Package ‘GOSemSim’

May 2, 2024

Type Package
Title GO-terms Semantic Similarity Measures
Version 2.30.0
Maintainer Guangchuang Yu <guangchuangyu@gmail.com>
Description The semantic comparisons of Gene Ontology (GO) annotations provide quantitative ways to compute similarities between genes and gene groups, and have became important basis for many bioinformatics analysis approaches. GOSemSim is an R package for semantic similarity computation among GO terms, sets of GO terms, gene products and gene clusters. GOSemSim implemented five methods proposed by Resnik, Schlicker, Jiang, Lin and Wang respectively.
Depends R (>= 3.5.0)
LinkingTo Rcpp
Imports AnnotationDbi, GO.db, methods, rlang, stats, utils, yulab.utils
Suggests AnnotationHub, BiocManager, clusterProfiler, DOSE, knitr, org.Hs.eg.db, prettydoc, readr, rmarkdown, testthat, tidyselect, ROCR
VignetteBuilder knitr
ByteCompile true
License Artistic-2.0
Encoding UTF-8
BugReports https://github.com/YuLab-SMU/GOSemSim/issues
biocViews Annotation, GO, Clustering, Pathways, Network, Software
RoxygenNote 7.3.0
git_url https://git.bioconductor.org/packages/GOSemSim
git_branch RELEASE_3_19
git_last_commit f155e5a
git_last_commit_date 2024-04-30
**Repository**  Bioconductor 3.19

**Date/Publication**  2024-05-02

**Author**  Guangchuang Yu [aut, cre],
Alexey Stukalov [ctb],
Pingfan Guo [ctb],
Chuanle Xiao [ctb],
Lluís Revilla Sancho [ctb]

## Contents

- GOSemSim-package .................................................. 2
- buildGOmap ......................................................... 3
- clusterSim .......................................................... 4
- combineScores ..................................................... 5
- geneSim ............................................................. 5
- godata ............................................................... 6
- GOSemSimDATA-class .............................................. 7
- goSim ............................................................... 8
- go_term_table ...................................................... 9
- infoContentMethod ................................................ 9
- load_OrgDb ....................................................... 10
- mclusterSim ...................................................... 10
- mgeneSim ........................................................ 11
- mgoSim ........................................................... 12
- read.blast2go .................................................... 13
- read.gaf .......................................................... 14
- tcss_cutoff ........................................................ 14
- termSim ........................................................... 16
- wangMethod_internal ............................................. 16

**Index**  18

---

**GOSemSim-package**  

**GOSemSim: GO-terms Semantic Similarity Measures**

---

**Description**

The semantic comparisons of Gene Ontology (GO) annotations provide quantitative ways to compute similarities between genes and gene groups, and have became important basis for many bioinformatics analysis approaches. GOSemSim is an R package for semantic similarity computation among GO terms, sets of GO terms, gene products and gene clusters. GOSemSim implemented five methods proposed by Resnik, Schlicker, Jiang, Lin and Wang respectively.
**buildGOmap**

**Author(s)**

**Maintainer**: Guangchuang Yu <guangchuangyu@gmail.com>

Other contributors:

- Alexey Stukalov <astukalov@gmail.com> [contributor]
- Pingfan Guo <1178431277@qq.com> [contributor]
- Chuanle Xiao <xiaochuanle@126.com> [contributor]
- Lluís Revilla Sancho <lluis.revilla@gmail.com> [contributor]

**See Also**

Useful links:

- [https://yulab-smu.top/biomedical-knowledge-mining-book/](https://yulab-smu.top/biomedical-knowledge-mining-book/)
- Report bugs at [https://github.com/YuLab-SMU/GOSemSim/issues](https://github.com/YuLab-SMU/GOSemSim/issues)

---

**Description**

Adding indirect GO annotation

**Usage**

`buildGOmap(TERM2GENE)`

**Arguments**

- `TERM2GENE` data.frame with two or three columns of GO TERM, GENE and ONTOLOGY (optional)

**Details**

provided by a data.frame of GO TERM (column 1), GENE (column 2) and ONTOLOGY (optional) that describes GO direct annotation, this function will add indirect GO annotation of genes.

**Value**

data.frame, GO annotation with direct and indirect annotation

**Author(s)**

Yu Guangchuang
clusterSim

Semantic Similarity Between Two Gene Clusters

Description

Given two gene clusters, this function calculates semantic similarity between them.

Usage

```
clusterSim(
  cluster1,cluster2,semData,
  measure = "Wang",
  drop = "IEA",
  combine = "BMA"
)
```

Arguments

- `cluster1`: A set of gene IDs.
- `cluster2`: Another set of gene IDs.
- `semData`: GOSemSimDATA object
- `measure`: One of "Resnik", "Lin", "Rel", "Jiang", "TCSS" and "Wang" methods.
- `drop`: A set of evidence codes based on which certain annotations are dropped. Use NULL to keep all GO annotations.
- `combine`: One of "max", "avg", "rcmax", "BMA" methods, for combining semantic similarity scores of multiple GO terms associated with protein or multiple proteins associated with protein cluster.

Value

`similarity`

References


See Also

gosim mgoSim geneSim mgeneSim mclusterSim
**combineScores**

Combining Similarity Matrix to Similarity Score

**Examples**

```r
d <- godata('org.Hs.eg.db', ont="MF", computeIC=FALSE)
c1 <- c("835", "5261", "241", "994")
c2 <- c("307", "308", "317", "321", "506", "378", "388", "396")
clusterSim(cluster1, cluster2, semData=d, measure="Wang")
```

**Description**

Functions for combining similarity matrix to similarity score

**Usage**

```r
combineScores(SimScores, combine)
```

**Arguments**

- `SimScores`: Similarity matrix
- `combine`: Combine method

**Value**

Similarity value

**Author(s)**

Guangchuang Yu [http://guangchuangyu.github.io](http://guangchuangyu.github.io)

---

**geneSim**

Semantic Similarity Between Two Genes

**Description**

Given two genes, this function will calculate the semantic similarity between them, and return their semantic similarity and the corresponding GO terms

**Usage**

```r
geneSim(gene1, gene2, semData, measure = "Wang", drop = "IEA", combine = "BMA")
```
Arguments

gene1 Entrez gene id.
gene2 Another entrez gene id.
semData GOSemSimDATA object
measure One of "Resnik", "Lin", "Rel", "Jiang" "TCSS" and "Wang" methods.
drop A set of evidence codes based on which certain annotations are dropped. Use NULL to keep all GO annotations.
combine One of "max", "avg", "rcmax", "BMA" methods, for combining semantic similarity scores of multiple GO terms associated with protein or multiple proteins associated with protein cluster.

Value

list of similarity value and corresponding GO.

References


See Also
goSim mgoSim mgeneSim clusterSim mclusterSim

Examples

d <- godata('org.Hs.eg.db', ont="MF", computeIC=FALSE)
geneSim("241", "251", semData=d, measure="Wang")

godata  godata

description

prepare GO DATA for measuring semantic similarity

Usage

godata(
    OrgDb = NULL,
    annoDb = NULL,
    keytype = "ENTREZID",
    ont,
    computeIC = TRUE,
GOSemSimDATA-class

```
processTCSS = FALSE,
cutoff = NULL
```

Arguments

- `OrgDb` OrgDb object (will be removed in future, please use annoDb instead)
- `annoDb` GO annotation database, can be OrgDb or a data.frame contains three columns of ‘GENE’, ‘GO’ and ‘ONTOLOGY’.
- `keytype` keytype
- `ont` one of ‘BP’, ‘MF’, ‘CC’
- `computeIC` logical, whether compute IC
- `processTCSS` logical, whether to process TCSS
- `cutoff` cutoff of TCSS

Value

GOSemSimDATA object

Author(s)

Guangchuang Yu

---

Class “GOSemSimDATA” This class stores IC and gene to go mapping for semantic similarity measurement

Description

Class "GOSemSimDATA" This class stores IC and gene to go mapping for semantic similarity measurement

Slots

- `keys` gene ID
- `ont` ontology
- `IC` IC data
- `geneAnno` gene to GO mapping
- `tcssdata` tcssdata
- `metadata` metadata
goSim

Semantic Similarity Between Two GO Terms

Description

Given two GO IDs, this function calculates their semantic similarity.

Usage

goSim(GOID1, GOID2, semData, measure = "Wang")

Arguments

- GOID1: GO ID 1.
- GOID2: GO ID 2.
- semData: GOSemSimDATA object
- measure: One of "Resnik", "Lin", "Rel", "Jiang", "TCSS" and "Wang" methods.

Value

similarity

References


See Also

mgoSim geneSim mgeneSim clusterSim mclusterSim

Examples

d <- godata('org.Hs.eg.db', ont="MF", computeIC=FALSE)
goSim("GO:0004022", "GO:0005515", semData=d, measure="Wang")
go_term_table

<table>
<thead>
<tr>
<th>go_term_table</th>
<th>Information content of GO terms</th>
</tr>
</thead>
</table>

**Description**

These datasets are the information contents of GO terms.

**References**


<table>
<thead>
<tr>
<th>infoContentMethod</th>
<th>information content based methods</th>
</tr>
</thead>
</table>

**Description**

Information Content Based Methods for semantic similarity measuring

**Usage**

```
infoContentMethod(ID1, ID2, method, godata)
```

**Arguments**

- **ID1**: Ontology Term
- **ID2**: Ontology Term
- **method**: one of "Resnik", "Jiang", "Lin" and "Rel", "TCSS".
- **godata**: GOSemSimDATA object

**Details**

implemented for methods proposed by Resnik, Jiang, Lin and Schlicker.

**Value**

semantic similarity score

**Author(s)**

Guangchuang Yu https://guangchuangyu.github.io
**Description**

load OrgDb

**Usage**

load_OrgDb(OrgDb)

**Arguments**

- **OrgDb**
  - OrgDb object or OrgDb name

**Value**

OrgDb object

**Author(s)**

Guangchuang Yu

---

**mclusterSim**

Pairwise Semantic Similarities for a List of Gene Clusters

**Description**

Given a list of gene clusters, this function calculates pairwise semantic similarities.

**Usage**

mclusterSim(clusters, semData, measure = "Wang", drop = "IEA", combine = "BMA")

**Arguments**

- **clusters**
  - A list of gene clusters.
- **semData**
  - GOSemSimDATA object
- **measure**
  - One of "Resnik", "Lin", "Rel", "Jiang", "TCSS" and "Wang" methods.
- **drop**
  - A set of evidence codes based on which certain annotations are dropped. Use NULL to keep all GO annotations.
- **combine**
  - One of "max", "avg", "rcmax", "BMA" methods, for combining semantic similarity scores of multiple GO terms associated with protein or multiple proteins associated with protein cluster.
mgeneSim

Value

similarity matrix

References


See Also

goSim mgoSim geneSim mgeneSim clusterSim

Examples

d <- godata('org.Hs.eg.db', ont="MF", computeIC=FALSE)
cluster1 <- c("835", "5261", "241")
cluster2 <- c("578", "582")
cluster3 <- c("307", "308", "317")
clusters <- list(a=cluster1, b=cluster2, c=cluster3)
mclusterSim(clusters, semData=d, measure="Wang")

mgeneSim

Pairwise Semantic Similarity for a List of Genes

Description

Given a list of genes, this function calculates pairwise semantic similarities.

Usage

mgeneSim(
  genes,
  semData,
  measure = "Wang",
  drop = "IEA",
  combine = "BMA",
  verbose = TRUE
)

Arguments

genesis A list of entrez gene IDs.
semData A GOSemSimDATA object
measure One of "Resnik", "Lin", "Rel", "Jiang", "TCSS" and "Wang" methods.
mgoSim

Description

Given two GO term sets, this function will calculate the semantic similarity between them, and return their semantic similarity.

Usage

mgoSim(GO1, GO2, semData, measure = "Wang", combine = "BMA")

Arguments

GO1 A set of go terms.
GO2 Another set of go terms.
semData GOSemSimDATA object
measure One of "Resnik", "Lin", "Rel", "Jiang", "TCSS" and "Wang" methods.
combine One of "max", "avg", "rcmax", "BMA" methods, for combining semantic similarity scores of multiple GO terms associated with protein or multiple proteins associated with protein cluster.

Value

similarity matrix

References


See Also
goSIm mgoSim geneSim clusterSim mclusterSim

Examples

d <- godata('org.Hs.eg.db', ont="MF", computeIC=FALSE)
mgeneSim(c("835", "5261","241"), semData=d, measure="Wang")
References


See Also

`goSim`, `geneSim`, `mgeneSim`, `clusterSim`, `mclusterSim`

Examples

```r
d <- godata('org.Hs.eg.db', ont="MF", computeIC=FALSE)
go1 <- c("GO:0004022", "GO:0004024", "GO:0004023")
go2 <- c("GO:0009055", "GO:0020037")
mgoSim("GO:0003824", go2, semData=d, measure="Wang")
mgoSim(go1, go2, semData=d, measure="Wang")
```

Description

given a BLAST2GO file, this function extracts the information from it and make it use for TERM2GENE.

Usage

```r
read.blast2go(file, add_indirect.GO = FALSE)
```

Arguments

- `file` : BLAST2GO file
- `add_indirect.GO` : whether add indirect GO annotation

Value

a data frame with three columns: GENE, GO and ONTOLOGY
**read.gaf**

**Description**

parse GAF files

**Usage**

```r
read.gaf(file, asis = FALSE, add_indirect_GO = FALSE)
```

```r
parse_gff(file, asis = FALSE, add_indirect_GO = FALSE)
```

**Arguments**

- `file`  
  GAF file
- `asis`  
  logical, whether output the original contains of the file and only works if `add_indirect_GO = FALSE`
- `add_indirect_GO`  
  whether to add indirect GO annotation

**Details**

given a GAF file, this function extracts the information from it

**Value**

A data.frame. Original table if `asis` works, otherwise contains 3 columns of 'GENE', 'GO' and 'ONTOLOGY'

---

**tcss_cutoff**

**determine the topological cutoff for TCSS method**

**Description**

determine the topological cutoff for TCSS method

**Usage**

```r
tcss_cutoff(
  OrgDb = NULL,
  keytype = "ENTREZID",
  ont,
  combine_method = "max",
  ppidata
)
```
Arguments

OrgDb
OrgDb object

keytype
keytype

ont
ontology: "BP", "MF", "CC"

combine_method
"max", "BMA", "avg", "rcmax", "rcmax.avg"

ppidata
A data.frame contains positive set and negative set. Positive set is PPI pairs that already verified. ppidata has three columns, column 1 and 2 are character, column 3 must be logical value: TRUE/FALSE.

Value

numeric, topological cutoff for given parameters

Examples

```r
## Not run:
library(org.Hs.eg.db)
library(STRINGdb)
string_db <- STRINGdb$new(version = "11.0", species = 9606,
score_threshold = 700)
string_proteins <- string_db$get_proteins()

#get relationship
ppi <- string_db$get_interactions(string_proteins$protein_external_id)

ppi$from <- vapply(ppi$from, function(e)
  strsplit(e, "9606.")[[1]][2], character(1))
ppi$to <- vapply(ppi$to, function(e)
  strsplit(e, "9606.")[[1]][2], character(1))

len <- nrow(ppi)

#select length
s_len <- 100
pos_1 <- sample(len, s_len, replace = T)

#negative set
pos_2 <- sample(len, s_len, replace = T)
pos_3 <- sample(len, s_len, replace = T)

#union as ppidata
ppidata <- data.frame(pro1 = c(ppi$from[pos_1], ppi$from[pos_2]),
  pro2 = c(ppi$to[pos_1], ppi$to[pos_3]),
  label = c(rep(TRUE, s_len), rep(FALSE, s_len)),
  stringsAsFactors = FALSE)

cutoff <- tcss_cutoff(OrgDb = org.Hs.eg.db, keytype = "ENSEMBLPROT",
  ont = "BP", combine_method = "max", ppidata)

## End(Not run)
```
Description
measuring similarities between two term vectors.

Usage
termSim(
t1,
t2,
semData,
method = c("Wang", "Resnik", "Rel", "Jiang", "Lin", "TCSS")
)

Arguments
t1 term vector
t2 term vector
semData GOSemSimDATA object
method one of "Wang", "Resnik", "Rel", "Jiang", and "Lin", "TCSS".

Details
provide two term vectors, this function will calculate their similarities.

Value
score matrix

Author(s)
Guangchuang Yu http://guangchuangyu.github.io

Description
Method Wang for semantic similarity measuring

Usage
wangMethod_internal(ID1, ID2, ont = "BP")
Arguments

<table>
<thead>
<tr>
<th>ID1</th>
<th>Ontology Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID2</td>
<td>Ontology Term</td>
</tr>
<tr>
<td>ont</td>
<td>Ontology</td>
</tr>
</tbody>
</table>

Value

semantic similarity score

Author(s)

Guangchuang Yu [http://ygc.name](http://ygc.name)
Index

* **classes**
  GOSemSimDATA-class, 7

* **datasets**
  go_term_table, 9

* **internal**
  GOSemSim-package, 2

* **manip**
  clusterSim, 4
  geneSim, 5
  goSim, 8
  mclusterSim, 10
  mgeneSim, 11
  mgoSim, 12

  buildGOmap, 3

  clusterSim, 4, 6, 8, 11–13
  combineScores, 5

  geneSim, 4, 5, 8, 11–13
  GO (go_term_table), 9
  go_term_table, 9
  godata, 6
  GOSemSim (GOSemSim-package), 2
  GOSemSim-package, 2
  GOSemSimDATA-class, 7
  goSim, 4, 6, 8, 11–13
  gotbl (go_term_table), 9

  infoContentMethod, 9

load_OrgDb, 10

mclusterSim, 4, 6, 8, 10, 12, 13
mgeneSim, 4, 6, 8, 11, 11, 13
mgoSim, 4, 6, 8, 11, 12, 12

parse_gff (read.gaf), 14

read.blast2go, 13
read.gaf, 14

show, GOSemSimDATA-method (GOSemSimDATA-class), 7

tcss_cutoff, 14

termSim, 16

wangMethod_internal, 16