# Package 'goSTAG'

October 17, 2017

**Type** Package

**Title** A tool to use GO Subtrees to Tag and Annotate Genes within a set

**Version** 1.0.1

**Date** 2017-06-02

**Author** Brian D. Bennett and Pierre R. Bushel

**Maintainer** Brian D. Bennett <brian.bennett@nih.gov>

**Description** Gene lists derived from the results of genomic analyses are rich in biological information. For instance, differentially expressed genes (DEGs) from a microarray or RNA-Seq analysis are related functionally in terms of their response to a treatment or condition. Gene lists can vary in size, up to several thousand genes, depending on the robustness of the perturbations or how widely different the conditions are biologically. Having a way to associate biological relatedness between hundreds and thousands of genes systematically is impractical by manually curating the annotation and function of each gene. Over-representation analysis (ORA) of genes was developed to identify biological themes. Given a Gene Ontology (GO) and an annotation of genes that indicate the categories each one fits into, significance of the over-representation of the genes within the ontological categories is determined by a Fisher's exact test or modeling according to a hypergeometric distribution. Comparing a small number of enriched biological categories for a few samples is manageable using Venn diagrams or other means for assessing overlaps. However, with hundreds of enriched categories and many samples, the comparisons are laborious. Furthermore, if there are enriched categories that are shared between samples, trying to represent a common theme across them is highly subjective. goSTAG uses GO subtrees to tag and annotate genes within a set. goSTAG visualizes the similarities between the over-representation of DEGs by clustering the p-values from the enrichment statistical tests and labels clusters with the GO term that has the most paths to the root within the subtree generated from all the GO terms in the cluster.

**License** GPL-3

**Depends** R (>= 3.4)

**Imports** AnnotationDbi, biomaRt, GO.db, graphics, memoise, stats, utils

**Suggests** BiocStyle, knitr, rmarkdown, testthat

**VignetteBuilder** knitr

**LazyData** true

**biocViews** GeneExpression, DifferentialExpression, GeneSetEnrichment, Clustering, Microarray, mRNAMicroarray, RNASeq, Visualization, GO

**NeedsCompilation** no

# R topics documented:



---

goSTAG-package *goSTAG Package*

---

## Description

A tool to use GO Subtrees to Tag and Annotate Genes within a set.

## Details

| | |
|---|---|
| Package: | goSTAG |
| Type: | Package |

## Author(s)

Brian D. Bennett
Pierre R. Bushel

## References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

## Examples

```
data( goSTAG_example_gene_lists )
go_terms <- loadGOTerms()
enrichment_matrix <- performGOEnrichment( goSTAG_example_gene_lists, go_terms )
hclust_results <- performHierarchicalClustering( enrichment_matrix )
clusters <- groupClusters( hclust_results )
cluster_labels <- annotateClusters( clusters )
plotHeatmap( enrichment_matrix, hclust_results, clusters, cluster_labels )
```

annotateClusters *Annotate Clusters*

## Description

Annotates each cluster of GO terms with the term containing the most paths to the root.

## Usage

```
annotateClusters(clusters)
```

## Arguments

| | |
|---|---|
| clusters | A list of vectors. Each element of the list corresponds to a cluster, and each vector contains the GO terms in the cluster. |

## Details

All of the GO terms within all of the clusters must belong to the same GO domain (either BP, CC, or MF).

## Value

A vector of cluster labels.

## Author(s)

Brian D. Bennett
Pierre R. Bushel

## References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

## Examples

```
data( goSTAG_example_gene_lists )
go_terms <- loadGOTerms()
enrichment_matrix <- performGOEnrichment( goSTAG_example_gene_lists, go_terms )
hclust_results <- performHierarchicalClustering( enrichment_matrix )
clusters <- groupClusters( hclust_results )

cluster_labels <- annotateClusters( clusters )
head( cluster_labels )
```

goSTAG_example_gene_lists *goSTAG Example Gene Lists*

## Description

An example set of gene lists.

## Usage

```
data(goSTAG_example_gene_lists)
```

## Value

A list of vectors. Each element of the list corresponds to a gene list, and each vector contains the genes in the gene list.

## Author(s)

Brian D. Bennett
Pierre R. Bushel

## References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

## Examples

```
data( goSTAG_example_gene_lists )
lapply( head( goSTAG_example_gene_lists ), head )
```

goSTAG_go_genes_human *goSTAG GO Genes (Human)*

## Description

A list of GO terms and the human genes associated with them.

## Usage

```
data(goSTAG_go_genes_human)
```

## Details

This data is meant for internal goSTAG use.

## Value

A list of vectors. Each element of the list corresponds to a GO term, and each vector contains the genes associated with the GO term. The list also has an additional element named "ALL", which is a vector that contains all annotated genes.

### Author(s)

Brian D. Bennett
Pierre R. Bushel

### References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

### Examples

```
data( goSTAG_go_genes_human )
lapply( head( goSTAG_go_genes_human ), head )
```

---

## goSTAG_go_genes_mouse *goSTAG GO Genes (Mouse)*

---

### Description

A list of GO terms and the mouse genes associated with them.

### Usage

```
data(goSTAG_go_genes_mouse)
```

### Details

This data is meant for internal goSTAG use.

### Value

A list of vectors. Each element of the list corresponds to a GO term, and each vector contains the genes associated with the GO term. The list also has an additional element named "ALL", which is a vector that contains all annotated genes.

### Author(s)

Brian D. Bennett
Pierre R. Bushel

### References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

### Examples

```
data( goSTAG_go_genes_mouse )
lapply( head( goSTAG_go_genes_mouse ), head )
```

goSTAG_go_genes_rat *goSTAG GO Genes (Rat)*

## Description

A list of GO terms and the rat genes associated with them.

## Usage

```
data(goSTAG_go_genes_rat)
```

## Details

This data is meant for internal goSTAG use.

## Value

A list of vectors. Each element of the list corresponds to a GO term, and each vector contains the genes associated with the GO term. The list also has an additional element named "ALL", which is a vector that contains all annotated genes.

## Author(s)

Brian D. Bennett
Pierre R. Bushel

## References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

## Examples

```
data( goSTAG_go_genes_rat )
lapply( head( goSTAG_go_genes_rat ), head )
```

---

groupClusters *Group Clusters*

## Description

Groups similar leaves of a hierarchical cluster analysis into clusters.

## Usage

```
groupClusters(hclust_results, distance_threshold = 0.2)
```

### Arguments

| | |
|---|---|
| hclust_results | An object of class "hclust" containing a tree produced by hierarchical clustering. |
| distance_threshold | The distance threshold at which to group leaves into clusters. Leaves whose distance is less than or equal to this threshold will be grouped together. A larger distance threshold will produce fewer clusters with more members, whereas a smaller one will produce more clusters with fewer members. |

### Value

A list of vectors. Each element of the list corresponds to a cluster, and each vector contains the GO terms in the cluster.

### Author(s)

Brian D. Bennett
Pierre R. Bushel

### References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

### Examples

```
data( goSTAG_example_gene_lists )
go_terms <- loadGOTerms()
enrichment_matrix <- performGOEnrichment( goSTAG_example_gene_lists, go_terms )
hclust_results <- performHierarchicalClustering( enrichment_matrix )

clusters <- groupClusters( hclust_results )
lapply( head( clusters ), head )
```

---

## loadGeneLists *Load Gene Lists*

---

### Description

Loads gene lists from a file or directory.

### Usage

```
loadGeneLists(location, type = "GMT", sep = "\t", header = FALSE, col = 1)
```

### Arguments

| | |
|---|---|
| location | The location of the GMT file or the directory containing the gene lists. |
| type | A value indicating whether to load the gene lists from a single GMT file or a directory containing a separate file for each gene list. Acceptable options are "GMT" or "DIR". |
| sep | The field separator character by which values on each line are separated. |

| | |
|---|---|
| header | A logical value indicating whether the gene list files contain a header that should be ignored. Only applicable when type = "DIR". |
| col | The column in the gene lists files containing the genes. Only applicable when type = "DIR". |

## Value

A list of vectors. Each element of the list corresponds to a gene list, and each vector contains the genes in the gene list.

## Author(s)

Brian D. Bennett
Pierre R. Bushel

## References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

## Examples

```
tf <- tempfile()
writeLines( c(  "Gene_List_1\tNA\tGene1\tGene4",
                "Gene_List_2\tNA\tGene2\tGene7\tGene5",
                "Gene_List_3\tNA\tGene4\tGene8" ), tf )
gene_lists <- loadGeneLists( tf )
lapply( head( gene_lists ), head )

td <- tempdir()
unlink( paste( sep="/", td, list.files(td) ) )
writeLines( c( "Gene1", "Gene4" ), paste(sep="/",td,"Gene_List_1.txt") )
writeLines( c( "Gene2", "Gene7", "Gene5" ), paste(sep="/",td,"Gene_List_2.txt") )
writeLines( c( "Gene4", "Gene8" ), paste(sep="/",td,"Gene_List_3.txt") )
gene_lists <- loadGeneLists( td, type = "dir" )
lapply( head( gene_lists ), head )
```

---

# loadGOTerms *Load GO Terms*

---

## Description

Loads a list of GO terms and the genes associated with them.

## Usage

```
loadGOTerms(species = "human", domain = "BP", min_num_genes = 5, use_archived = TRUE)
```

### Arguments

| | |
|---|---|
| species | The species to use for associating genes with GO terms. Available options are "human", "mouse", or "rat". |
| domain | The GO domain to use. Acceptable options are "BP" (biological process), "CC" (cellular component), or "MF" (molecular function). |
| min_num_genes | The minimum number of genes required to be associated with each GO term. Any GO terms with fewer genes will be excluded. |
| use_archived | A logical value indicating whether to use an archived version of the gene associations. If FALSE, this function will use BioMart to generate the latest version of the gene associations, but may take several minutes. If TRUE (default), this function will use a previously generated version of the gene associations, which takes very little time. |

### Value

A list of vectors. Each element of the list corresponds to a GO term, and each vector contains the genes associated with the GO term. The list also has an additional element named "ALL", which is a vector that contains all annotated genes.

### Author(s)

Brian D. Bennett
Pierre R. Bushel

### References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

### Examples

```
go_terms <- loadGOTerms()
lapply( head( go_terms ), head )
```

---

## performGOEnrichment *Perform GO Enrichment*

---

### Description

Performs a GO enrichment analysis on a set of gene lists and GO terms. Will filter out GO terms having no significant enrichment in any of the gene lists.

### Usage

```
performGOEnrichment(gene_lists, go_terms, filter_method = "pval", significance_threshold = 0.05, p
```

### Arguments

| | |
|---|---|
| gene_lists | A list of vectors. Each element of the list corresponds to a gene list, and each vector contains the genes in the gene list. |
| go_terms | A list of vectors. Each element of the list corresponds to a GO term, and each vector contains the genes associated with the GO term. The list must also have an additional element named "ALL", which is a vector that contains all annotated genes. |
| filter_method | A value indicating whether to filter out GO terms based on a nominal p-value or an adjusted p-value. Acceptable options are "pval" (nominal p-value) or "p.adjust" (adjusted p-value). |
| significance_threshold | The significance threshold at which to filter out GO terms. GO terms with a nominal or adjusted p-value (indicated by "filter_method") of greater than or equal to this threshold in all gene lists will be filtered out. |
| p.adjust_method | The correction method used to adjust the p-values. Available options are those of the "method" argument for the "p.adjust" function, which are currently "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", or "none". Only applicable when filter_method = "p.adjust". |

### Value

A matrix of enrichment scores. Rows correspond to GO terms and columns correspond to gene lists.

### Author(s)

Brian D. Bennett
Pierre R. Bushel

### References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

### Examples

```
data( goSTAG_example_gene_lists )
go_terms <- loadGOTerms()

enrichment_matrix <- performGOEnrichment( goSTAG_example_gene_lists, go_terms )
head( enrichment_matrix )
```

---

## performHierarchicalClustering

*Perform Hierarchical Clustering*

---

### Description

Performs a hierarchical clustering analysis on a GO enrichment matrix.

## Usage

```
performHierarchicalClustering(enrichment_matrix, feature = "row", distance_method = "correlation"
```

## Arguments

| | |
|---|---|
| enrichment_matrix | A matrix of enrichment scores. Rows correspond to GO terms and columns correspond to gene lists. |
| feature | A value indicating whether to cluster the rows or the columns. Acceptable options are "row" or "col". |
| distance_method | The distance measure to use when generating the distance matrix. If "correlation" (default), this function will use one minus the absolute value of the correlation to measure distance. Otherwise, this function will use the "dist" function to measure distance. Available options are those of the "method" argument for the "dist" function, which are currently "euclidean", "maximum", "manhattan", "canberra", "binary", or "minkowski". |
| clustering_method | The agglomeration method to use when performing the hierarchical clustering. Available options are those of the "method" argument for the "hclust" function, which are currently "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median", or "centroid". |

## Value

An object of class "hclust" containing a tree produced by hierarchical clustering.

## Author(s)

Brian D. Bennett
Pierre R. Bushel

## References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

## Examples

```
data( goSTAG_example_gene_lists )
go_terms <- loadGOTerms()
enrichment_matrix <- performGOEnrichment( goSTAG_example_gene_lists, go_terms )

hclust_results <- performHierarchicalClustering( enrichment_matrix )
sample_hclust_results <- performHierarchicalClustering( enrichment_matrix, feature = "col" )
```

plotHeatmap *Plot Heatmap*

## Description

Plots a heatmap of goSTAG analysis results.

## Usage

```
plotHeatmap(enrichment_matrix, hclust_results, clusters, cluster_labels, sample_hclust_results =
```

## Arguments

| | |
|---|---|
| enrichment_matrix | A matrix of enrichment scores. Rows correspond to GO terms and columns correspond to gene lists. |
| hclust_results | An object of class "hclust" containing a tree produced by hierarchical clustering. This tree is for the GO terms (rows) in the enrichment matrix. |
| clusters | A list of vectors. Each element of the list corresponds to a cluster, and each vector contains the GO terms in the cluster. |
| cluster_labels | A vector of cluster labels. |
| sample_hclust_results | An object of class "hclust" containing a tree produced by hierarchical clustering. This tree is for the samples (columns) in the enrichment matrix. If NULL (default), the samples will appear in their original order with no dendrogram. |
| min_num_terms | The minimum number of GO terms required to plot a cluster label. Any cluster with fewer GO terms will not be labeled. |
| maximum_heatmap_value | The maximum value in the heatmap. Any values in the enrichment matrix greater than this value will be made equal to this value. This is to increase contrast in the heatmap by controlling outliers. |
| heatmap_colors | The color range for the heatmap. Available options are "auto", "extra", or a vector containing the color range. If "auto" (default), this function will use a color range from grey to red. If "extra", this function will use a color range from grey to yellow to red. |
| sample_colors | A vector of colors for the sample labels. If NULL (default), the sample labels will not have any color. |
| margin_size | The size of the margin between the elements of the plot, as a percentage of the entire width of the plot. |
| dendrogram_width | The size of the GO term dendrogram (including the cluster labels), as a percentage of the entire width of the plot (excluding the margins). |
| cluster_label_width | The size of the cluster labels, as a percentage of the GO term dendrogram. |
| header_height | The size of the header (including the sample labels), as a percentage of the entire height of the plot (excluding the margins). |
| sample_label_height | The size of the sample labels, as a percentage of the header. |

dendrogram_lwd The width of the GO term dendrogram lines.

header_lwd The width of the sample dendrogram lines.

cluster_label_cex
A value that scales the cluster label text size.

sample_label_cex
A value that scales the sample label text size.

## Value

None, the function is invoked for its side effect.

## Author(s)

Brian D. Bennett
Pierre R. Bushel

## References

Bennett BD and Bushel PR. goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set. Source Code Biol Med. 2017 Apr 13.

## Examples

```
data( goSTAG_example_gene_lists )
go_terms <- loadGOTerms()
enrichment_matrix <- performGOEnrichment( goSTAG_example_gene_lists, go_terms )
hclust_results <- performHierarchicalClustering( enrichment_matrix )
clusters <- groupClusters( hclust_results )
cluster_labels <- annotateClusters( clusters )

plotHeatmap( enrichment_matrix, hclust_results, clusters, cluster_labels )
```

---

# rat_cancer_therapeutics_gene_lists
*Rat Cancer Therapeutics Gene Lists*

---

## Description

Differentially expressed genes from gene expression analysis (Affymetrix GeneChip Rat Genome 230 2.0 arrays) of samples acquired from the bone marrow of rats exposed to cancer therapeutic drugs (topotecan in combination with oxaliplatin) for 1, 6, or 24 hrs. Comparisons were treated samples to time-matched controls using limma with an FDR < 0.05 and absolute fold change > 2.0. Details of the analysis are as previously described (Davis et al., 2015). The raw data are available in the Gene Expression Omnibus under accession number GSE63902.

## Usage

```
data(rat_cancer_therapeutics_gene_lists)
```

## Value

A list of vectors. Each element of the list corresponds to a gene list, and each vector contains the genes in the gene list.

## References

Davis M, Li J, Knight E, Eldridge SR, Daniels KK, Bushel PR. Toxicogenomics profiling of bone marrow from rats treated with topotecan in combination with oxaliplatin: a mechanistic strategy to inform combination toxicity. Front Genet. 2015 Feb 12;6:14.

## Examples

```
data( rat_cancer_therapeutics_gene_lists )
lapply( head( rat_cancer_therapeutics_gene_lists ), head )
```

# Index