Package 'DOSim'

June 21, 2010

Type Package

Title Computation of functional similarities between DO terms and gene products; DO enrichment analysis;Disease Ontology annotation

Version 1.1

Date 2010-05-18

Author Jiang Li <riverlee2008@gmail.com>

Maintainer Jiang Li <riverlee2008@gmail.com>

Depends RBGL, graph, stats

Description This package implements several functions useful for computing similarities between DO terms and gene products based on their DO annotation. Moreover it allows for computing a DO enrichment analysis and provide basic disease ontology annotation.

License GPL (>= 2)

URL http://bioinfo.hrbmu.edu.cn/dosim/

LazyLoad yes

Repository CRAN

R topics documented:

`DOSim-package` *DOSim package*

Description

This package implements several functions useful for computing similarities between DO terms and gene products based on their DO annotation. Moreover it allows for computing a DO enrichment analysis and provide basic disease ontology annotation.

Details

Package:	DOSim
Type:	Package
Version:	1.0
Date:	2010-04-05
License:	GPL (>= 2)
LazyLoad:	yes

Author(s)

Jiang Li

Maintainer: Jiang Li <<riverlee2008@gmail.com>>

Examples

```
##################################
#example
terms<-c("DOID:1579","DOID:945")
tsim<-getTermSim(terms)

print(tsim)
```

DOEnrichment *DO enrichment analysis*

Description

This function performs a DO enrichment analysis using Hypergeometric Test or Fisher's Exact Test.

Usage

```
DOEnrichment(genelist, method = "hypertest", filter = 5, cutoff = 0.05)
```

Arguments

genelist	character vector of Entrez Gene IDs
method	one of ("hypertest","fisher")
filter	indicates that DO terms must have at least 'filter' genes annotated
cutoff	significant cutoff for DO enrichment analysis

Details

Currently the following methods for DO enrichment are implemented:

"hypertest" Using Hypergeometric Test

"fisher" Using Fisher's Exact Test

Value

Return a data.frame object with 6 columns.Details are below:

"DOID" enriched DO ID name

"pvalue" correspdonding pvalue of enriched DO term

"odds" Calculated by $\frac{m/n}{M/N}$ where 'm' stands for the gene number covered by DO in the list,'n' for inputed gene list number,'M' for gene number covered by DO among whole human genes,'N' stands for the gene number of whole human beings.

"genenum1" Gene number annotated to this DO term among whole human genes

"genenum2" Gene number annotated to this DO term in the inputed gene list

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

Examples

```
##############################
#Examples

  genelist=as.character(1:100)
  res<-DOEnrichment(genelist,filter=50)
  print(res)
```

DOSimEnv *Disease Ontology enviroment object*

Description

Disease Ontology enviroment object used for other functions

Usage

```
data(DOSimEnv)
```

Format

The format is: <environment: 0xbe1834c>

Source

The original data is came from John D.Osborne's work.URL:http://projects.bioinformatics.northwestern.edu/do_rif/

Examples

```
data(DOSimEnv)
## maybe str(DOSimEnv) ; plot(DOSimEnv) ...
ls(DOSimEnv)
```

---

filterDO *Filter DO*

Description

Filter out genes from a list not mapping to the disease ontology.

Usage

```
filterDO(genelist)
```

Arguments

genelist	character vector of Entrez gene IDs

Details

Filter out genes from a list not mapping to the disease ontology,and return a list which the genes are in the disease ontology.

Value

List with items

"genename"	gene ID
"annotation"	character vector of DO IDs mapping to the gene

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

Examples

```
################################
#Example
genelist<-1:10
res<-filterDO(genelist)
print(res)
```

---

getAncestors *Get a list of all ancestors associated to each DO term*

Description

Returns the list of all ancestors associated to each DO term.

Usage

```
getAncestors(dolist, verbose = TRUE)
```

Arguments

dolist	character vector of DO IDs
verbose	print out some information

Value

List with entry names for each DO ID. Each entry contains a charcter vector with the ancestor DO IDs

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

See Also

getOffsprings, getChildren, getParents

Examples

```
################################
#Example

terms<-c("DOID:934","DOID:1579")
res<-getAncestors(terms)
print(res)
```

---

getChildren *Get a list of all direct children of each DO term*

Description

Returns the list of all direct children associated to each DO term.

Usage

```
getChildren(dolist, verbose = TRUE)
```

Arguments

dolist	character vector of DO IDs
verbose	print out some information

Value

List with entry names for each DO ID. Each entry contains a character vector with the direct children of DO IDs.

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

See Also

getOffsprings, getParents, getAncestors

Examples

```
################################
#Example

terms<-c("DOID:934","DOID:1579")
res<-getChildren(terms)
print(res)
```

`getDisjCommAnc` *Get disjoint common ancestors.*

Description

Returns the DO terms representing the disjoint common ancestors of two DO terms.

Usage

```
getDisjCommAnc(term1, term2)
```

Arguments

`term1`	DO term 1
`term2`	DO term 2

Value

Character vector of DO terms

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

References

Couto, F.; Silva, M. & Coutinho, P., Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors, Conference in Information and Knowledge Management, 2005

Examples

```
if(require(graph))
getDisjCommAnc("DOID:934","DOID:95")
```

getDoAnno *Get gene list associated to each DO term*

Description

Get gene list associated to each DO term

Usage

```
getDoAnno(dolist)
```

Arguments

dolist	character vector of DO IDs

Value

List with entry names for each DO ID. Each entry contains a character vector with associated Entrez Gene IDs.

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

See Also

getDoTerm

Examples

```
################################
#Example

terms<-c("DOID:934","DOID:1579")
res<-getDoAnno(terms)
print(res)
```

getDOGraph *Get DO graph with specified DO terms at its leave.*

Description

The function getDOGraph returns a graphNEL object representing the DO graph with leaves specified in the argument.

Usage

```
getDOGraph(term, prune = Inf)
```

Arguments

term	character vector of DO term IDs
prune	do not show the complete graph, but prune it after the specified number of ancestors

Value

graphNEL object(s)

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

Examples

```
if(require(graph)){
g<-getDOGraph(c("DOID:95","DOID:8"))
if(require(Rgraphviz)){
plot(g)
}
}
```

getDoTerm *Get DO term's name*

Description

Returns the list of DO term's name associated to each DO ID.

Usage

```
getDoTerm(dolist)
```

Arguments

dolist	character vector of DO IDs

Value

List with entry names for each DO ID. Each entry contains a character represents DOID's term name.

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

See Also

getDoAnno

Examples

```
################################
#Example

terms<-c("DOID:934","DOID:1579")
res<-getDoTerm(terms)
print(res)
```

---

getGeneSim *Compute functional similarity for genes*

Description

Calculate the pairwise functional similarities for a list of genes using different strategies

Usage

```
getGeneSim(genelist, similarity = "funSimMax", similarityTerm = "relevance", normal
```

Arguments

genelist	character vector of Entrez gene IDs
similarity	method to calculate the functional similarity between gene products
similarityTerm	method to compute the similarity of DO terms
normalization	normalize similarities yes/no

method	"sqrt": normalize sim(x,y) <- sim(x,y)/sqrt(sim(x,x)*sim(y,y)); "Lin": normalize sim(x,y) <- 2*sim(x,y)/(sim(x,x) + sim(y,y)); "Tanimoto": normalize sim(x,y) <- sim(x,y)/(sim(x,x) + sim(y,y) - sim(x,y)). NOTE: normalization does not have any effect, if similarity = "funSimMax", "funSimAvg" or similarity = "OA" and avg=TRUE
avg	standardize the OA kernel by the maximum number of DO terms for both genes
verbose	print out some information

Details

The method to calculate the pairwise functional similarity between gene products can either be:

"max" the maximum similarity between any two DO terms

"mean" the average similarity between any two DO terms

funSimMax the average of best matching DO term similarities. Take the maximum of the scores achieved by assignments of DO terms from gene 1 to gene 2 and vice versa. [2]

funSimAvg the average of best matching DO term similarities. Take the average of the scores achieved by assignments of DO terms from gene 1 to gene 2 and vice versa. [2]

"OA" the optimal assignment (maximally weighted bipartite matching) of DO terms associated to the gene having fewer annotation to the DO terms of the other gene. [1]

"hausdorff" the translation of the Hausdorff distance between two sets: Let A and B be the two sets of DO terms associated to two genes. Then $sim(A, B) = min(min(max_{x \in A}(x, y)), min(max_{y \in B}(x, y)))$. [3]

"dot" the dot product between feature vectors describing the absence/presence of each DO term. The absence of each DO term is weighted by its information content. Depending on the type of later normalization one can arrive at the cosine similarity (method="sqrt") or at the Tanimoto coefficient (method="Tanimoto").[4]

"Wang" Assume gene1 have m DO annoated(DO1={do1,do2,...,dom}) and gene2 have n DO annotated(DO2={do1,do2,...,don}).Define Sum1 is the sum of maximum of the scores achieved by assignments of each DO in DO1 to DO2,same for Sum2, and the Sim(g1,g2)=(Sum1+Sum2)/(m+n) [5]

Value

n x n similarity matrix (n = number of genes)

References

[1] H. Froehlich, N. Speer, C. Spieth, A. Zell, Kernel Based Functional Gene Grouping, Proc. Int. Joint Conf. on Neural Networks (IJCNN), 6886 - 6891, 2006.

[2] A. Schlicker, F. Domingues, J. Rahnenfuehrer, T. Lengauer, A new measure for functional similarity of gene products based on Gene Ontology, BMC Bioinformatics, 7, 302, 2006.

[3] A. del Pozo, F. Pazos, A. Valencia, Defining functional distances over Gene Ontology, BMC Bioinformatics, 9:50, 2008.

[4] M. Mistry, P Pavlidis, Gene Ontology term overlap as a measure of gene functional similarity, BMC Bioinformatics, 9:327, 2008.

[5] James Z.Wang,Zhidian Du, et al. A new method to measure the semantic similarity of GO terms.Bioinformatics 2007,Vol 23,1274-1281.

See Also

getTermSim

Examples

```
####################################
#Example
genelist=1:10
gsim<-getGeneSim(genelist)
print(gsim)
```

`getMinimumSubsumer` *Compute minimum subsumer of two DO terms*

Description

Returns the minimum subsumer(i.e. the common ancestor having the maximal information content) of two DO terms

Usage

```
getMinimumSubsumer(term1, term2)
```

Arguments

`term1`	DO term 1
`term2`	DO term 2

Details

The result is computed base on current disease ontology

Value

DO term representing the minmum subsumber. If there is no minumum subsumer,the result is the string "NA".

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

References

P. Resnik, Using Information Content to evaluate semantic similarity in a taxonomy, Proc. 14th Int. Conf. Artificial Intel., 1995

Examples

```
############################
#Example
term1="DOID:8"
term2="DOID:95"
getMinimumSubsumer(term1,term2)
```

`getOffsprings` *Get all offspring associated with each DO term*

Description

Returns the list of all offspring associated to each DO term.

Usage

```
getOffsprings(dolist, verbose = TRUE)
```

Arguments

`dolist`	character vector of DO IDs
`verbose`	print out some information

Value

List with entry names for each DO ID. Each entry contains a character vector with the offspring DO IDs.

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

See Also

`getChildren`, `getParents`, `getAncestors`

Examples

```
################################
#Example

terms<-c("DOID:934","DOID:1579")
res<-getOffsprings(terms)
print(res)
```

`getParents` *Get direct parents for each DO term*

Description

Returns a list of all direct parents associated to each DO term.

Usage

```
getParents(dolist, verbose = TRUE)
```

Arguments

`dolist`	character vector of DO IDs
`verbose`	print out some information

Value

List with entry names for each DO ID. Each entry contains a character vector with the direct parent of DO IDs.

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

See Also

`getOffsprings`, `getChildren`, `getAncestors`

Examples

```
################################
#Example

terms<-c("DOID:934","DOID:1579")
res<-getParents(terms)
print(res)
```

getShortestPath *Get the shortest path between two terms*

Description

Get the shortest path between two terms.

Usage

```
getShortestPath(term1, term2)
```

Arguments

term1	DO term 1
term2	DO term 2

Value

return the shortest path between two terms, if two term are not connect, the return value is Inf

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

Examples

```
#####################################
#example
term1="DOID:8"
term2="DOID:5"
getShortestPath(term1,term2)
#return 1
```

getTermSim *Get pairwise DO term similarities.*

Description

Returns the pairwise similarities between DO terms based on different methods.

Usage

```
getTermSim(termlist, method = "relevance", verbose = TRUE)
```

Arguments

`termlist`	character vector of DO terms
`method`	one ot the supported methods for DO term similairty(see below)
`verbose`	print out various information or not

Details

Currently the following methods for computing DO term similarities are implemented:

"Resnik" information content of minimum subsumer (ICms) [1]

"JiangConrath" $1 - \min(1, IC(term1) - 2ICms + IC(term2))$[2]

"Lin" $\frac{2ICms}{(IC(term1)+IC(term2))}$ [3]

"CoutoEnriched" FuSSiMeg enriched term similarity by Couto et al. [4].

"CoutoResnik" average information content of common disjunctive ancestors of `term1` and `term2` (ICshare) [5]

"CoutoJiangConrath" $1 - \min(1, IC(term1) - 2ICshare + IC(term2))$ [5]

"CoutoLin" $\frac{2ICshare}{(IC(term1)+IC(term2))}$ [5]

"relevance" sim_Lin * (1 - exp(-ICms)) [6]

"GIC" summed information content of common ancestors divided by summed information content of all ancestors of term1 and term2 [7]

"simIC" sim_Lin * (1 - 1/(1 + ICms)) [8]

"path" $\frac{1}{p}$ where p is the number of nodes on the shortest path [9][11]

"lch" $-\log(\frac{p}{2.depth})$ where depth is maximum depth of the hierachy [10][11]

"Wang" $Sim(term1, term2) = \frac{\sum_{t \in T_{term1} \cap T_{term2}} (S_{term1}(t) + S_{term2}(t))}{SV(term1)+SV(term2)}$ [12]

Value

n x n matrix (n = number of DO terms) with similarities between DO terms.

Note

All calculations use normalized information contents for each DO term. Normalization is achieved by dividing each information content by the maximum information content.

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

References

[1] P. Resnik, Using Information Content to evaluate semantic similarity in a taxonomy, Proc. 14th Int. Conf. Artificial Intel., 1995

[2] J. Jiang, D. Conrath, Semantic Similarity based on Corpus Statistics and Lexical Taxonomy, Proc. Int. Conf. Research in Comp. Ling., 1998

[3] D. Lin, An Information-Theoretic Definition of Similarity, Proc. 15th Int. Conf. Machine Learning, 1998

[4] F. Couto, M. Silva, P. Coutinho, Implementation of a Functional Semantic Similarity Measure between Gene-Products, DI/FCUL TR 03-29, Department of Informatics, University of Lisbon, 2003

[5] Couto, F.; Silva, M. & Coutinho, P., Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors, Conference in Information and Knowledge Management, 2005

[6] A. Schlicker, F. Domingues, J. Rahnenfuehrer, T. Lengauer, A new measure for functional similarity of gene products based on Gene Ontology, BMC Bioinformatics, 7, 302, 2006.

[7] C. Pesquita, D. Faria, H. Bastos, A. Falcao, F. Couto, Evaluating GO-based Semantic Similarity Measures, In: Proc. 10th Annual Bio-Ontologies Meeting 2007, 37 - 40, 2007

[8] B. Li, J. Wang, A. Feltus, J. Zhou, F. Luo, Effectively Integrating Information Content and Structural Relationship to Improve the GO-based Similarity Measure Between Proteins, BMC Bioinformatics, 2009.

[9] Wu Z, Palmer M. Verb semantics and lexical selection.In:Proceedings of the 32nd annual meeting of the association for computational linguistics. Las Cruces, NM; 1994.p.133-8.

[10] Leacock C,Chodorow M. Combining local context and WordNet similarity for word sense identification. In: Fellbaum C, editor. WordNet: An electronic lexical database. Cambridge, MA: MIT Press;1998.p.265-83.

[11] Pedersen, T., S. V. S. Pakhomov, et al. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 2007 40(3): 288-299.

[12] James Z.Wang,Zhidian Du, et al. A new method to measure the semantic similarity of GO terms.Bioinformatics 2007,Vol 23,1274-1281.

See Also

`getMinimumSubsumer`, `getDisjCommAnc`

Examples

```
################################
#Example
getTermSim(c("DOID:8","DO:1117"),method="Lin")
```

internal *internal functions*

Description

internal functions: do not call these functions directly.

Usage

```
various
```

Arguments

various

Value

various

Author(s)

Jiang Li </emailriverlee2008@gmail.com>

plotCluster *Plot a cluster Dendrogram*

Description

Plot a cluster Dendrogram

Usage

```
plotCluster(hier, h = 0.9, minsize = 5, main = "Cluster Dendrogram", ...)
```

Arguments

hier	an hclust object
h	height cut-off for the branches
minsize	cluster size cut-off
main	plot's name
...	other parameters

Value

show a picture with top part a cluster dendrogram and below is a barplot which the same color indicates a module defined by argument "h" and "minsize"

Author(s)

Jiang Li<<riverlee2008@gmail.com>>

Examples

```
require(graphics)
hc <- hclust(dist(USArrests), "ave")
plotCluster(hc,h=50,minsize=5,hang=-1)
```

Index