Package ‘DTSEAdata’

November 4, 2022

Type Package

Title Data collections for DTSEA package

Version 0.0.3

Maintainer Junwei Han <hanjunwei1981@163.com>

Description This package provides supplementary data for DTSEA.

License GPL (>= 2)

Depends R (>= 4.0.0)

Imports dplyr

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.1

R topics documented:

DTSEAdata-package *Data collections for Drug set enrichment analysis (DTSEA)*

Description

This package provides supplementary data for **DTSEA**. The data package contains full disease-related gene list of GSE183071, full network list, and other useful supplementary data for exploring DTSEA.

We also provide a list of all the supplementary data files.

Data name	Description
`deletion`	Drug prediction results on trimmed graphs after enrichment
`drug_indications`	The drug indications data from ChEMBL
`drug_predicted`	The drug target data predicted by ChEMBL
`drug_targets`	The combined drug target data
`expr_raw`	The raw expression profile of GSE164805
`graph`	The human gene functional interaction network we used
`ncbi_list`	The COVID-19-related genes provided by NCBI

Details

DTSEAdata

`deletion` *Drug prediction results on trimmed graphs after enrichment*

Description

We we simulated 5 node failures and 5 link failures with 50 repeats in each condition to measure the robustness of the graph.

Usage

```
deletion
```

Format

A list with 500 items, including 10%, 20%, 30%, 40%, 50% failures, and node and link failures.

Examples

```
library(DTSEAdata)
data("deletion")

# Get length
length(deletion)
```

disease_related *The disease-related nodes used in the DTSEA paper*

Description

The list was integrated the significantly differentially expressed genes (DEGs) of GEO dataset GSE183071 and the work from Feng, Song, Guo, and et al.

Usage

```
disease_related
```

Format

An object of class character of length 113.

References

Gómez-Carballa A, Rivero-Calle I, Pardo-Seco J, Gómez-Rial J, Rivero-Velasco C, Rodríguez-Núñez N, Barbeito-Castiñeiras G, Pérez-Freixo H, Cebey-López M, Barral-Arca R, Rodriguez-Tenreiro C, Dacosta-Urbieta A, Bello X, Pischedda S, Currás-Tuala MJ, Viz-Lasheras S, Martinón-Torres F, Salas A; GEN-COVID study group. A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity. Environ Res. 2022 Jul;210:112890. doi: 10.1016/j.envres.2022.112890. Epub 2022 Feb 22. PMID: 35202626; PMCID: PMC8861187.

Feng S, Song F, Guo W, Tan J, Zhang X, Qiao F, Guo J, Zhang L, Jia X. Potential Genes Associated with COVID-19 and Comorbidity. Int J Med Sci. 2022 Jan 24;19(2):402-415. doi: 10.7150/ijms.67815. PMID: 35165525; PMCID: PMC8795808.

Examples

```
library(DTSEAdata)
data("disease_related")

# Get the count of the vector
length(disease_related)
```

drug_indications *The drug indications data used in the DTSEA paper*

Description

We downloaded the drug indications data from ChEMBL, one of the largest drug information databases in the world. In our provided data, we deleted the drugs without a DrugBank ID.

Usage

```
drug_indications
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 32758 rows and 8 columns.

Details

A data frame with 32758 rows and 8 columns:

- `drugbank_id`: The DrugBank ID of the drugs.
- `drug_name`: the drug name.
- `max_phase`: refers to the phase a drug achieved in a clinical trial. Phases are typically ordered from **0** to **4**:
  0. Preclinical research;

  1. Clinical phase 1;
  2. Clinical phase 2;
  3. Clinical phase 3;
  4. Approved by FDA or EMCA.
- `mesh_id`: the Mesh ID of the specific indications.
- `mesh_heading`: the Mesh Heading of the specific indications.
- `efo_id`: the EFO ID of the specific indications.
- `efo_term`: the EFO Term of the specific indications.
- `indication`: the text-based main indications of the drugs.

References

Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Res. 2017 Jan 4;45(D1):D945-D954. doi: 10.1093/nar/gkw1074. Epub 2016 Nov 28. PMID: 27899562; PMCID: PMC5210557.

Examples

```
library(DTSEAdata)
data("drug_indications")

# Get Aspirin
library(dplyr)
filter(drug_indications, drug_name == 'Acetylsalicylic acid') %>%
  select(-indication)
```

drug_predicted *The predicted drug target data used in the DTSEA paper*

Description

We manually downloaded the predicted drug target data on the *Target Predictions* subsections of the Compound Report Card in each drug. Also, you can run the prediction in your local machine using docker image `chembl/mcp`.

Usage

```
drug_predicted
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 2401 rows and 10 columns.

Details

A data frame with 2401 rows and 10 columns:

- `drugbank_id`: the DrugBank ID
- `drug_name`: the name of each drug
- `gene_target`: the targets of drugs
- `drugbank_id`: the DrugBank ID.
- `drug_name`: the name of each drug.
- `chembl_id`: the ChEMBL ID of each drug.
- `uniport_id`: the UniPort ID of each protein.
- `hgnc_symbol`: the HGNC symbol of each gene.
- `target`: the ChEMBL ID of each target.
- `seventy`: the 70% confidence level of the prediction. The results are in four levels: **active**, **inactive**, **empty**, and **both**. The four levels represent the following:
  - **Active**: the drug is predicted to interact with the target.
  - **Inactive**: the drug is not predicted to interact with the target.
  - **Empty**: the model can not predict the compound interacts with the target.
  - **Both**: The model can not conclude the interaction.
- `eighty`: the 80% confidence level of the prediction. The results are in four levels: **active**, **inactive**, **empty**, and **both**. The four levels represent the following:
  - **Active**: the drug is predicted to interact with the target.
  - **Inactive**: the drug is not predicted to interact with the target.
  - **Empty**: the model can not predict the compound interacts with the target.
  - **Both**: The model can not conclude the interaction.
- `ninty`: the 90% confidence level of the prediction. The results are in four levels: **active**, **inactive**, **empty**, and **both**. The four levels represent the following:
  - **Active**: the drug is predicted to interact with the target.
  - **Inactive**: the drug is not predicted to interact with the target.
  - **Empty**: the model can not predict the compound interacts with the target.
  - **Both**: The model can not conclude the interaction.
- `threshold`: activity threshold.

References

Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Res. 2017 Jan 4;45(D1):D945-D954. doi: 10.1093/nar/gkw1074. Epub 2016 Nov 28. PMID: 27899562; PMCID: PMC5210557.

Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform. 2019 Jan 10;11(1):4. doi: 10.1186/s13321-018-0325-4. PMID: 30631996; PMCID: PMC6690068.

Examples

```
library(DTSEAdata)
data("drug_predicted")

# Get the drug count
library(dplyr)
pull(drug_predicted, drug_name) %>%
  unique() %>%
  length()
```

drug_targets *The drug target data used in the DTSEA paper*

Description

Drug-target interactions were downloaded and integrated from DrugBank and ChEMBL.

Usage

```
drug_targets
```

Format

An object of class `data.frame` with 17160 rows and 3 columns.

Details

A data frame with 17160 rows and 3 columns:

- `drugbank_id`: the DrugBank ID
- `drug_name`: the name of each drug
- `gene_target`: the targets of drugs

References

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D1074-D1082. doi: 10.1093/nar/gkx1037. PMID: 29126136; PMCID: PMC5753335.

Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Res. 2017 Jan 4;45(D1):D945-D954. doi: 10.1093/nar/gkw1074. Epub 2016 Nov 28. PMID: 27899562; PMCID: PMC5210557.

Examples

```
library(DTSEAdata)
data("drug_targets")

# Get the drug count and target count
library(dplyr)
pull(drug_targets, drugbank_id) %>%
  unique() %>%
  length() %>%
  cat('Drug count:', .)

pull(drug_targets, gene_target) %>%
  unique() %>%
  length() %>%
  cat('Target count:', .)
```

`expr_raw` *The raw expression profile used in the DTSEA paper*

Description

We provide the processed expression profile of GSE164805.

Usage

```
expr_raw
```

Format

An object of class `data.frame` with 20180 rows and 15 columns.

Details

A data frame contains 20180 rows and 15 columns.

- column: the patient label ID. IDs greater than **0** are disease samples, and less than **0** are controls.
- row: the gene symbol. The probes were converted into gene symbol and aggregated the repeated ones by mean.

References

Zhang Q, Meng Y, Wang K, Zhang X, Chen W, Sheng J, Qiu Y, Diao H, Li L. Inflammation and Antiviral Immune Response Associated With Severe Progression of COVID-19. Front Immunol. 2021 Feb 18;12:631226. doi: 10.3389/fimmu.2021.631226. PMID: 33679778; PMCID: PMC7930228.

Examples

```
library(DTSEAdata)
data("expr_raw")

# Get the data length
library(dplyr)
dim(expr_raw) %>%
  cat(sep = '*')
```

---

graph — *The human gene functional interaction network object used in the DTSEA paper*

Description

We extracted the gene functional interaction network from multiple sources with experimental evidence and then integrated them.

Usage

```
graph
```

Format

An `igraph` object with 221,353 edges and 12,836 nodes.

References

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021 Jan 8;49 (D1):D545-D551. doi: 10.1093/nar/gkaa970. PMID: 33125081; PMCID: PMC7779016.

Examples

```
library(DTSEAdata)
data("graph")

# Get graph characteristics
library(igraph)
cat(length(V(graph)), 'nodes,', length(E(graph)), 'edges. \n')
```

`ncbi_list` *The COVID-19-related genes provided by NCBI*

Description

NCBI provided a COVID-19-related node list (https://www.ncbi.nlm.nih.gov/gene/?term=coronavirus+related+%5Bprop
The WayBackMachine did not snapshot the page on July 2022, so we manually took a snapshot and obtained this gene list.

Usage

```
ncbi_list
```

Format

A `list` object with 190 gene symbols.

Examples

```
library(DTSEAdata)
data("ncbi_list")

# Get vector size
length(ncbi_list)
```

Index