

# Package ‘SEMdata’

January 5, 2021

**Type** Package

**Title** Datasets for SEM-based network analysis using the SEMgraph R package

**Version** 0.1.0

**Date** 2020-12-01

**Author** Mario Grassi [aut], Fernando Palluzzi [aut, cre], Daniele Pepe [ctb]

**Maintainer** Fernando Palluzzi <fernando.palluzzi@gmail.com>

**Description** This package contains a collection of datasets and reference interaction networks for graph analysis in computational biology.

**License** GPL-3

**Encoding** UTF-8

**LazyData** TRUE

**Depends** R (>= 4.0)

**Suggests** BiocStyle, graphite, igraph, knitr, org.Hs.eg.db, rmarkdown, SEMgraph

**VignetteBuilder** knitr

**RoxygenNote** 7.1.1

**NeedsCompilation** no

## R topics documented:

alsData . . . . .	2
ftdDName . . . . .	3
kegg . . . . .	4
kegg.pathways . . . . .	5
string . . . . .	6
<b>Index</b>	<b>7</b>

---

`alsData`*Amyotrophic Lateral Sclerosis (ALS) dataset*

---

**Description**

Expression profiling through high-throughput sequencing (RNA-seq) of 139 ALS patients and 21 healthy controls (HCs), from Tam et al. (2019).

**Usage**`alsData`**Format**

`alsData` is a list of 4 objects:

1. "graph", ALS graph as the largest connected component of the "Amyotrophic lateral sclerosis (ALS)" pathway from KEGG database;
2. "exprs", a matrix of 160 rows (subjects) and 17695 columns (genes). Raw data from the GEO dataset GSE124439 (Tam et al., 2019) were pre-processed applying batch effect correction, using the `sva` R package (Leek et al., 2012), to remove data production center and brain area biases. Using multidimensional scaling-based clustering, ALS-specific and an HC-specific clusters were generated. Misclassified samples were blacklisted and removed from the current dataset;
3. "group", a binary group vector of 139 ALS subjects (1) and 21 healthy controls (0);
4. "details", a data.frame reporting information about included and blacklisted samples.

**Source**

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124439>

**References**

Tam OH, Rozhkov NV, Shaw R, Kim D et al. (2019). Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Reppts*, 29(5):1164-1177.e5. DOI: <https://doi.org/10.1016/j.celrep.2019.09.066>

Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey (2012). The `sva` package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. Mar 15; 28(6): 882-883. <https://doi.org/10.1093/bioinformatics/bts034>

**Examples**

```
alsData$graph
dim(alsData$exprs)
table(alsData$group)
```

---

ftdDName	<i>Frontotemporal Dementia (FTD) DNA methylation dataset</i>
----------	--

---

## Description

DNA methylation (DName) profiling by genome tiling array of 105 FTD patients and 150 healthy controls from peripheral blood samples (Li et al., 2014).

## Usage

```
ftdDName
```

## Format

ftdDName is a list of 2 objects:

1. "pc1", a data matrix of 256 rows (subjects) and 16560 columns (genes) containing the value of the first principal component of DName levels, obtained applying a principal component analysis to methylated CpG sites within the promoter region, for each gene (genes with unmethylated CpGs in both conditions were discarded);
2. "group", a binary group vector of 105 FTD patients (1) and 150 healthy controls (0).

## Source

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53740>

## References

Li Y, Chen JA, Sears RL, Gao F et al. An epigenetic signature in peripheral blood associated with the haplotype on 17q21.31, a risk factor for neurodegenerative tauopathy. PLoS Genet 2014 Mar;10(3):e1004211. <https://doi.org/10.1371/journal.pgen.1004211>

## Examples

```
dim(ftdDName$pc1)
table(ftdDName$group)
```

---

kegg

*KEGG interactome*

---

## Description

Interactome generated by merging KEGG pathways extracted using the graphite R package (update: October, 2019).

## Usage

kegg

## Format

"kegg" is an igraph network object of 5934 nodes and 77158 edges corresponding to the union of 306 KEGG pathways.

## Source

<https://www.kegg.jp/kegg>

## References

Kanehisa M, Goto S (1999). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acid Research 28(1): 27-30. <https://doi.org/10.1093/nar/27.1.29>

Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 41(1):e19. <https://doi.org/10.1093/nar/gks866>.

## Examples

```
kegg
summary(kegg)

# KEGG degrees of freedom
vcount(kegg)*(vcount(kegg) - 1)/2 - ecount(kegg)

# KEGG average shortest path length
mean_distance(kegg)
```

---

kegg.pathways*KEGG signaling pathways collection*

---

**Description**

Collection of KEGG signaling pathways extracted using the graphite R package (update: October, 2019).

**Usage**

```
kegg.pathways
```

**Format**

"kegg.pathways" is a list of 306 KEGG signaling pathways, stored as igraph objects.

**Source**

<https://www.kegg.jp/kegg>

**References**

Kanehisa M, Goto S (1999). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acid Research 28(1): 27-30. <https://doi.org/10.1093/nar/27.1.29>

Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 41(1):e19. <https://doi.org/10.1093/nar/gks866>.

**Examples**

```
## NOT RUN ## {

# Number of nodes per pathway
kegg.nodes <- unlist(lapply(kegg.pathways, graph::numNodes))
# Number of edges per pathway
kegg.edges <- unlist(lapply(kegg.pathways, graph::numEdges))
# Gene list per pathway
kegg.genes <- unlist(lapply(kegg.pathways, graph::nodes))
quantile(kegg.nodes)
quantile(kegg.edges)
length(unique(kegg.genes)) # Number of unique genes within the dataset

# Loading breast cancer KEGG network
i <- which(names(kegg.pathways) == "Breast cancer")
bc.graph <- kegg.pathways[[i]]
bc.graph

## }
```

---

`string`

---

*STRING interactome***Description**

STRING interactome version 10.5.

**Usage**

`string`

**Format**

"string" is an igraph network object of 9725 nodes and 170987 edges corresponding to the STRING interactome (version 10.5).

**Source**

<https://string-db.org>

**References**

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, von Mering C (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res., 47: D607-613. <https://doi.org/10.1093/nar/gky1131>

**Examples**

```
string
summary(string)

# STRING degrees of freedom
vcount(string)*(vcount(string) - 1)/2 - ecount(string)

# STRING average shortest path length
mean_distance(string)
```

# Index

alsData, [2](#)

ftdDName, [3](#)

kegg, [4](#)

kegg.pathways, [5](#)

string, [6](#)