Package ‘PPInfer’

March 30, 2024

Type Package

Title Inferring functionally related proteins using protein interaction networks

Description Interactions between proteins occur in many, if not most, biological processes. Most proteins perform their functions in networks associated with other proteins and other biomolecules. This fact has motivated the development of a variety of experimental methods for the identification of protein interactions. This variety has in turn ushered in the development of numerous different computational approaches for modeling and predicting protein interactions. Sometimes an experiment is aimed at identifying proteins closely related to some interesting proteins. A network based statistical learning method is used to infer the putative functions of proteins from the known functions of its neighboring proteins on a PPI network. This package identifies such proteins often involved in the same or similar biological functions.

Version 1.28.0

Date 2022-11-17

Author Dongmin Jung, Xijin Ge

Maintainer Dongmin Jung <dmdmjung@gmail.com>

Depends biomaRt, fgsea, kernlab, ggplot2, igraph, STRINGdb, yeastExpData

Imports httr, grDevices, graphics, stats, utils

Suggests

License Artistic-2.0

biocViews Software, StatisticalMethod, Network, GraphAndNetwork, GeneSetEnrichment, NetworkEnrichment, Pathways

NeedsCompilation no

git_url https://git.bioconductor.org/packages/PPInfer

git_branch RELEASE_3_18

git_last_commit 8fc4265

git_last_commit_date 2023-10-24

Repository Bioconductor 3.18

Date/Publication 2024-03-29


\section*{R topics documented:}

- PPIInfer-package ........................................ 2
- enrich.net ............................................... 3
- GSEA.barplot ........................................... 4
- net.infer .................................................. 6
- net.infer.ST ............................................. 7
- net.kernel ................................................ 9
- ORA ....................................................... 10
- ORA.barplot ............................................. 11
- ppi.infer.human ......................................... 12
- ppi.infer.mouse ......................................... 13
- self.train.kernel ....................................... 15

\section*{Index}

17

\begin{center}
\begin{tabular}{ll}
\textbf{PPIInfer-package} & \textit{Inferring functionally related proteins using protein interaction networks} \\
\end{tabular}
\end{center}

\section*{Description}

Interactions between proteins occur in many, if not most, biological processes. Most proteins perform their functions in networks associated with other proteins and other biomolecules. This fact has motivated the development of a variety of experimental methods for the identification of protein interactions. This variety has in turn ushered in the development of numerous different computational approaches for modeling and predicting protein interactions. Sometimes an experiment is aimed at identifying proteins closely related to some interesting proteins. A network based statistical learning method is used to infer the putative functions of proteins from the known functions of its neighboring proteins on a PPI network. This package identifies such proteins often involved in the same or similar biological functions.

\section*{Details}

The DESCRIPTION file: This package was not yet installed at build time.

Index: This package was not yet installed at build time.

\section*{Author(s)}

Dongmin Jung, Xijin Ge

Maintainer: Dongmin Jung <dmdmjung@gmail.com>
enrich.net

enrich.net Visualize network for the functional enrichment analysis

Description
The connection between nodes depends on the proportion of overlapping genes between two categories.

Usage
enrich.net(x, gene.set, node.id, node.name = node.id, pvalue,
  n = 50, numChar = NULL, pvalue.cutoff = 0.05,
  edge.cutoff = 0.05, degree.cutoff = 0,
  edge.width = function(x) {10*x^2},
  node.size = function(x) {2.5*log10(x)},
  group = FALSE, group.color = c('red', 'green'),
  group.shape = c('circle', 'square'),
  legend.parameter = list('topright'),
  show.legend = TRUE, ...
)

Arguments
x a result with category and p-value of gene sets
gene.set gene sets which is already used for functional enrichment
node.id name of gene sets
node.name label of nodes in the network (default: node.id)
pvalue pvalues for categories
n number of top categories (default: 50)
numChar the maximal number of characters of the label of gene sets
pvalue.cutoff nodes with p-values which are greater than pvalue.cutoff are removed (default: 0.05)
edge.cutoff edges with the proportion which is less than edge.cutoff are removed (default: 0.05)
degree.cutoff nodes with the degrees which are less than degree.cutoff are removed (default: 0)
edge.width width of edges
node.size size of nodes
group variable for group
group.color color for group (default: red and green for 2 groups)
group.shape shape for group (default: circle and square for 2 groups)
legend.parameter list of parametres for the legend
show.legend show the legend (default: TRUE)
... additional parameters for the igraph
Value

plot for the network. The size of nodes is proportional to the size of gene sets. The more significant categories are, the less transparent their nodes are.

Author(s)

Dongmin Jung, Xijin Ge

References


See Also

igraph

Examples

data(examplePathways)
data(exampleRanks)
set.seed(1)
result.GSEA <- fgsea(examplePathways, exampleRanks, nperm = 1000)
enrich.net(result.GSEA, examplePathways, node.id = 'pathway',
  pvalue = 'pval', edge.cutoff = 0.6, degree.cutoff = 1,
  n = 50, vertex.label.cex = 0.75, show.legend = FALSE,
  edge.width = function(x) {5*sqrt(x)},
  layout = igraph::layout.kamada.kawai)
Arguments

object  a table with category, enrichment score and p-value of gene sets
category name of gene sets
score    enrichment score
pvalue   p-value of gene sets
top      the number of top categories (default: 10)
sort     a variable used for sorting data
decreasing logical indicating whether ascending or descending order (default: FALSE)
numChar  the maximal number of characters of the name of gene sets
title    title for the plot
transparency transparency (default: 0.5)
plot     return plot when plot is true, otherwise return table (default: TRUE)

Value

GSEA barplot

Author(s)

Dongmin Jung, Xijin Ge

References


See Also

ggplot2

Examples

data(examplePathways)
data(exampleRanks)
set.seed(1)
result.GSEA <- fgsea(examplePathways, exampleRanks, nperm = 1000)
GSEA.barplot(result.GSEA, category = 'pathway', score = 'NES',
pvalue = 'pval', sort = 'NES', decreasing = TRUE)
net.infer

Inferring functionally related proteins using networks

Description

Proteins can be classified by using networks to identify functionally closely related proteins.

Usage

```r
net.infer(target, kernel, top = NULL, cross = 0,
C = 1, nu = 0.2, epsilon = 0.1, cache1 = 40,
tol1 = 0.001, shrinking1 = TRUE, cache2 = 40,
tol2 = 0.001, shrinking2 = TRUE)
```

Arguments

- `target`: set of interesting proteins or target class
- `kernel`: the regularized Laplacian matrix for a graph
- `top`: number of top proteins most closely related to target class (default: all proteins except for target and pseudo-absence class)
- `cross`: if an integer value k>0 is specified, a k-fold cross validation on the training data is performed to assess the quality of the model
- `C`: cost of constraints violation for SVM (default: 1)
- `nu`: The nu parameter for OCSVM (default: 0.2)
- `epsilon`: epsilon in the insensitive-loss function for OCSVM (default: 0.1)
- `cache1`: cache memory in MB for OCSVM (default: 40)
- `tol1`: tolerance of termination criterion for OCSVM (default: 0.001)
- `shrinking1`: option whether to use the shrinking-heuristics for OCSVM (default: TRUE)
- `cache2`: cache memory in MB for SVM (default: 40)
- `tol2`: tolerance of termination criterion for SVM (default: 0.001)
- `shrinking2`: option whether to use the shrinking-heuristics for SVM (default: TRUE)

Value

- `list`: list of a target class used in the model
- `error`: training error
- `CVerror`: cross validation error, (when cross > 0)
- `top`: top proteins
- `score`: decision values for top proteins

Author(s)

Dongmin Jung, Xijin Ge
References


See Also

ksvm

Examples

# example 1
## Not run:
string.db.9606 <- STRINGdb$new(version = '11', species = 9606, score_threshold = 999)
string.db.9606.graph <- string.db.9606$get_graph()
K.9606 <- net.kernel(string.db.9606.graph)
rownames(K.9606) <- substring(rownames(K.9606), 6)
colnames(K.9606) <- substring(colnames(K.9606), 6)
target <- colnames(K.9606)[1:100]
infer <- net.infer(target, K.9606, 10)

## End(Not run)

# example 2
data(litG)
litG <- igraph.from.graphNEL(litG)
sg <- decompose(litG, min.vertices = 50)
sg <- sg[[1]]
K <- net.kernel(sg)
litG.infer <- net.infer(names(V(sg))[1:10], K, top=20)

net.infer.ST

Inferring functionally related proteins with self training

Description

This function is the self-training version of net.infer. The function net.infer is the special case of net.infer.ST where a single iteration is conducted.

Usage

net.infer.ST(target, kernel, top = NULL, C = 1, nu = 0.2, epsilon = 0.1, cache1 = 40, tol1 = 0.001, shrinking1 = TRUE, cache2 = 40, tol2 = 0.001, shrinking2 = TRUE, thrConf = 0.9, maxIts = 10, percFull = 1, verbose = FALSE)
Arguments

**target**
set of interesting proteins or target class

**kernel**
the regularized Laplacian matrix for a graph

**top**
number of top proteins most closely related to target class (default: all proteins except for target and pseudo-absence class)

**C**
cost of constraints violation for SVM (default: 1)

**nu**
The nu parameter for OCSVM (default: 0.2)

**epsilon**
epsilon in the insensitive-loss function for OCSVM (default: 0.1)

**cache1**
cache memory in MB for OCSVM (default: 40)

**tol1**
tolerance of termination criterion for OCSVM (default: 0.001)

**shrinking1**
option whether to use the shrinking-heuristics for OCSVM (default: TRUE)

**cache2**
cache memory in MB for SVM (default: 40)

**tol2**
tolerance of termination criterion for SVM (default: 0.001)

**shrinking2**
option whether to use the shrinking-heuristics for SVM (default: TRUE)

**thrConf**
A number between 0 and 1, indicating the required classification confidence for an unlabelled case to be added to the labelled data set with the label predicted by the classification algorithm (default: 0.9)

**maxIts**
The maximum number of iterations of the self-training process (default: 10)

**percFull**
A number between 0 and 1. If the percentage of labelled cases reaches this value the self-training process is stopped (default: 1)

**verbose**
A boolean indicating the verbosity level of the function. (default: FALSE)

Value

**list**
list of a target class used in the model

**error**
training error

**top**
top proteins

**score**
decision values for top proteins

Author(s)

Dongmin Jung, Xijin Ge

See Also

self.train

Examples

data(litG)
litG <- igraph.from.graphNEL(litG)
sg <- decompose(litG, min.vertices = 50)
sg <- sg[[1]]
K <- net.kernel(sg)
litG.infer.ST <- net.infer.ST(names(V(sg))[1:10], K, top=20)
net.kernel

Kernel matrix for a graph

Description
This function gives the regularized Laplacian matrix for a graph.

Usage
net.kernel(g, decay = 0.5)

Arguments
g       graph
decay    decaying constant (default: 0.5)

Value
the regularized Laplacian matrix

Author(s)
Dongmin Jung, Xijin Ge

See Also
laplacian_matrix

Examples

# example 1
## Not run:
string.db.9606 <- STRINGdb$new(version = '11', species = 9606,  
                           score_threshold = 999)
string.db.9606.graph <- string.db.9606$get_graph()
K.9606 <- net.kernel(string.db.9606.graph)
## End(Not run)

# example 2
data(litG)
litG <- igraph.from.graphNEL(litG)
sg <- decompose(litG, min.vertices=50)
sg <- sg[[1]]
K <- net.kernel(sg)
Over-representation Analysis

**Description**

the result from the over-representation analysis

**Usage**

```r
ORA(pathways, gene.id, minSize = 1, maxSize = Inf, 
    p.adjust.methods = NULL)
```

**Arguments**

- `pathways` list of gene sets
- `gene.id` set of genes
- `minSize` Minimal size of a gene set
- `maxSize` Maximal size of a gene set
- `p.adjust.methods` a correction method

**Value**

ORA result

**Author(s)**

Dongmin Jung, Xijin Ge

**See Also**

fisher.test

**Examples**

```r
data(examplePathways)
data(exampleRanks)
geneNames <- names(exampleRanks)
set.seed(1)
gene.id <- sample(geneNames, 100)
ORA(examplePathways, gene.id)
```
**Description**

For the functional enrichment analysis, we can visualize the result from the over-representation analysis.

**Usage**

```r
ORA.barplot(object, category, size, count, pvalue, top = 10,
             sort = NULL, decreasing = FALSE, p.adjust.methods = NULL,
             numChar = NULL, title = NULL, transparency = 0.5,
             plot = TRUE)
```

**Arguments**

- **object**: a table with category, size, count and p-value of gene sets
- **category**: name of gene sets
- **size**: size of gene sets
- **count**: count of gene sets
- **pvalue**: p-value of gene sets
- **top**: the number of top categories (default: 10)
- **sort**: a variable used for sorting data
- **decreasing**: logical indicating whether ascending or descending order (default: FALSE)
- **p.adjust.methods**: a correction method
- **numChar**: the maximal number of characters of the name of gene sets
- **title**: title for the plot
- **transparency**: transparency (default: 0.5)
- **plot**: return plot when plot is true, otherwise return table (default: TRUE)

**Value**

ORA barplot

**Author(s)**

Dongmin Jung, Xijin Ge

**References**

ppi.infer.human

Inferring functionally related proteins using protein networks for human

Description

This function is designed for human protein-protein interaction from STRING database. Default format is 'hgnc'. The number of proteins is 10 in default. Note that the number of proteins used as a target may be different from the number of proteins in the input since mapping between formats is not always one-to-one in getBM.

Usage

ppi.infer.human(target, kernel, top = 10, classifier = net.infer,
                 input = "hgnc_symbol", output = "hgnc_symbol", ...)

Arguments

target    set of interesting proteins or target class
kernel    the regularized Laplacian matrix for a graph
top       number of top proteins most closely related to target class (default: 10)
classifier net.infer or net.infer.ST (default: net.infer)
input     input format
output    output format
...       additional parameters for the chosen classifier

Value

list      list of a target class used in the model
error     training error
CError    cross validation error, (when cross > 0 in net.infer)
top       top proteins
score     decision values for top proteins

See Also

p.adjust, ggplot2

Examples

data(examplePathways)
data(exampleRanks)
geneNames <- names(exampleRanks)
set.seed(1)
gene.id <- sample(geneNames, 100)
result.ORA <- ORA(examplePathways, gene.id)
ORA.barplot(result.ORA, category = "Category", size = "Size",
            count = "Count", pvalue = "pvalue", sort = "pvalue")
Author(s)

Dongmin Jung, Xijin Ge

See Also

net.infer, net.infer.ST, getBM

Examples

# example 1
string.db.9606 <- STRINGdb$new(version = '11', species = 9606, 
    score_threshold = 999)
string.db.9606.graph <- string.db.9606$get_graph()
K.9606 <- net.kernel(string.db.9606.graph)
rownames(K.9606) <- substring(rownames(K.9606), 6)
colnames(K.9606) <- substring(colnames(K.9606), 6)
target <- colnames(K.9606)[1:100]
infer.human <- ppi.infer.human(target, K.9606, input = "ensembl_peptide_id")

## Not run:
# example 2
library(graph)
data(apopGraph)
target <- nodes(apopGraph)
apoptosis.infer <- ppi.infer.human(target, K.9606, 100)

# example 3
library(KEGGgraph)
library(KEGG.db)
pName <- "p53 signaling pathway"
pId <- mget(pName, KEGGPATHNAME2ID)[[1]]
getKGMLurl(pId, organism = "hsa")
p53 <- system.file("extdata/hsa04115.xml", package="KEGGgraph")
p53graph <- parseKGML2Graph(p53, expandGenes=TRUE)
entrez <- translateKEGGID2GeneID(nodes(p53graph))
http::set_config(http::config(ssl_verifypeer = FALSE))
human.ensembl <- useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
target <- getBM(attributes=c('entrezgene', 'hgnc_symbol'),
    filter = 'entrezgene', values = entrez,
    mart = human.ensembl)[,2]
p53.infer <- ppi.infer.human(target, K.9606, 100)

## End(Not run)
Description

This function is designed for mouse protein-protein interaction from STRING database. Default format is 'mgi'. The number of proteins is 10 in default. Note that the number of proteins used as a target may be different from the number of proteins in the input since mapping between formats is not always one-to-one in getBM.

Usage

```r
ppi.infer.mouse(target, kernel, top = 10, classifier = net.infer,
                 input = "mgi_symbol", output = "mgi_symbol", ...)
```

Arguments

- `target`: set of interesting proteins or target class
- `kernel`: the regularized Laplacian matrix for a graph
- `top`: number of top proteins most closely related to target class (default: 10)
- `classifier`: net.infer or net.infer.ST (default: net.infer)
- `input`: input format
- `output`: output format
- `...`: additional parameters for the chosen classifier

Value

- `list`: list of a target class used in the model
- `error`: training error
- `CVerror`: cross validation error, (when cross > 0 in net.infer)
- `top`: top proteins
- `score`: decision values for top proteins

Author(s)

Dongmin Jung, Xijin Ge

See Also

- net.infer, net.infer.ST, getBM

Examples

```r
string.db.10090 <- STRINGdb$new(version = '11', species = 10090,
                                 score_threshold = 999)
string.db.10090.graph <- string.db.10090$get_graph()
K.10090 <- net.kernel(string.db.10090.graph)
rownames(K.10090) <- substring(rownames(K.10090), 7)
colnames(K.10090) <- substring(colnames(K.10090), 7)
target <- colnames(K.10090)[1:100]
infer.mouse <- ppi.infer.mouse(target, K.10090, input="ensembl_peptide_id")
```
**Description**

This function can be used for classification of semi-supervised data by using the kernel support vector machine.

**Usage**

```r
self.train.kernel(K, y, type = 'response', C = 1, cache = 40,
                   tol = 0.001, shrinking = TRUE, thrConf = 0.9,
                   maxIts = 10, percFull = 1, verbose = FALSE)
```

**Arguments**

- `K`: kernel matrix  
- `y`: label vector  
- `type`: one of response, probabilities, votes, decision indicating the type of output (default: response)  
- `C`: cost of constraints violation for SVM (default: 1)  
- `cache`: cache memory in MB for SVM (default: 40)  
- `tol`: tolerance of termination criterion for SVM (default: 0.001)  
- `shrinking`: option whether to use the shrinking-heuristics for OCSVM (default: TRUE)  
- `thrConf`: A number between 0 and 1, indicating the required classification confidence for an unlabelled case to be added to the labelled data set with the label predicted predicted by the classification algorithm (default: 0.9)  
- `maxIts`: The maximum number of iterations of the self-training process (default: 10)  
- `percFull`: A number between 0 and 1. If the percentage of labelled cases reaches this value the self-training process is stopped (default: 1)  
- `verbose`: A boolean indicating the verbosity level of the function (default: FALSE)

**Value**

prediction from the SVM

**Author(s)**

Dongmin Jung, Xijin Ge

**References**

Examples

data(litG)
litG <- igraph.from.graphNEL(litG)
sg <- decompose(litG, min.vertices = 50)
sg <- sg[[1]]
K <- net.kernel(sg)
y <- rep(NA, length(V(sg)))
y[1:10] <- 1
y[11:20] <- 0
y <- factor(y)
self.train.kernel(K, y)
Index

enrich.net, 3
GSEA.barplot, 4
net.infer, 6
net.infer.ST, 7
net.kernel, 9
ORA, 10
ORA.barplot, 11
ppi.infer.human, 12
ppi.infer.mouse, 13
PPInfer (PPInfer-package), 2
PPInfer-package, 2
self.train.kernel, 15