Package ‘GOCompare’

April 30, 2022

**Title**  Comprehensive GO Terms Comparison Between Species

**Version**  1.0.1

**Description**  Supports the assessment of functional enrichment analyses obtained for several lists of genes and provides a workflow to analyze them between two species via weighted graphs. Methods are described in Sosa et al. (2022) (to be submitted).

**URL**  https://github.com/ccsosa/GOCompare

**BugReports**  https://github.com/ccsosa/GOCompare/issues

**Depends**  R (>= 4.0.0),

**Imports**  base (>= 3.5), utils (>= 3.5), methods (>= 3.5), stats, grDevices, ape, vegan, ggplot2, ggrepel, igraph, parallel, stringr

**License**  GPL-3

**LazyData**  true

**Encoding**  UTF-8

**RoxygenNote**  7.1.2

**Suggests**  testthat (>= 3.0.0)

**Config/testthat/edition**  3

**NeedsCompilation**  no

**Author**  Chrystian Camilo Sosa [aut, cre, cph]

Diana Carolina Clavijo-Buriticá [aut],
Mauricio Alberto Quimbaya [aut],
Maria Victoria Díaz [ctb],
Camila Riccio Rengifo [ctb],
Nicolas López-Rozo [ctb],
Arlen James Mosquera [ctb],
Andrés Álvarez [ctb]

**Maintainer**  Chrystian Camilo Sosa <ccsosaa@javerianacali.edu.co>

**Repository**  CRAN

**Date/Publication**  2022-04-29 23:10:02 UTC
A _thaliana

**R topics documented:**

- GOCompare-package ................................................................. 2
- A _thaliana ............................................................................ 2
- A _thaliana_compress ............................................................... 3
- compareGOspecies ................................................................. 4
- comparison_ex_compress ....................................................... 5
- comparison_ex_compress_CH ................................................... 6
- evaluateCAT_species ............................................................... 7
- evaluateGO_species ............................................................... 8
- graphGOspecies ................................................................. 9
- graph_two_GOspecies ............................................................ 11
- H _sapiens ........................................................................... 12
- H _sapiens_compress ............................................................. 13
- mostFrequentGOs ................................................................. 14

**Index**

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOCompare-package</td>
<td><em>GOCompare: An R package to compare GO terms of a gene list and their orthologs</em></td>
</tr>
</tbody>
</table>

**Description**

GOCompare is an R package used to compare a GO terms list between two species.

**Details**

- **Package:** GOCompare
- **Type:** Package
- **Version:** 1.0.0
- **Date:** 2021-07-14
- **License:** GPL-3

**A _thaliana**

*A thaliana functional enrichment analysis of 2224 ortholog genes related to cancer-hallmarks*

**Description**

This dataset is the original dataset obtained for Clavijo-Buriticá (In preparation)
Usage

A_thaliana

Format

A data frame with 4063 rows and 6 variables:

- **Enrichment_FDR**: Numeric: False discovery rate values for the GO term
- **Genes_in_list**: numeric: Number of genes in the list of genes for a given GO term
- **Total_genes**: numeric: Number of genes in the genome of a species for a given GO term
- **Functional_Category**: character: GO term name or GO term id
- **Genes**: character: Genes found for a given GO term
- **feature**: character: A column representing the belonging of a group of comparison

Source

https://data.mendeley.com/datasets/myyy2wxd59/1

References

Clavijo-Buriticá, Sosa, C.C., Mosquera, A.J. Álvarez, A., Medina, J. Quimbaya, M.A. A systematic comparison of the molecular machinery associated with Cancer-Hallmarks between plants and humans reveals Arabidopsis thaliana as a useful model to understand specific carcinogenic events (to be submitted, Target journal: Plos Biology)

---

**A_thaliana_compress**

A thaliana functional enrichment analysis results for "AID","DCE","RCD","SPS" cancer-hallmarks

Description

This dataset is a subset of the original dataset obtained for Clavijo-Buriticá (In preparation)

Usage

A_thaliana_compress

Format

A data frame with 120 rows and 6 variables (30 GO terms per cancer hallmark):

- **Enrichment_FDR**: Numeric: False discovery rate values for the GO term
- **Genes_in_list**: numeric: Number of genes in the list of genes for a given GO term
- **Total_genes**: numeric: Number of genes in the genome of a species for a given GO term
- **Functional_Category**: character: GO term name or GO term id
- **Genes**: character: Genes found for a given GO term
- **feature**: character: A column representing the belonging of a group of comparison
**compareGOspecies**

**Source**
https://data.mendeley.com/datasets/myyy2wxd59/1

**References**
Clavijo-Buriticá, Sosa, C.C., Mosquera, A.J. Álvarez, A., Medina, J. Quimbaya, M.A. A systematic comparison of the molecular machinery associated with Cancer-Hallmarks between plants and humans reveals Arabidopsis thaliana as a useful model to understand specific carcinogenic events (to be submitted, Target journal: Plos Biology)

**Description**

compareGOspecies function provides a simple workflow to compare results of functional enrichment analysis for two species.

To use this function you will need two matrices with a column which, represents the features to be compared (e.g. feature). This function will extract the unique GO terms for two matrices and it will generate a presence-absence matrix where rows will represent a combination of categories and species (e.g H.sapiens AID) and columns will represent the GO terms analyzed. Further, this function will calculate Jaccard distances and it will provide as outputs a list with four slots: 1.) A principal coordinates analysis (PCoA) 2.) The Jaccard distance matrix 3.) A list of shared GO terms between species 4.) Finally, a list of the unique GO terms and the belonging to the respective species.

**Usage**

```
compareGOspecies(df1, df2, GOterm_field, species1, species2)
```

**Arguments**

- **df1**: A data frame with the results of a functional enrichment analysis for the species 1 with an extra column "feature" with the features to be compared
- **df2**: A data frame with the results of a functional enrichment analysis for the species 2 with an extra column "feature" with the features to be compared
- **GOterm_field**: This is a string with the column name of the GO terms (e.g; "Functional_Category")
- **species1**: This is a string with the species name for species 1 (e.g; "H. sapiens")
- **species2**: This is a string with the species name for species 2 (e.g; "A. thaliana")

**Value**

This function will return a list with four slots: graphics, distance shared.GO_list, and unique.GO_list
**Examples**

```r
# Loading example datasets
data(H_sapiens_compress)
data(A_thaliana_compress)

# Defining the column with the GO terms to be compared
GOterm_field <- "Functional_Category"

# Defining the species names
species1 <- "H. sapiens"
species2 <- "A. thaliana"

# Running function
x <- compareGOspecies(df1=H_sapiens_compress,
df2=A_thaliana_compress,
GOterm_field=GOterm_field,
species1=species1,
species2=species2)

## Not run:
# Displaying PCoA results
x$graphics

# Checking shared GO terms between species
print(tapply(x$shared_GO_list$feature,x$shared_GO_list$feature,length))

## End(Not run)
```

---

**Description**

This dataset is the results of running the compareGOspecies species and it is composed of four slots:

- **graphics**  PCoA graphics
- **distance** numeric: Jaccard distance matrix
- **shared_GO_list** data.frame with shared GO terms between species
- **unique_GO_list** data.frame with unique GO terms and their belonging two each species

**Usage**

```r
comparison_ex_compress
```

**Format**

An object of class list of length 4.
Source

https://data.mendeley.com/datasets/myy2wxd59/1

References

Clavijo-Buriticá, Sosa, C.C., Mosquera, A.J. Álvarez, A., Medina, J. Quimbaya, M.A. A systematic comparison of the molecular machinery associated with Cancer-Hallmarks between plants and humans reveals Arabidopsis thaliana as a useful model to understand specific carcinogenic events (to be submitted, Target journal: Plos Biology)

---

Functional enrichment analysis comparison between H. sapiens and A. thaliana for "DCE", and "RCD" cancer-hallmarks. This dataset contains 10 GO terms per category to allow a fast run of the function graph_two_GOspecies.

---

Description

This dataset is the results of running the compareGOspecies species and it is composed of three slots:

- **distance** numeric: Jaccard distance matrix
- **shared.GO_list** data.frame with shared GO terms between species
- **unique.GO_list** data.frame with unique GO terms and their belonging two each species

Usage

comparison_ex_compress_CH

Format

An object of class list of length 3.

Source

https://data.mendeley.com/datasets/myy2wxd59/1

References

Clavijo-Buriticá, Sosa, C.C., Mosquera, A.J. Álvarez, A., Medina, J. Quimbaya, M.A. A systematic comparison of the molecular machinery associated with Cancer-Hallmarks between plants and humans reveals Arabidopsis thaliana as a useful model to understand specific carcinogenic events (to be submitted, Target journal: Plos Biology)
evaluateCAT_species

Comprehensive comparison between species using categories and Pearson’s Chi-squared Tests

Description

evaluateGO_species provides a simple function to compare results of functional enrichment analysis for two species through the use of proportion tests or Pearson’s Chi-squared Tests and a False discovery rate correction

Usage

evaluateCAT_species(df1, df2, species1, species2, GOterm_field, test = "prop")

Arguments

df1
A data frame with the results of a functional enrichment analysis for the species 1 with an extra column "feature" with the features to be compared

df2
A data frame with the results of a functional enrichment analysis for the species 2 with an extra column "feature" with the features to be compared

species1
This is a string with the species name for the species 1 (e.g; "H. sapiens")

species2
This is a string with the species name for the species 2 (e.g; "A. thaliana")

GOterm_field
This is a string with the column name of the GO terms (e.g; "Functional_Category")

test
This is a string with the hypothesis test to be performed. Two options are provided, "prop" and "chi-squared" (default value="prop")

Value

This function will return a data.frame with the following fields:

<table>
<thead>
<tr>
<th>CAT</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>pvalue</td>
<td>p-value obtained through the use of Pearson’s Chi-squared Test</td>
</tr>
<tr>
<td>FDR</td>
<td>Multiple comparison correction for the p-value column</td>
</tr>
</tbody>
</table>

Examples

#Loading example datasets
data(H_sapiens)
data(A_thaliana)

#Defining the column with the GO terms to be compared
GOterm_field <- "Functional_Category"

#Defining the species names
species1 <- "H. sapiens"
species2 <- "A. thaliana"
evaluateGO_species

Comprehensive comparison between species using GO terms and Pearson's Chi-squared Tests

Description

evaluateGO_species provides a simple function to compare results of functional enrichment analysis for two species through the use of proportion tests or Pearson's Chi-squared Tests and a False discovery rate correction.

Usage

evaluateGO_species(df1, df2, species1, species2, GOterm_field, test = "prop")

Arguments

df1: A data frame with the results of a functional enrichment analysis for the species 1 with an extra column "feature" with the features to be compared.
df2: A data frame with the results of a functional enrichment analysis for the species 2 with an extra column "feature" with the features to be compared.
species1: This is a string with the species name for the species 1 (e.g; "H. sapiens").
species2: This is a string with the species name for the species 2 (e.g; "A. thaliana").
GOterm_field: This is a string with the column name of the GO terms (e.g; "Functional_Category").
test: This is a string with the hypothesis test to be performed. Two options are provided, "prop" and "chi-squared" (default value="prop").

Value

This function will return a data.frame with the following fields:

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO</td>
<td>GO term analyzed</td>
</tr>
<tr>
<td>pvalue</td>
<td>p-value obtained through the use of Pearson's Chi-squared Test</td>
</tr>
<tr>
<td>FDR</td>
<td>Multiple comparison correction for the p-value column</td>
</tr>
</tbody>
</table>

Examples
#Loading example datasets
data(H_sapiens)
data(A_thaliana)

#Defining the column with the GO terms to be compared
GOterm_field <- "Functional_Category"

#Defining the species names
species1 <- "H. sapiens"
species2 <- "A. thaliana"

#Running function
x <- evaluateGO_species(df1= H_sapiens, 
df2=A_thaliana, 
species1=species1, 
species2=species2, 
GOterm_field=GOterm_field, 
test="prop")

print(x)

---

**graphGOspecies**  
**Undirected network representation for the results of functional enrichment analysis for one species**

### Description

**graphGOspecies** is a function to create undirected graphs using two options:

1.) Nodes are GO terms such as biological processes and the edges are features. First, edges weights are calculated as the intersection where $\text{cat}(U) \cap \text{cat}(V)$ represents categories where the GO terms $U$ and $V$ are. While $nBP$ is the total number of biological processes represented by the GO terms (1). Finally, node weights are calculated as the sum of all $w(e)$ where the node is participant (2) (Please be patient, it requires a long time to finish).

$$w(e) = \frac{|\text{cat}(U) \cap \text{cat}(V)|}{nBP}$$

$$K_w(U) = \sum (w(U, V))$$

2.) Nodes are features, the edges are the number of GO terms such as biological processes in your gene lists. In this case the edge weights are calculated as the number of biological processes shared by a category expressed as $\text{BP}(U) \cap \text{BP}(V)$ $nBP$ is the total number of biological processes (3). Finally, the node weights is calculated as the sum of all $w(e)$ where the node is participant (4).

$$w(e) = \frac{|\text{BP}(U) \cap \text{BP}(V)|}{nBP}$$

$$K_w(U) = \sum (w(U, V))$$
Usage

graphGOspecies(
  df,
  GOterm_field,
  option = "Categories",
  numCores = 2,
  saveGraph = FALSE,
  outdir = NULL
)

Arguments

df A data frame with the results of a functional enrichment analysis for a species
with an extra column "feature" with the features to be compared

GOterm_field This is a string with the column name of the GO terms (e.g: "Functional.Category")

option (values: "GO" or "Categories"). This option allows create either a graph where
nodes are GO terms and edges are features or alternatively a graph where nodes
are features and edges are GO terms (default value="Categories")

numCores numeric, Number of cores to use for the process (default value numCores=2).
For the example below, only one core will be used

saveGraph logical, if TRUE the function will allow save the graph in graphml format

outdir This parameter will allow save the graph file in a folder described here (e.g:
"D:").This parameter only works when saveGraph=TRUE

Value

This function will return a list with two slots: edges and nodes. Edges represents an edge list and
their weights and nodes which represents the nodes and their respective weights

Examples

#Loading example datasets
data(H_sapiens_compress)

GOterm_field <- "Functional_Category"

#Running function
x <- graphGOspecies(df=H_sapiens_compress,
  GOterm_field=GOterm_field,
  option = "Categories",
  numCores=1,
  saveGraph=FALSE,
  outdir = NULL)
`graph_two.GOspecies` is a function to create undirected graphs to compare GO terms between two species using two options: 1.) Nodes are GO terms such as biological processes and the edges represent features for a species since the method creates a graph per species as well as shared GO terms between them. Edge weights are calculated as the intersection where \( \text{cat}(U) \cap \text{cat}(V) \) represents categories where the GO terms U and V are. \( nBP \) is the total number of biological processes represented by the GO terms (1). Node weights are calculated as the sum of all \( w(e) \) where the node is a participant (2) in each species and a shared GO terms(k) graphs.

\[
w(e) = \frac{\left| \text{cat}(U) \cap \text{cat}(V) \right|}{\left| nBP \right|}
\]

\[
K_w(U) = \sum (\sum (w(U,V)_{k = 1, k}))
\]

2.) Nodes are features and edges are GO terms available in the set of graphs (k) which consist of each species graphs and a shared GO terms graph (k). Two edges weights are calculated. First, edges weights are calculated as number of BP in the feature in comparison with the total number of GO terms available (3). Second, a shared weight is calculated for interactions shared between two species. Finally, node weights are calculated as the sum of all \( w(e) \) where the node is a participant (2) in each species and a shared GO terms(k) graphs.

\[
w(e) = \frac{\left| \text{BP}(U) \cap \text{BP}(V) \right|}{\left| nBP \right|}
\]

\[
K_w(U) = \sum (\sum (w(U,V)_{k = 1, k}))
\]

**Usage**

```r
graph_two.GOspecies(
  x,
  species1,
  species2,
  GOterm_field,
  saveGraph = FALSE,
  option = "Categories",
  numCores = 2,
  outdir = NULL
)
```
**Arguments**

- **x**: is a list obtained as output of the compareOspecies function.
- **species1**: This is a string with the species name for species 1 (e.g; "H. sapiens").
- **species2**: This is a string with the species name for species 2 (e.g; "A. thaliana").
- **GOterm_field**: This is a string with the column name of the GO terms (e.g; "Functional_Category").
- **saveGraph**: logical, if TRUE the function will allow save the graph in graphml format.
- **option**: (values: "Categories or "GO"). This option allows create either a graph where nodes are GO terms and edges are features and GO as well as species belonging are edges attributes or a graph where nodes are GO terms and edges are species belonging (default value="Categories").
- **numCores**: numeric, Number of cores to use for the process (default value numCores=2). For the example below, only one core will be used.
- **outdir**: This parameter will allow save the graph file in a folder described here (e.g; "D:"). This parameter only works when saveGraph=TRUE.

**Value**

This function will return a list with two slots: edges and nodes. Edges represent an edge list and their weights and nodes which represent the nodes and their respective weights (weights, shared).

**Examples**

```r
GOterm_field <- "Functional_Category"
data(comparison_ex_compress_CH)
#Defining the species names
species1 <- "H. sapiens"
species2 <- "A. thaliana"
x_graph <- graph_two_GOspecies(x=comparison_ex_compress_CH, 
species1=species1, 
species2=species2, 
GOterm_field=GOterm_field, 
numCores=1, 
saveGraph = FALSE, 
option= "Categories", 
outdir = NULL)
```

---

**H_sapiens**  
*H. sapiens* functional enrichment analysis of 5494 genes related to cancer-hallmarks

**Description**

This dataset is a subset of the original dataset obtained for Clavijo-Buriticá (In preparation).
Usage

H_sapiens

Format

A data frame with 5000 rows and 6 variables:

- **Enrichment_FDR** Numeric: False discovery rate values for the GO term
- **Genes_in_list** numeric: Number of genes in the list of genes for a given GO term
- **Total_genes** numeric: Number of genes in the genome of a species for a given GO term
- **Functional_Category** character: GO term name or GO term id
- **Genes** character: Genes found for a given GO term
- **feature** character: A column representing the belonging of a group of comparison

Source

https://data.mendeley.com/datasets/myy2wxd59/1

References

Clavijo-Buriticá, Sosa, C.C., Mosquera, A.J. Álvarez, A., Medina, J. Quimbaya, M.A. A systematic comparison of the molecular machinery associated with Cancer-Hallmarks between plants and humans reveals Arabidopsis thaliana as a useful model to understand specific carcinogenic events (to be submitted, Target journal: Plos Biology)
**Source**

https://data.mendeley.com/datasets/myyy2wxd59/1

**References**

Clavijo-Buriticá, Sosa, C.C., Mosquera, A.J. Álvarez, A., Medina, J. Quimbaya, M.A. A systematic comparison of the molecular machinery associated with Cancer-Hallmarks between plants and humans reveals Arabidopsis thaliana as a useful model to understand specific carcinogenic events (to be submitted, Target journal: Plos Biology)

---

**mostFrequentGOs**

**Most frequent GO terms among groups for a data.frame**

**Description**

Provides an easy way to get the frequency of GO terms such as biological processes for a data frame and a series of features

**Usage**

```r
mostFrequentGOs(df, GOterm_field)
```

**Arguments**

- `df`: A data frame with the results of a functional enrichment analysis for a species with an extra column "feature" with the features to be compared
- `GOterm_field`: This is a string with the column name of the GO terms (e.g: "Functional.Category")

**Value**

This function will return a table with the frequency of GO terms per feature

**Examples**

```r
#Loading example datasets
data(H_sapiens)
#Defining the column with the GO terms to be compared
GOterm_field <- "Functional_Category"
#Running function
x <- mostFrequentGOs(df=H_sapiens, GOterm_field=GOterm_field)
#Displaying results
head(x)
```
Index

* datasets
  A_thaliana, 2
  A_thaliana_compress, 3
  comparison_ex_compress, 5
  comparison_ex_compress_CH, 6
  H_sapiens, 12
  H_sapiens_compress, 13
* package
  GOCompare-package, 2

A_thaliana, 2
A_thaliana_compress, 3

compareGOspecies, 4
comparison_ex_compress, 5
comparison_ex_compress_CH, 6

evaluateCAT_species, 7
evaluateGO_species, 8

GOCompare (GOCompare-package), 2
GOCompare-package, 2
graph_two_GOspecies, 11
graphGOspecies, 9

H_sapiens, 12
H_sapiens_compress, 13

mostFrequentGOs, 14