

# Package ‘GOSemSim’

November 14, 2025

**Type** Package

**Title** GO-terms Semantic Similarity Measures

**Version** 2.36.0

**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 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, DBI, digest, GO.db, methods, rlang, R.utils, stats, utils, yulab.utils (>= 0.2.1)

**Suggests** AnnotationHub, BiocManager, clusterProfiler, DOSE, knitr, org.Hs.eg.db, prettydoc, readr, rmarkdown, testthat, tidy, 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** 7.3.2

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

**git\_branch** RELEASE\_3\_22

**git\_last\_commit** c1bf5c5

**git\_last\_commit\_date** 2025-10-29

**Repository** Bioconductor 3.22

**Date/Publication** 2025-11-13

**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          | 3         |
| combineScores       | 4         |
| geneSim             | 5         |
| godata              | 6         |
| GOSemSimDATA-class  | 6         |
| goSim               | 7         |
| go_term_table       | 8         |
| infoContentMethod   | 8         |
| load_OrgDb          | 9         |
| mclusterSim         | 9         |
| mgeneSim            | 10        |
| mgoSim              | 11        |
| read.blast2go       | 12        |
| read.gaf            | 12        |
| tcss_cutoff         | 13        |
| termSim             | 14        |
| wangMethod_internal | 15        |
| <b>Index</b>        | <b>16</b> |

---

|                  |  |
|------------------|--|
| GOSemSim-package | <i>GOSemSim: GO-terms Semantic Similarity Measures</i> |
|------------------|--|

---

## 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.

## 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/>
- Report bugs at <https://github.com/YuLab-SMU/GOSemSim/issues>

---

buildGOMap

*buildGOMap*

---

**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**

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**

|         |  |
|---------|--|
| 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**

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

*godata*


---

### Description

prepare GO DATA for measuring semantic similarity

### Usage

```
godata(
  OrgDb = NULL,
  annoDb = NULL,
  keytype = "ENTREZID",
  ont,
  computeIC = TRUE,
  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 computer IC  |
| processTCSS | logical, whether to process TCSS  |
| cutoff      | cutoff of TCSS  |

### Value

GOSemSimDATA object

### Author(s)

Guangchuang Yu

---

 GOSemSimDATA-class

*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(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

*Information content of GO terms*


---

### 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

*information content based methods*


---

### 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 | <i>load_OrgDb</i> |
|------------|-------------------|

---

**Description**

load OrgDb

**Usage**

```
load_OrgDb(OrgDb)
```

**Arguments**

OrgDb                      OrgDb object or OrgDb name

**Value**

OrgDb object

**Author(s)**

Guangchuang Yu <https://yulab-smu.top>

---

|             |   |
|-------------|---|
| mclusterSim | <i>Pairwise Semantic Similarities for a List of Gene Clusters</i> |
|-------------|---|

---

**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", "rcmax", "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**

|         |  |
|---------|--|
| 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

**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](#) [mgeneSim](#) [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

*read.blast2go*

---

**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

*read.gaf*

---

**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 <https://yulab-smu.top>

# Index

- \* **classes**
  - GOSemSimDATA-class, 6
- \* **datasets**
  - go\_term\_table, 8
- \* **internal**
  - GOSemSim-package, 2
- \* **manip**
  - clusterSim, 3
  - geneSim, 5
  - goSim, 7
  - mclusterSim, 9
  - mgeneSim, 10
  - mgoSim, 11
- buildGOMap, 3
- clusterSim, 3, 5, 7, 10–12
- combineScores, 4
- geneSim, 4, 5, 7, 10–12
- GO (go\_term\_table), 8
- go\_term\_table, 8
- godata, 6
- GOSemSim (GOSemSim-package), 2
- GOSemSim-package, 2
- GOSemSimDATA-class, 6
- goSim, 4, 5, 7, 10–12
- gotbl (go\_term\_table), 8
- infoContentMethod, 8
- load\_OrgDb, 9
- mclusterSim, 4, 5, 7, 9, 11, 12
- mgeneSim, 4, 5, 7, 10, 10, 12
- mgoSim, 4, 5, 7, 10, 11, 11
- parse\_gff (read.gaf), 12
- read.blast2go, 12
- read.gaf, 12
- show, GOSemSimDATA-method  
(GOSemSimDATA-class), 6
- tcss\_cutoff, 13
- termSim, 14
- wangMethod\_internal, 15