2022 character values, the descriptions for the given pathways.

return_pin_path *Return The Path to Given Protein-Protein Interaction Network (PIN)*

Description

This function returns the path/to/PIN.sif. While the default PINs are Biogrid, GeneMania, IntAct and KEGG, the user can choose to use any other PIN by specifying the path/to/PIN.sif. All PINs to be used in this workflow must have 3 columns with no header and be tab-separated. Columns 1 and 3 must be interacting proteins' HGNC gene symbols, column 2 must be a column with all rows consisting of "pp".

Usage

```
return_pin_path(pin_name_path = "Biogrid")
```

Arguments

pin_name_path	Name of the chosen PIN or path/to/PIN.sif. If PIN name, must be one of c("Biogrid", "GeneMania", "IntAct", "KEGG"). If path/to/PIN.sif, the file must comply with the PIN specifications. Defaults to "Biogrid".

Value

A character value that contains the path to chosen PIN.

See Also

See run_pathfindR for the wrapper function of the pathfindR workflow

Examples

```
pin_path <- return_pin_path("Biogrid")
pin_path <- return_pin_path("KEGG")
```

run_pathfindR *Wrapper Function for pathfindR Workflow*

Description

run_pathfindR is the wrapper function for the pathfindR workflow

Usage

```
run_pathfindR(input, p_val_threshold = 0.05,
  enrichment_threshold = 0.05, adj_method = "bonferroni",
  search_method = "GR", use_all_positives = FALSE, saTemp0 = 1,
  saTemp1 = 0.01, saIter = 10000, gaPop = 400, gaIter = 10000,
  gaThread = 5, gaMut = 0, grMaxDepth = 1, grSearchDepth = 1,
  grOverlap = 0.5, grSubNum = 1000, iterations = 10,
  n_processes = NULL, pin_name_path = "Biogrid", score_thr = 3,
  sig_gene_thr = 2, gene_sets = "KEGG", custom_genes = NULL,
  custom_pathways = NULL, bubble = TRUE,
  output_dir = "pathfindR_Results", list_active_snw_genes = FALSE,
  silent_option = TRUE)
```

Arguments

input	the input data that pathfindR uses. The input must be a data frame with three columns: 1. Gene Symbol (HGNC Gene Symbol) 2. Change value, e.g. log(fold change) 3. adjusted p value associated with test, e.g. differential expression/methylation
p_val_threshold	the adjusted-p value threshold to use when filtering the input data frame. Must a numeric value between 0 and 1.
enrichment_threshold	threshold used when filtering individual pathway enrichment results
adj_method	correction method to be used for adjusting p-values of pathway enrichment results (Default: 'bonferroni')
search_method	algorithm to use when performing active subnetwork search. Options are greedy search (GR), simulated annealing (SA) or genetic algorithm (GA) for the search (Default:GR. Can be one of c("GR", "SA", "GA"))
use_all_positives	if TRUE: in GA, adds an individual with all positive nodes. In SA, initializes candidate solution with all positive nodes. (Default = FALSE)
saTemp0	initial temperature for SA (Default: 1.0)
saTemp1	final temperature for SA (Default: 0.01)
saIter	iteration number for SA (Default: 10000)
gaPop	population size for GA (Default: 400)
gaIter	iteration number for GA (Default: 10000)
gaThread	number of threads to be used in GA (Default: 5)
gaMut	the mutation rate for GA (Default: 0)
grMaxDepth	sets max depth in greedy search. set to 0 for no limit (Default: 1)
grSearchDepth	sets search depth in greedy search (Default: 1)
grOverlap	sets overlap threshold for results of greedy search (Default: 0.5)

`grSubNum` sets number of subnetworks to be presented in the results (Default: 1000)

`iterations` number of iterations for active subnetwork search and enrichment analyses (Default = 10. Gets set to 1 for Genetic Algorithm)

`n_processes` optional argument for specifying the number of processes used by foreach. If not specified, the function determines this automatically (Default == NULL. Gets set to 1 for Genetic Algorithm)

`pin_name_path` Name of the chosen PIN or path/to/PIN.sif. If PIN name, must be one of c("Biogrid", "GeneMania", "IntAct", "KEGG"). If path/to/PIN.sif, the file must comply with the PIN specifications. Defaults to "Biogrid".

`score_thr` active subnetwork score threshold (Default = 3)

`sig_gene_thr` threshold for minimum number of significant genes (Default = 2)

`gene_sets` the gene sets to be used for enrichment analysis. Available gene sets are KEGG, Reactome, BioCarta, GO-BP, GO-CC, GO-MF or Custom. If "Custom", the arguments custom_genes and custom pathways must be specified. (Default = "KEGG")

`custom_genes` a list containing the genes involved in each custom pathway. Each element is a vector of gene symbols located in the given pathway. Names correspond to the ID of the pathway.

`custom_pathways`
A list containing the descriptions for each custom pathway. Names of the list correspond to the ID of the pathway.

`bubble` boolean value. If TRUE, a bubble chart displaying the enrichment results is plotted. (default = TRUE)

`output_dir` the directory to be created under the current working directory where the output and intermediate files are saved (default: "pathfindR_Results")

`list_active_snw_genes`
boolean value indicating whether or not to report the non-DEG active subnetwork genes for the active subnetwork which was enriched for the given pathway with the lowest p value (default = FALSE)

`silent_option` boolean value indicating whether or not to print to the console (FALSE) or print to a file (TRUE) during active subnetwork search (default = TRUE)

Details

This function takes in a data frame consisting of Gene Symbol, log-fold-change and adjusted-p values. After input testing, any gene symbols that are not in the PIN are converted to alias symbols if the alias is in the PIN. Next, active subnetwork search is performed. Pathway enrichment analysis is performed using the genes in each of the active subnetworks. Pathways with adjusted-p values lower than `enrichment_threshold` are discarded. The lowest adjusted-p value (over all subnetworks) for each pathway is kept. This process of active subnetwork search and enrichment is repeated for a selected number of `iterations`, which is done in parallel. Over all iterations, the lowest and the highest adjusted-p values, as well as number of occurrences are reported for each enriched pathway.

Value

Data frame of pathfindR enrichment results. Columns are:

ID KEGG ID of the enriched pathway

Pathway Description of the enriched pathway

Fold_Enrichment Fold enrichment value for the enriched pathway

occurrence the number of iterations that the given pathway was found to enriched over all iterations

lowest_p the lowest adjusted-p value of the given pathway over all iterations

highest_p the highest adjusted-p value of the given pathway over all iterations

non_DEG_Active_Snw_Genes (OPTIONAL) the non-DEG active subnetwork genes, comma-separated

Up_regulated the up-regulated genes in the input involved in the given pathway, comma-separated

Down_regulated the down-regulated genes in the input involved in the given pathway, comma-separated

The function also creates an HTML report with the pathfindR enrichment results linked to the visualizations of the pathways in addition to the table of converted gene symbols. This report can be found in "'output_dir'/results.html" under the current working directory.

Optionally, a bubble chart of enrichment results are plotted. The x-axis corresponds to fold enrichment values while the y-axis indicates the enriched pathways. Size of the bubble indicates the number of DEGs in the given pathway. Color indicates the -log10(lowest-p) value; the more red it gets, the more significant the pathway is.

Warning

Depending on the protein interaction network of your choice, active subnetwork finding component of pathfindR may take a very long time to finish.

See Also

`input_testing` for input testing, `input_processing` for input processing, `parseActiveSnwSearch` for parsing an active subnetwork search output, `enrichment` for pathway enrichment analysis and `pathmap` for annotation of involved genes and visualization of pathways. See `foreach` for details on parallel execution of looping constructs. See `choose_clusters` for clustering the resulting enriched pathways and partitioning into clusters.

Examples

```
## Not run:
run_pathfindR(RA_input)

## End(Not run)
```

Index