Package 'NetAct'

April 19, 2023

Type Package

Title NetAct -- a computational platform to construct core transcription factor regulatory networks using gene activity

Version 1.0.6

Author Kenong Su, Vivek Kohar, Danya Gordin, Mingyang Lu

Maintainer Yukai You <you.yu@northeastern.edu>, Danya Gordin <Danya.Gordin@gmail.com>

Description NetAct R package for gene network construction and modeling.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

VignetteBuilder knitr

NeedsCompilation yes

LinkingTo Rcpp

Imports Rcpp, edgeR, Biobase, qvalue, fgsea, reshape2, fastmatch, methods, mclust, entropy, infotheo, circlize, visNetwork, limma, DESeq2, org.Mm.eg.db, org.Hs.eg.db, AnnotationDbi, ComplexHeatmap

Suggests R.utils, knitr, sRACIPE

RoxygenNote 7.2.3

R topics documented:

`Activity_heatmap` *Plotting TF gene expresion & activity heatmap*

Description

Plotting TF gene expresion & activity heatmap

Usage

```
Activity_heatmap(new_activity, eset)
```

Arguments

`new_activity`	Matrix. TF activity matrix
`eset`	ExpressionSet of gene expression data

Value

Heatmap plotting object

allNet *Compile a GSDB to a matrix with 2 columns*

Description

Compile a GSDB to a matrix with 2 columns

Usage

```
allNet(GSDB)
```

Arguments

GSDB	List of list. Gene set database of interactions

Value

matrix. Matrix, each row containing regulators ("from"), targets ("to")

applyDPI *Apply data processing inequality*

Description

Remove the interactions from a triangle which have lowest interaction score.

Usage

```
applyDPI(tfLinks = tfLinks, miMat = miMat, miDiff = 0, minMiTh = 0.5)
```

Arguments

tfLinks	Data.frame. containing the interactions as source (character), target (character), type (integer).
miMat	numeric matrix. Interaction scores based on mutual information or correlation.
miDiff	numeric (0-1). Default 0.0 (optional) Minimum difference between mutual information of a traingle for the edge to be removed.
minMiTh	numeric (0-1). Default 0.5. Minimum value of MI for an interaction which will not be removed.

Value

data.frame. containing the filtered interactions.

calculateMI *Calculate mutual information*

Description

Mutual information between all pairs based on entropy package.

Usage

```
calculateMI(actMat = actMat, nbins = 16, method = "2d")
```

Arguments

actMat	numeric matrix.
nbins	integer (optional). Number of bins Default 16
method	MI calculation method: "2d": 2D discretization with entropy (default) "1d": 1D discretization with infotheo

Value

numeric matrix (0-1). Matrix containing mutual information values

cal_activity *The core function to compute the activity profile of an TF*

Description

The core function to compute the activity profile of an TF

Usage

```
cal_activity(
  gs_remain,
  tmp_data,
  tmp_sign,
  ind,
  with_weight,
  DE_weights,
  tf_exprs,
  useCorSign = useCorSign
)
```

Arguments

gs_remain	a vector of target genes after filtering
tmp_data	gene expression of target genes
tmp_sign	sign of target genes (+1 for one group, -1 for the other)
ind	Hill coefficient parameter used in the weighting factors (default: 1/5, recommend to use 0 < ind < 1/4)
with_weight	whether weighting factors (based on DEG p-values) are used to compute TF activity (default: TRUE)
DE_weights	a vector of the input for computing DE weighting factors (typically, adjusted p-values from DEG analysis)
tf_exprs	a vector of gene expression of the TF
useCorSign	allow the option to use the TF gene expression correlation to flip signs (default: TRUE)

Value

a list of results: activity: matrix of TF activity. sign: grouping scheme of all TF gene sets.

DEG_Analysis_Micro *Helper Function For DEG Analysis of microArray Data (for a single comparison)*

Description

Helper Function For DEG Analysis of microArray Data (for a single comparison)

Usage

```
DEG_Analysis_Micro(eset, qval = 0.05)
```

Arguments

eset	Processed gene expression data in the ExpressionSet format batch & experimental conditions are provided in pData.
qval	q-value cutoff for DEG analysis (default: 0.05)

Value

DEG result in the format of a list containing: table: table of DEG results. rank_vector: a vector of t-statistics for every gene. degs: a vector of gene symbols for DEGs.

filterDB *Filtered gene set database based on minimum sizes*

Description

Filtered gene set database based on minimum sizes

Usage

```
filterDB(GSDB, geneList, minSize = 5)
```

Arguments

GSDB	list of list. gene set database
geneList	a vector of available genes
minSize	minimum number of genes of a gene set (default: 5)

Value

DB: list of list. filtered gene set database

getAdjacencyMat *Obtain the adjacency matrix from a matrix of tf-target relationships*

Description

Obtain the adjacency matrix from a matrix of tf-target relationships

Usage

```
getAdjacencyMat(tfLinks = tfLinks)
```

Arguments

tfLinks	matrix. Matrix of tf-target relationships

Value

adjMat, matrix. adjacency matrix

GSEA_permut_R *Compute ES scores for Gene Set Enrichment Analysis (GSEA) with a new permutation method (using the original GSEA algorithm)*

Description

The function uses the original GSEA enrichment score calculation but using the new permutation method. Here, the gene symbols/names are permutated without changing the ranking vector (stats_vector).

Usage

```
GSEA_permut_R(sim_all, gs, stats_vector)
```

Arguments

sim_all	a matrix of permutated gene lists
gs	a vector of genes in the gene set
stats_vector	a vector of DEG statistics for every gene in gene_list (rank_vector in the DEG results) Absolute values of the t-statistics are required for the desired performance.

Value

tmp_sim_sgeas: a vector of ES values for all permutated gene lists

GSEA_permut_R_revised *Compute ES scores for Gene Set Enrichment Analysis (GSEA) with a new permutation method (using a revised algorithm)*

Description

To improve computational efficiency, we devised a new permutation approach by swapping stats_vector. Here, the gene symbols/names are permutated without changing the ranking vector (stats_vector). This function becomes unused in NetAct, as a much faster c++ implementation (GSEA_permute) is provided.

Usage

```
GSEA_permut_R_revised(sim_all, gene_set, stats_vector, N)
```

Arguments

sim_all	a vector of genes in the expression data
gene_set	a vector of genes in the gene set
stats_vector	a vector of DEG statistics for every gene in gene_list (rank_vector in the DEG results); Absolute values of the t-statistics are required for the desired performance.
N	total number of genes (size of sim_all)

Value

ES: enrichment score

`GSEA_proc_R` *Gene Set Enrichment Analysis (GSEA) with a new permutation method – implementation in R*

Description

To improve computational efficiency, we devised a new permutation approach by swapping stats_vector. Here, the gene symbols/names are permutated without changing the ranking vector (stats_vector).

Usage

```
GSEA_proc_R(GSDB, DElist, minSize = 5, nperm = 1000)
```

Arguments

`GSDB`	gene set database (a list of gene sets, each of which is comprised of a vector genes)
`DElist`	a vector of DEG statistics for every gene in gene_list (rank_vector in the DEG results)
`minSize`	the minimum number of overlapping genes required for each gene set (a gene set filtering parameter, default: 5)
`nperm`	the number of gene list permutations (default: 1000)

Value

data.frame(rslt_mat): a table of GSEA results: tf: TF (gene set name). es: ES score. lens: number of overlapping genes in each gene set. pvals: p-value by counting. z: z-score. qvals: q-value from pvals.

`GSEA_proc_RC` *Gene Set Enrichment Analysis (GSEA) with a new permutation method – implementation in R/c++*

Description

To improve computational efficiency, we devised a new permutation approach by swapping stats_vector. Here, the gene symbols/names are permutated without changing the ranking vector (stats_vector). A much faster c++ implementation (GSEA_permute) is used.

Usage

```
GSEA_proc_RC(GSDB, DElist, minSize = 5, nperm = 1000)
```

Arguments

GSDB	gene set database (a list of gene sets, each of which is comprised of a vector genes)
DElist	a vector of DEG statistics for every gene in gene_list (rank_vector in the DEG results)
minSize	the minimum number of overlapping genes required for each gene set (a gene set filtering parameter, default: 5)
nperm	the number of gene list permutations (default: 1000)

Value

data.frame(rslt_mat): a table of GSEA results: tf: TF (gene set name). es: ES score. lens: number of overlapping genes in each gene set. pvals: p-value by counting. z: z-score. qvals: q-value from pvals.

GSEA_score *Compute the enrichment score (ES) from Gene Set Enrichment Analysis*

Description

In this specific implementation, we modified two aspects: (1) the stats_vector takes the absolute value of the t-statistics; This ensures that the GSEA works for the case where part of genes are up-regulated and others are down-regulated. (2) permutation of the gene symbols of the ranking vector, instead of gene expression values.

Usage

```
GSEA_score(gene_list, gene_set, stats_vector)
```

Arguments

gene_list	a vector of genes in the expression data
gene_set	a vector of genes in the gene set
stats_vector	a vector of DEG statistics for every gene in gene_list (rank_vector in the DEG results) Absolute values of the t-statistics are required for the desired performance.

Value

ES: enrichment score

hDB *Human transcription factor target regulatory database*

Description

This data contains the literature-derived target genes of transcription factors for human genome. For further details, see the bioRxiv preprint https://doi.org/10.1101/2022.05.06.487898

Usage

```
hDB
```

Format

A list of 875 TFs, each element is an array of gene symbols of target genes.

Hill *Hill function for the gene weight*

Description

Hill function for the gene weight

Usage

```
Hill(x, ind)
```

Arguments

x	value (adj p-value)
ind	Hill coefficient

Value

Hill function of x

mDB *Mouse transcription factor target regulatory database*

Description

This data contains the literature-derived target genes of transcription factors for mouse genome. For further details, see the bioRxiv preprint https://doi.org/10.1101/2022.05.06.487898

Usage

```
mDB
```

Format

A list of 895 TFs, each element is an array of gene symbols of target genes.

MicroDegs *Helper Function For DEG Analysis of microArray Data (for all cases, including single and multiple comparisons)*

Description

Helper Function For DEG Analysis of microArray Data (for all cases, including single and multiple comparisons)

Usage

```
MicroDegs(eset, compList, qval = 0.05)
```

Arguments

eset	Processed gene expression data in the ExpressionSet format batch & experimental conditions are provided in pData.
compList	a vector of multiple comparisons in the format of contrasts in limma (e.g. c("A-B", "A-C", "B-C"))
qval	q-value cutoff for DEG analysis (default: 0.05)

Value

DEresult: a list of DEG results, including those for each single comparison and those for the overall comparison. Each DEG result is in the format of A list containing: table: table of DEG results. rank_vector: a vector of t-statistics for every gene. degs: a vector of gene symbols for DEGs.

plot_network *Plotting gene network*

Description

Plotting gene network

Usage

```
plot_network(tf_links = tf_links)
```

Arguments

tf_links	a data frame of networ interactions

Value

visNetwork object

Preprocess_counts *RNA-seq data pre processing*

Description

NetAct uses edgeR to load the count data and the group information for experimental conditions, It also coverts gene symbols and remove duplicates.

Usage

```
Preprocess_counts(counts, groups, mouse = FALSE)
```

Arguments

counts	raw count matrix
groups	group information for experimental conditions
mouse	use mouse genome or not (default: FALSE)

Value

x$counts: processed count matrix

rem_data *Remove Non-informative genes*

Description

Remove Non-informative genes

Usage

```
rem_data(x)
```

Arguments

x	gene expression matrix

Value

x: gene expression matrix without containing non-informative genes

Reselect_TFs — *Reselecting TFs using gene set enrichement analysis (GSEA) using an adjusted set of parameters (work together with TF_Selection)*

Description

Reselecting TFs using gene set enrichement analysis (GSEA) using an adjusted set of parameters (work together with TF_Selection)

Usage

```
Reselect_TFs(GSEArslt, qval = 0.05, combine_TFs = TRUE, ntop = NULL)
```

Arguments

GSEArslt	GSEA results from TF_Selection
qval	q-value cutoff (default: 0.05)
combine_TFs	whether combine selected TFs from multiple comparisons or not (default: TRUE)
ntop	the number of top genes (selection by the top genes) (default: NULL, no selection by the top genes)

Value

tfs: a vector of selected TFs

RNAseqDegs_DESeq — *Helper Function For DEG Analysis of RNA-seq Data using DESeq*

Description

Helper Function For DEG Analysis of RNA-seq Data using DESeq

Usage

```
RNAseqDegs_DESeq(counts, phenodata, complist, qval = 0.05)
```

Arguments

counts	Processed gene expression count data
phenodata	pData that provides batch & experimental conditions
complist	a vector of multiple comparisons in the format of contrasts in limma (e.g. c("A-B", "A-C", "B-C"))
qval	q-value cutoff for DEG analysis (default: 0.05)

Value

DEresult: a list of DEG results, including those for each single comparison and those for the overall comparison. Each DEG result is in the format of A list containing: table: table of DEG results. rank_vector: a vector of t-statistics for every gene. degs: a vector of gene symbols for DEGs. e: expression data (CPM).

RNAseqDegs_limma — *Helper Function For DEG Analysis of RNA-seq Data using limma + Voom*

Description

Helper Function For DEG Analysis of RNA-seq Data using limma + Voom

Usage

```
RNAseqDegs_limma(counts, phenodata, complist, lfc, qval = 0.05)
```

Arguments

counts	Processed gene expression count data
phenodata	pData that provides batch & experimental conditions
complist	a vector of multiple comparisons in the format of contrasts in limma (e.g. c("A-B", "A-C", "B-C"))
lfc	(optional) log fold change constraints for DEGs
qval	q-value cutoff for DEG analysis (default: 0.05)

Value

DEresult: a list of DEG results, including those for each single comparison and those for the overall comparison. Each DEG result is in the format of A list containing: table: table of DEG results. rank_vector: a vector of t-statistics for every gene. degs: a vector of gene symbols for DEGs. e: expression data (CPM). e_batch: batch corrected expression.

row_norm — *Row normalization (standardization)*

Description

Row normalization (standardization)

Usage

```
row_norm(data)
```

Arguments

data	gene expression matrix

Value

norm_data: standardized gene expression matrix

TF_Activity *Inference of TF activity*

Description

Inference of TF activity

Usage

```
TF_Activity(
  tfs,
  GSDB,
  eset,
  DErslt,
  with_weight = TRUE,
  if_module = FALSE,
  ind = 1/5,
  useCorSign = TRUE
)
```

Arguments

tfs	a vector of selected tfs
GSDB	gene set database (a list of gene sets, each of which is comprised of a vector genes)
eset	expression set of gene expression data or gene expression matrix
DErslt	DEG results
with_weight	whether weighting factors (based on DEG p-values) are used to compute TF activity (default: TRUE)
if_module	whether the grouping scheme (activation or inhibition) depends on module detection algorithm (default: FALSE, no need to change)
ind	Hill coefficient parameter used in the weighting factors (default: 1/5, recommend to use 0 < ind < 1/4)
useCorSign	allow the option to use the TF gene expression correlation to flip signs (default: TRUE)

Value

a list of results: all_list: grouping scheme of all TF gene sets. all_activity: matrix of TF activity.

TF_Filter *Generate network*

Description

Network calculated using activity and interaction database. Uses mutual information to find possible interactions and keeps the interactions if they are available in the database. Sign of interaction is assigned based on the correlation between the activities.

Usage

```
TF_Filter(
  actMat,
  GSDB,
  miTh = 0.4,
  maxTf = 75,
  maxInteractions = 300,
  nbins = 16,
  miMethod = "2d",
  corMethod = "spearman",
  useCor = FALSE,
  removeSignalling = FALSE,
  DPI = FALSE,
  nameFile = NULL,
  ...
)
```

Arguments

actMat	numeric matrix. Matrix containing the activities
GSDB	List of list. Gene set database of interactions
miTh	numeric. Mutual information threshold
maxTf	integer (optional). Default 75. Maximum number of transcription factors in the network. If removeSignalling is TRUE the actual number will be less.
maxInteractions	integer (optional). Default 300. Maximum number of interactions in the network.
nbins	integer (optional). Number of bins Default 16.
miMethod	MI calculation method: "2d": 2D discretization with entropy (default) "1d": 1D discretization with infotheo
corMethod	character (optional). Method to compute correlation.
useCor	Logical (optional). Whether to use correlation instead of mutual information to find possible interactions.
removeSignalling	logical (optional). Whether to remove the Tfs which are not the target of any other Tfs. Default TRUE. It is not recursive and the generated network might still contain some signalling tfs.
DPI	logical (optional). Default FALSE. Whether to apply the data processing inequality to remove weak edges from triangles.

nameFile	character (optional). Ouput file name. Default NULL (no file output).
...	two additional parameters passed from applyDPI (default: miDiff = 0, minMiTh = 0.5)

Value

data.frame. Contains the interactions in a dataframe listing. source tf, target tf and interaction type (1-activation, 2-inhibition).

`TF_Filter_addgene` *Generate network (an extension of TF_Filter)*

Description

Network calculated using activity and interaction database. Uses mutual information to find possible interactions and keeps the interactions if they are available in the database. Sign of interaction is assigned based on the correlation between the activities. An extension of TF_Filter. Add a list of genes of interest.

Usage

```
TF_Filter_addgene(
  actMat,
  GSDB,
  genes,
  DEgenes,
  eset,
  miTh = 0.4,
  maxTf = 75,
  maxInteractions = 300,
  nbins = 16,
  miMethod = "e",
  corMethod = "spearman",
  useCor = FALSE,
  removeSignalling = FALSE,
  DPI = FALSE,
  ...
)
```

Arguments

actMat	numeric matrix. Matrix containing the activities
GSDB	List of list. Gene set database of interactions
genes	vector. a vector of gene symbols of genes of interest
DEgenes	vector. a vector of gene symbols of DE genes
eset	expression set of gene expression data or gene expression matrix
miTh	numeric. Mutual information threshold
maxTf	integer (optional). Default 75. Maximum number of transcription factors in the network. If `removeSignalling` is `TRUE` the actual number will be less.

maxInteractions
: integer (optional). Default 300. Maximum number of interactions in the network.

nbins
: integer (optional). Number of bins Default 16.

miMethod
: MI calculation method: e: entropy (default) or i: infotheo

corMethod
: character (optional). Method to compute correlation.

useCor
: Logical (optional). Whether to use correlation instead of mutual information to find possible interactions. Default FALSE

removeSignalling
: logical (optional). Whether to remove the Tfs which are not the target of any other Tfs. Default FALSE. It is not recursive and the generated network might still contain some signalling tfs.

DPI
: logical (optional). Default FALSE. Whether to apply the data processing inequality to remove weak edges from triangles.

...
: two additional parameters passed from applyDPI (default: miDiff = 0, minMiTh = 0.5)

Value

List of data.frame. Contains the interactions in a data frame listing. source tf, target tf and interaction type (1-activation, 2-inhibition). tf_links: network interactions. new_links: new interactions associated with the genes of interest.

TF_GSEA *A unified Gene Set Enrichment Analysis (GSEA) function for three methods*

Description

A unified Gene Set Enrichment Analysis (GSEA) function for three methods

Usage

```
TF_GSEA(GSDB, DErslt, minSize = 5, nperm = 1000, method = "binary")
```

Arguments

GSDB
: gene set database (a list of gene sets, each of which is comprised of a vector genes)

DErslt
: DEG results

minSize
: the minimum number of overlapping genes required for each gene set (a gene set filtering parameter, default: 5)

nperm
: the number of gene list permutations (default: 1000)

method
: fast: fgsea; r: R implementation of GSEA with a new permutation method; binary: R/C++ implementation for fast speed

Value

gseaRes: a table of GSEA results: tf: TF (gene set name). es: ES score. lens: number of overlapping genes in each gene set. pvals: p-value by counting. z: z-score. qvals: q-value from pvals.

TF_Selection *Identifying enriched TFs using Gene Set Enrichment Analysis (GSEA) – a wrapper function with many options*

Description

Identifying enriched TFs using Gene Set Enrichment Analysis (GSEA) – a wrapper function with many options

Usage

```
TF_Selection(
  GSDB,
  DErslt,
  minSize = 5,
  nperm = 5000,
  method = "binary",
  qval = 0.05,
  compList = NULL,
  ntop = NULL,
  nameFile = NULL
)
```

Arguments

GSDB	gene set database (a list of gene sets, each of which is comprised of a vector genes)
DErslt	DEG results
minSize	the minimum number of overlapping genes required for each gene set (a gene set filtering parameter, default: 5)
nperm	the number of gene list permutations (default: 1000)
method	fast: fgsea; R: r implementation of GSEA with a new permutation method; binary: R/C++ implementation for fast speed
qval	q-value cutoff (default: 0.05)
compList	a vector of comparisons, it needs to be consistent with DErslt from MicroDegs, RNAseqDegs_limma, and RNAseqDegs_DESeq. GSEA is applied to each comparison
ntop	the number of top genes (selection by the top genes) (default: NULL, no selection by the top genes)
nameFile	file name to save the GSEA results (default: NULL, no output to a file). The saved results can be reused later to adjust the TF selection parameters

Value

a list of results: GSEArslt: a dataframe of GSEA results (see TF_GSEA). tfs: a vector of selected TFs.

`toCPM` *convert to log10 (CPM) measurement in the RNA-Seq matrix*

Description

convert to log10 (CPM) measurement in the RNA-Seq matrix

Usage

```
toCPM(ctMat)
```

Arguments

`ctMat`	Matrix of gene expression counts

Value

mat: Matrix of CPM gene expression

Index