

# Package ‘GOSemSim’

April 15, 2024

**Type** Package

**Title** GO-terms Semantic Similarity Measures

**Version** 2.28.1

**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 become 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,  
readr, rmarkdown, org.Hs.eg.db, prettydoc, testthat, tidyr,  
tidyselect, ROCR

**VignetteBuilder** knitr

**ByteCompile** true

**License** Artistic-2.0

**Encoding** UTF-8

**URL** <https://yulab-smu.top/biomedical-knowledge-mining-book/>

**BugReports** <https://github.com/YuLab-SMU/GOSemSim/issues>

**biocViews** Annotation, GO, Clustering, Pathways, Network, Software

**RoxygenNote** 7.2.3

**git\_url** <https://git.bioconductor.org/packages/GOSemSim>

**git\_branch** RELEASE\_3\_18

**git\_last\_commit** 84618af

**git\_last\_commit\_date** 2024-01-16

**Repository** Bioconductor 3.18

**Date/Publication** 2024-04-15

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

## R topics documented:

GOSemSim-package . . . . .	2
buildGOMap . . . . .	3
clusterSim . . . . .	4
combineScores . . . . .	5
geneSim . . . . .	6
godata . . . . .	7
GOSemSimDATA-class . . . . .	8
goSim . . . . .	8
go_term_table . . . . .	9
infoContentMethod . . . . .	9
load_OrgDb . . . . .	10
mclusterSim . . . . .	10
mgeneSim . . . . .	11
mgoSim . . . . .	13
read.blast2go . . . . .	14
read.gaf . . . . .	14
tcss_cutoff . . . . .	15
termSim . . . . .	16
wangMethod_internal . . . . .	17
<b>Index</b>	<b>18</b>

---

GOSemSim-package

*Gene Ontology-based Semantic Similarity Measures*

---

### Description

Implementation of semantic similarity measures to estimate the functional similarities among Gene Ontology terms and gene products

### Details

Quantitative measure of functional similarities among gene products is important for post-genomics study. and widely used in gene function prediction, cluster analysis and pathway modeling. This package is designed to estimate the GO terms' and genes' semantic similarities. Implemented five methods proposed by Resnik, Schlicker, Jiang, Lin and Wang respectively. Support many species,

including Anopheles, Arabidopsis, Bovine, Canine, Chicken, Chimp, E coli strain K12 and strain Sakai, Fly, Human, Malaria, Mouse, Pig, Rhesus, Rat, Worm, Xenopus, Yeast, Zebrafish.

Package: GOSemSim  
Type: Package  
Version: 2.0.0  
Date: 09-11-2012  
biocViews: GO, Clustering, Pathways, Anopheles\_gambiae, Arabidopsis\_thaliana, Bos\_taurus, Caenorhabditis\_elegans, Ca  
Depends:  
Imports: methods, AnnotationDbi, GO.db  
Suggests: clusterProfiler, DOSE  
License: Artistic-2.0

### Author(s)

Guangchuang Yu

Maintainer: Guangchuang Yu <guangchuangyu@gmail.com>

### References

Yu et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products *Bioinformatics* (Oxford, England), 26:7 976–978, April 2010. ISSN 1367-4803 <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976> PMID: 20179076

### See Also

[goSim](#) [mgoSim](#) [geneSim](#) [mgeneSim](#) [clusterSim](#) [mclusterSim](#)

---

buildGOMap

*buildGOMap*

---

### Description

Adding indirect GO annotation

### Usage

```
buildGOMap(x)
```

### Arguments

x data.frame with two or three columns of GENE, GO and ONTOLOGY (optional)

### Details

provided by a data.frame of GENE (column 1), GO (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**

Yu et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products *Bioinformatics* (Oxford, England), 26:7 976–978, April 2010. ISSN 1367-4803 <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976> PMID: 20179076

**See Also**

[goSim](#) [mgoSim](#) [geneSim](#) [mgeneSim](#) [mclusterSim](#)

**Examples**

```
d <- godata('org.Hs.eg.db', ont="MF", computeIC=FALSE)
cluster1 <- c("835", "5261", "241", "994")
cluster2 <- c("307", "308", "317", "321", "506", "540", "378", "388", "396")
clusterSim(cluster1, cluster2, semData=d, measure="Wang")
```

---

combineScores	<i>combining similarity matrix to similarity score</i>
---------------	--

---

**Description**

Functions for combining similarity matrix to similarity score

**Usage**

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

**Arguments**

SimScores	similarity matrix
combine	combine method

**Value**

similarity value

**Author(s)**

Guangchuang Yu <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**

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

**Arguments**

<code>gene1</code>	Entrez gene id.
<code>gene2</code>	Another entrez gene id.
<code>semData</code>	GOSemSimDATA object
<code>measure</code>	One of "Resnik", "Lin", "Rel", "Jiang" "TCSS" and "Wang" methods.
<code>drop</code>	A set of evidence codes based on which certain annotations are dropped. Use NULL to keep all GO annotations.
<code>combine</code>	One of "max", "avg", "rmax", "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**

Yu et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products *Bioinformatics* (Oxford, England), 26:7 976–978, April 2010. ISSN 1367-4803 <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976> PMID: 20179076

**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	<i>godata</i>
--------	---------------

---

**Description**

prepare GO DATA for measuring semantic similarity

**Usage**

```
godata(  
  OrgDb = NULL,  
  keytype = "ENTREZID",  
  ont,  
  computeIC = TRUE,  
  processTCSS = FALSE,  
  cutoff = NULL  
)
```

**Arguments**

OrgDb	OrgDb object
keytype	keytype
ont	one of 'BP', 'MF', 'CC'
computeIC	logical, whether computer IC
processTCSS	logical, whether to process TCSS
cutoff	cutoff of TCSS

**Value**

GOSemSimDATA object

**Author(s)**

Guangchuang Yu

---

GOSemSimDATA-class	<i>Class "GOSemSimDATA" This class stores IC and gene to go mapping for semantic similarity measurement</i>
--------------------	---

---

### 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	<i>Semantic Similarity Between Two GO Terms</i>
-------	---

---

### Description

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

### Usage

```
goSim(GO1D1, GO1D2, semData, measure = "Wang")
```

### Arguments

GO1D1	GO ID 1.
GO1D2	GO ID 2.
semData	GOSemSimDATA object
measure	One of "Resnik", "Lin", "Rel", "Jiang", "TCSS" and "Wang" methods.

### Value

similarity

### References

Yu et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products *Bioinformatics* (Oxford, England), 26:7 976–978, April 2010. ISSN 1367-4803 <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976> PMID: 20179076



**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	<i>Information content of GO terms</i>
---------------	--

---

**Description**

These datasets are the information contents of GOterms.

**References**

Yu et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products *Bioinformatics* (Oxford, England), 26:7 976–978, April 2010. ISSN 1367-4803 <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976> PMID: 20179076

---

infoContentMethod	<i>information content based methods</i>
-------------------	--

---

**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>

---

load\_OrgDb

*load\_OrgDb*

---

**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.

**Value**

similarity matrix

**References**

Yu et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products *Bioinformatics* (Oxford, England), 26:7 976–978, April 2010. ISSN 1367-4803 <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976> PMID: 20179076

**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**

genes	A list of entrez 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", "rmax", "BMA" methods, for combining semantic similarity scores of multiple GO terms associated with protein or multiple proteins associated with protein cluster.
verbose	show progress bar or not.

**Value**

similarity matrix

**References**

Yu et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products *Bioinformatics* (Oxford, England), 26:7 976–978, April 2010. ISSN 1367-4803 <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976> PMID: 20179076

**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")
```

---

`mgoSim`*Semantic Similarity Between two GO terms lists*

---

**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**

G01	A set of go terms.
G02	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

**References**

Yu et al. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products *Bioinformatics* (Oxford, England), 26:7 976–978, April 2010. ISSN 1367-4803 <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976> PMID: 20179076

**See Also**

[goSim](#) [geneSim](#) [geneSim](#) [clusterSim](#) [mclusterSim](#)

**Examples**

```
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")
```

---

read.blast2go	<i>read.blast2go</i>
---------------	----------------------

---

**Description**

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

**Usage**

```
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	<i>read.gaf</i>
----------	-----------------

---

**Description**

parse GAF files

**Usage**

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

```
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	<i>determine the topological cutoff for TCSS method</i>
-------------	---

---

**Description**

determine the topological cutoff for TCSS method

**Usage**

```
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**

```
## 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)

```

---

termSim

*termSim*


---

## 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>

---

wangMethod\_internal    *wangMethod*

---

**Description**

Method Wang for semantic similarity measuring

**Usage**

```
wangMethod_internal(ID1, ID2, ont = "BP")
```

**Arguments**

ID1	Ontology Term
ID2	Ontology Term
ont	Ontology

**Value**

semantic similarity score

**Author(s)**

Guangchuang Yu <http://ygc.name>

# Index

- \* **classes**
  - GOSemSimDATA-class, 8
- \* **datasets**
  - go\_term\_table, 9
- \* **manip**
  - clusterSim, 4
  - geneSim, 6
  - goSim, 8
  - mclusterSim, 10
  - mgeneSim, 11
  - mgoSim, 13
- \* **package**
  - GOSemSim-package, 2

buildGOMap, 3

clusterSim, 3, 4, 6, 9, 11–13

combineScores, 5

geneSim, 3, 5, 6, 9, 11–13

GO (go\_term\_table), 9

go\_term\_table, 9

godata, 7

GOSemSim (GOSemSim-package), 2

GOSemSim-package, 2

GOSemSimDATA-class, 8

goSim, 3, 5, 6, 8, 11–13

gotbl (go\_term\_table), 9

infoContentMethod, 9

load\_OrgDb, 10

mclusterSim, 3, 5, 6, 9, 10, 12, 13

mgeneSim, 3, 5, 6, 9, 11, 11, 13

mgoSim, 3, 5, 6, 9, 11, 12, 13

parse\_gff (read.gaf), 14

read.blast2go, 14

read.gaf, 14

show, GOSemSimDATA-method  
(GOSemSimDATA-class), 8

tcss\_cutoff, 15

termSim, 16

wangMethod\_internal, 17