Package 'MetID'

November 15, 2024

Type Package

Title Network-based prioritization of putative metabolite IDs

Version 1.24.0

Author Zhenzhi Li <zzrickli@gmail.com>

Maintainer Zhenzhi Li <zzrickli@gmail.com>

URL https://github.com/ressomlab/MetID

Description This package uses an innovative network-based approach that will enhance our ability to determine the identities of significant ions detected by LC-MS.

License Artistic-2.0

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

Depends R (>= 3.5)

Imports utils (>= 3.3.1), stats (>= 3.4.2), devtools (>= 1.13.0), stringr (>= 1.3.0), Matrix (>= 1.2-12), igraph (>= 1.2.1), ChemmineR (>= 2.30.2)

Suggests knitr (>= 1.19), rmarkdown (>= 1.8)

VignetteBuilder knitr

biocViews AssayDomain, BiologicalQuestion, Infrastructure, ResearchField, StatisticalMethod, Technology, WorkflowStep, Network, KEGG

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

git_branch RELEASE_3_20

git_last_commit 5809773

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-11-14

Contents

`demo1` *Example of input dataset, in which colnames does not meet requirement.*

Description

A dataset which can be used as input dataset and its row names do not match the default row names.

Usage

```
demo1
```

Format

A data frame with 20 rows and 6 variables:

Query.Mass Mass of compounds.

Name Names of putative IDs.

Formula Formulas of putative IDs.

Exact.Mass Exact mass of putative IDs.

PubChem.CID PubChem IDs of putative IDs.

KEGG.ID KEGG IDs of putative IDs. ...

`demo2` *Example of input dataset, in which colnames does not meet requirement.*

Description

A dataset which can be used as input dataset and its row names do not match the default row names.

Usage

```
demo2
```

Format

A data frame with 3592 rows and 6 variables:

Query.Mass Mass of compounds.

Name Names of putative IDs.

Formula Formulas of putative IDs.

Exact.Mass Exact mass of putative IDs.

PubChem.CID PubChem IDs of putative IDs.

KEGG.ID KEGG IDs of putative IDs. ...

---

get_cleaned *Preprocess input file.*

Description

Preprocess input file.

Usage

```
get_cleaned(filename, type = c("data.frame", "csv", "txt"), na, sep)
```

Arguments

filename	the name of the file which the data are to be read from. Its type should be chosen in 'type' parameter. Also, it should have columns named exactly as 'metid' (IDs for peaks), 'query_m.z' (query mass of peaks), 'exact_m.z' (exact mass of putitative IDs), 'kegg_id' (IDs of putitative IDs from KEGG Database), 'pubchem_cid' (CIDs of putitative IDs from PubChem Database). Otherwise, this function would not work.
type	string indicating the type of the file. It can be a 'data.frame' which is already loaded into R, or some other types like a csv file.
na	a character vector of strings which are to be interpreted as NA values.
sep	a character value which seperates multiple IDs in kegg_id or pubchem_cid field, if there are multiple IDs.

Value

get_cleaned returns a list containing the following components:

df	a data frame which is the original input data.
clean_data	a data frame with unuseful observations and features removed.
mass	a data frame with unique query peak, along with query mass.
ID	a data frame with unique putitative IDs, along with PubChem ID, KEGG ID, exact mass.
index_na	a vector of row indexes which contains NA values.

get_kegg_network *Build network between identifications based on kegg network database.*

Description

Build network between identifications based on kegg network database.

Usage

```
get_kegg_network(kegg_id)
```

Arguments

kegg_id	a vector of strings indicating KEGG ID of putative ID.

Value

a binary matrix of network of KEGG IDs.

get_scores_for_LC_MS *Get scores for metabolite putative IDs by LC-MS .*

Description

Get scores for metabolite putative IDs by LC-MS .

Usage

```
get_scores_for_LC_MS(filename, type = c("data.frame", "csv", "txt"),
  na = "NA", sep = ";", mode = c("POS", "NEG"), Size = 2000,
  delta = 1, gamma_mass = 10, iterations = 500)
```

Arguments

filename	the name of the file which the data are to be read from. Its type should be chosen in 'type' parameter. Also, it should have columns named exactly 'metid' (IDs for peaks), 'query_m.z' (query mass of peaks), 'exact_m.z' (exact mass of putative IDs), 'kegg_id' (IDs of putative IDs from KEGG Database), 'pubchem_cid' (CIDs of putative IDs from PubChem Database). Otherwise, this function would not work.
type	string indicating the type of the file. It can be a 'data.frame' which is already loaded into R, or some other specified types like a csv file.
na	a character vector of strings which are to be interpreted as NA values.
sep	a character value which seperates multiple IDs in kegg_id or pubchem_cid field, if there are multiple IDs.
mode	string indicating the mode of metabolites. It can be positive mode (POS) or negative mode (NEG).

Size	an integer which indicates sample size in Gibbs sampling.
delta	a hyper-parameter representing the mean value of mass ratio.
gamma_mass	a hyper-parameter representing the accuracy of mass measurement.
iterations	ask user to input number of interations,default 500

Value

A dataframe which contains input data together with a column of scores in the end. In the score column, if the row contains NA values or does not has a PubChem cid, the score would be '-', which stands for missing value. Otherwise, each score would be from 0 to 1.

Examples

```
## check if colnames of dataset meet requirement
names(demo1)
## change colnames
colnames(demo1) <- c('query_m.z','name','formula','exact_m.z','pubchem_cid','kegg_id')
## get scores
out <- get_scores_for_LC_MS(demo1, type = 'data.frame', na='-', mode='POS')
```

---

get_tani_network *Build network between identifications based on tanimoto score.*

Description

Build network between identifications based on tanimoto score.

Usage

```
get_tani_network(pubchem_cid)
```

Arguments

pubchem_cid	a vector of strings indicating PubChem CID of putative ID.

Value

a binary matrix of network of tanimoto scores.

InchiKey — *Inchikey database.*

Description

A dataset containing PubChem CIDs, InchiKey in the PubChem database.

Usage

```
InchiKey
```

Format

A data frame with 101494 rows and 2 variables:

CID PubChem CIDs

InchiKey Inchikeys ...

kegg_network — *Pairs of kegg network.*

Description

A dataset containing kegg IDs in the KEGG database with all networks.

Usage

```
kegg_network
```

Format

A data frame with 57070 rows and 2 variables:

r1 KEGG IDs

r2 KEGG IDs, which have a connection with KEGG ID in the first column ...

MetID — *MetID: A package for Network-based prioritization of putative metabolite IDs.*

Description

The foo package provides one important functions: get_scores_for_LC_MS

Foo functions

get_scores_for_LC_MS: Get scores for metabolite putative IDs by LC-MS.

Index