CD4Ids  
**Entrez identifiers for CD4-TCells example**

**Description**

This dataset contains the entrez identifiers CD4EntrezIds and their associated GO Terms CD4GOTermsFrame and CD4GOTermsList corresponding to the list of differentially expressed genes in a study by Henkel et al.

**Usage**

```r
data(CD4Ids)
```

**Source**


**Examples**

```r
data(CD4Ids)
```

**GOTermsList**  
**Functions to create and manage lists of GO terms associated with a vector of 'Entrez' identifiers**

**Description**

These functions prepare data to be processed by the 'basicProfile' function. To create a profile a set of GOterms belonging to one or more ontologies is needed. The terms belonging to each gene must be given separately so that they can be counted. This function queries the environment 'GOENTREZID2GO' with the vector of Entrez terms and formats the output into a list whose components -one per Entrez term- contain the most specific GO identifiers associated with this term.
Usage

GOTermsList(LLids, onto = "any", evid = "any", na.rm = FALSE, orgPkg )
getAncestorsList(GOtermslist, onto, unique.ancestor=TRUE, na.rm=TRUE, combine=TRUE)
getGOLevel(onto, level)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLids</td>
<td>Character vector of Entrez (formerly Locuslink identifiers)</td>
</tr>
<tr>
<td>onto</td>
<td>ontology to be queried using the genes list</td>
</tr>
<tr>
<td>evid</td>
<td>type of evidence supporting the selected GO Terms</td>
</tr>
<tr>
<td>na.rm</td>
<td>flag indicating if those ids returning NA must be removed from the output</td>
</tr>
<tr>
<td>orgPkg</td>
<td>Organism annotation package ('org.Xx.eg.db') required to obtain the GO terms associated with the Entrez identifiers</td>
</tr>
</tbody>
</table>

GOtermslist  List produced by a call to function GOTermsList
unique.ancestor Flag to remove repeated ancestor identifiers
combine Flag to combine ancestors
level GO level at which the profile is built

Details

During the call to this function there may appear two types of NAs.
By one side if a name is not mapped in LocusLink this yields an NA that must be eliminated because nothing can be found through LL about this name
By another side if a gene is identified in LL but yields NA it seems to mean that it is not mapped in the GO
This may be eliminated but it may be worth the pity to keep track of them and to put these terms in an 'Seemingly unmannotated' category. In the case that its number was very high it migt suggest reviewing the list or reconsidering the results.

Value

A list whose components -one per Entrez term- are character vectors with the most specific GO identifiers associated with this term

Author(s)

Alex Sanchez

See Also

getAncestorsList

Examples

#data(CD4Ids)
#simpleLLids<- as.character(c(2189,5575,5569,11)) #1 is not a Locuslink identifier
#simpleGOlist<- GOTermsList (simpleLLids, orgPkg="org.Hs.eg.db")
#print(simpleGOlist.CC<-GOTermsList (simpleLLids,"CC", orgPkg="org.Hs.eg.db"))
#print(simpleGOlist.IEA<-GOTermsList (simpleLLids,evid="IEA",na.rm=TRUE, orgPkg="org.Hs.e..."
**basicProfile**  
*Builds basic functional profile*

### Description

Compute basic functional profile for a given list of genes/GO identifiers, a given ontology at a given level of the GO

### Usage

```r
basicProfile(genelist, idType = "Entrez", onto = "ANY", level = 2, orgPackage=NULL, anotPackage=NULL, ord = TRUE, multilevels = NULL, empty.cats = TRUE, cat.names = TRUE, na.rm = FALSE)
```

### Arguments

- **genelist**: List of genes on which the profile has to be based
- **idType**: Type of identifiers for the genes. May be 'Entrez' (default), BioCprobes or GoTermsFrame (see details below).
- **onto**: Ontology on which the profile has to be built
- **level**: Level of the ontology at which the profile has to be built
- **orgPackage**: Name of a Bioconductor's organism annotations package ('org.Xx-eg-db'). This field must be provided if the gene list passed to the function is either a character vector of 'Entrez' (NCBI) identifiers or a character vector of probe names
- **anotPackage**: Name of Bioconductor’s microarray annotations package. This field must be provided if the gene list passed to the function is a character vector of probe names
- **ord**: Set to 'TRUE' if the profile has to appear ordered by the category names
- **multilevels**: If it is not NULL it must be a vector of GO categories that defines the level at where the profile is built
- **empty.cats**: Set to 'TRUE' if empty categories should appear in the profile
- **cat.names**: Set to 'TRUE' if the profile has to contain the names of categories
- **na.rm**: Set to 'TRUE' if NAs should be removed

### Details

The function admits three types of entries: Entrez ('Entrez'), Bioconductor probe set names ('BioCprobes') or a special type of data frames ('GOTermsFrames'). If the identifier type are 'BioCprobes' then an annotation package name must be provided too.

### Value

An object of class GOProfile (one or more data frames in a list named by the ontologies)

### Author(s)

Alex Sanchez
References

See Also
expandedProfile

Examples
data(CD4Ids)
CD4.MF.Profiles <- basicProfile(genelist=CD4LLids, onto='MF', level=2, orgPackage="org.Hs.eg.db")
print(CD4.MF.Profiles)

compSummary
This function returns a brief summary of the comparison between two (expanded) profiles.

Description
Function to return a brief summary of the comparison between two (expanded) profiles.

Usage
compSummary(l, decs = 6)

Arguments
l A list of comparison results as returned by a call to compareGenelists
decs Number of decimal places to use in the output

Value
A data frame with the summarized results of each comparison. The values contained are: Sqr.Eucl.Dist: The squared euclidean distance, Standard Err: The standard error estimate, pValue p value of the test, low conf.int:Lower value for the desired confidence interval, up conf.int:Upper value for the desired confidence interval.

Author(s)
Alex Sanchez

Examples
data(sampleProfiles)
comparedMF <- compareGOProfiles (pn=expandedWelsh01[['MF']],
qm = expandedSingh01[['MF']],
pqn0 = commonExpandedWelsh01Singh01[['MF']])
print(comparedMF)
print(compSummary(comparedMF))
**compareGOProfiles**  
*Comparison of lists of genes through their functional profiles*

**Description**

Compare two samples of genes in terms of their GO profiles \(pn\) and \(qm\). Both samples may share a common subsample of genes, with GO profile \(pqn0\). `compareGOProfiles` implements some inferential procedures based on asymptotic properties of the squared euclidean distance between the contracted versions of \(pn\) and \(qm\).

**Usage**

```r
compareGOProfiles(pn, qm = NULL, pqn0 = NULL, n = ngenes(pn), m = ngenes(qm), n0 = ngenes(pqn0), method = "lcombChisq", ab.approx = "asymptotic", confidence = 0.95, nsims = 10000, simplify = T, ...)
```

**Arguments**

- \(pn\): an object of class `ExpandedGOProfile` representing one or more "sample" expanded GO profiles for a fixed ontology (see the 'Details' section)
- \(qm\): an object of class `ExpandedGOProfile` representing one or more "sample" expanded GO profiles for a fixed ontology (see the 'Details' section)
- \(pqn0\): an object of class `ExpandedGOProfile` representing one or more "sample" expanded GO profiles for a fixed ontology (see the 'Details' section)
- \(n\): a numeric vector with the number of genes profiled in each column of \(pn\). This parameter is included to allow the possibility of exploring the consequences of varying sample sizes, other than the true sample size in \(pn\).
- \(m\): a numeric vector with the number of genes profiled in each column of \(qm\).
- \(n0\): a numeric vector with the number of genes profiled in each column of \(pqn0\).
- \(method\): the approximation method to the sampling distribution under the null hypothesis specifying that the samples \(pn\) and \(qm\) come from the same population. See the 'Details' section below
- \(confidence\): the confidence level of the confidence interval in the result
- \(ab.approx\): the approximation used for computing 'a' and 'b' coefficients (see details)
- \(nsims\): some inferential methods require a simulation step; the number of simulation replicates is specified with this parameter
- \(simplify\): should the result be simplified, if possible? See the 'Details' section
- \(...\): Other arguments needed

**Details**

An object of S3 class 'ExpandedGOProfile' is, essentially, a 'data.frame' object with each column representing the relative frequencies in all observed node combinations, resulting from profiling a set of genes, for a given and fixed ontology. The row.names attribute codifies the node combinations and each data.frame column (say, each profile) has an attribute, 'ngenes', indicating the number of profiled genes. The arguments 'pn', 'qm' and 'pqn0' are compared in a column by column wise, recycling columns, if necessary, in order to perform \(max(ncol(pn),ncol(qm),ncol(pqn0))\) comparisons (each comparison resulting in an object of class 'GOProfileHtest', an specialization of 'htest'). In order to be properly compared, these arguments are expanded by row, according to their row names.
That is, the data arguments can have unequal row numbers. Then, they are expanded adding rows with zero frequencies, in order to make them comparable.

In the i-th comparison (i from 1 to max(ncol(pn),ncol(qm),ncol(pqn0))), the parameters n, m and n0 are included to allow the possibility of exploring the consequences of varying sample sizes, other than the true sample sizes included as an attribute in pn, qm and pqn0.

When qm = NULL, the genes profiled in pn are compared with a subsample of them, those profiled in pqn0 (compare a set of genes with a restricted subset, e.g. those overexpressed under a disease). In this case we take qm=pqn0. When pqn0 = NULL, two profiles with no genes in common are compared.

Let Pn and Qm correspond to the contracted functional profiles (the total counts or relative frequencies of hits in each one of the s GO categories being compared) obtained from pn and qm. If P stands for the "population" profile originating the sample profile Pn[,j], Q for the profile originating Qm[,j] and d(,) for the squared euclidean distance, if P != Q, the distribution of sqrt(nm/(n+m))(d(Pn[,j],Qm[,j]) - d(P,Q))/se(d) is approximately standard normal, N(0,1). This provides the basis for the confidence interval in the result field icDistance. When P=Q, the asymptotic distribution of (nm/(n+m)) d(Pn[,j],Qm[,j]) corresponds to the distribution of a mixture of independent chi-square random variables, each one with one degree of freedom. The sampling distribution under H0 P=Q may be directly computed from this distribution (approximating it by simulation) (method="lcombChisq") or by a chi-square approximation to it, based on two correcting constants a and b (method="chi-square"). These constants are chosen to equate the first two moments of both distributions (the linear combination of chi-square random variables distribution and the approximating chi-square distribution). When method="chi-square", the returned test statistic value is the chi-square approximation (n d(Pn[,j],qm[,j]) - b) / a. Then, the result field 'parameter' is a vector containing the 'a' and 'b' values and the number of degrees of freedom, 'df'. Otherwise, the returned test statistic value is (nm/(n+m)) d(Pn[,j],Qm[,j]) and 'parameter' contains the coefficients of the linear combination of chi-squares.

Value

A list containing max(ncol(pn),ncol(qm),ncol(pqn0)) objects of class ‘GOPProfileHtest’, directly inheriting from 'htest' or a single ‘GOPProfileHtest’ object if max(ncol(pn),ncol(qm),ncol(pqn0))==1 and simplify == T. Each object of class 'GOPProfileHtest' has the following fields:

- **profilePn**: the first contracted profile to compute the squared Euclidean distance
- **profileQm**: the second contracted profile to compute the squared Euclidean distance
- **statistic**: test statistic; its meaning depends on the value of "method", see the 'Details' section.
- **parameter**: parameters of the sample distribution of the test statistic, see the 'Details' section.
- **p.value**: associated p-value to test the null hypothesis of profiles equality.
- **conf.int**: asymptotic confidence interval for the squared euclidean distance. Its attribute "conf.level" contains its nominal confidence level.
- **estimate**: squared euclidean distance between the contracted profiles. Its attribute "se" contains its standard error estimate.
- **method**: a character string indicating the method used to perform the test.
- **data.name**: a character string giving the names of the data.
- **alternative**: a character string describing the alternative hypothesis (always ‘true squared Euclidean distance between the contracted profiles is greater than zero’
compareGeneLists

Author(s)
Jordi Ocana

References

See Also
fitGOProfile, equivalentGOProfiles

Examples

data(sampleProfiles)
comparedMF <- compareGOProfiles (pn=expandedWelsh01['MF'],
                               qm = expandedSingh01['MF'],
                               pqn0= commonExpandedWelsh01Singh01['MF'])
print(comparedMF)
print(compSummary(comparedMF))

---

compareGeneLists  Compares two lists of genes by building (expanded) profiles and comparing them

Description
This function wraps all the needed steps to compare two lists of genes following the methodology developed by Sanchez, Salicru and Ocana (2007)

Usage

compareGeneLists(genelist1, genelist2, idType = "Entrez", onto = "ANY",
                 level = 2, orgPackage,
                 method = "lcombChisq", ab.approx = "asymptotic", confidence = 0.95, compareFunction = compareGOProfiles)

Arguments

genelist1   First gene set to be compared
genelist2   Second gene set to be compared
idType      Type of identifiers for the genes. May be 'Entrez' (default), BiocProbes or GoTermsFrame. See the 'Details' section below
onto        Ontology on which the profile has to be built
level       Level of the ontology at which the profile has to be built
orgPackage  Name of a Bioconductor's organism annotations package ('org.Xx-eg-db')
method      The approximation method to the sampling distribution under the null hypothesis specifying that the samples pn and qm come from the same population. See the 'Details' section below
compareProfilesLists

confi dence The confidence level of the confidence interval in the result
ab.approx The approximation used for computing ‘a’ and ‘b’ coefficients (see details)
compareFunction Allows to use ‘fitGOProfile’ (sample vs population) or ‘compareGOProfiles’ (sample1 vs sample2)

Value

The result of the comparison is a list with a variable number of arguments, depending for which ontologies has been performed the comparison. Each list member is an object of class ‘htest’ corresponding to the output of the function compareGOProfiles

Author(s)

Alex Sanchez

References


See Also

compareGOProfiles, basicProfile

Examples

data(prostateIds)
prostateCompared <- compareGeneLists (welsh01EntrezIDs[1:500],
singh01EntrezIDs[1:500], level=2, onto="MF", orgPackage="org.Hs.eg.db")
print(prostateCompared)
print(compSummary(prostateCompared))
compareProfilesLists

Arguments

expanded1  First expanded profile to be compared
expanded2  Second expanded profile to be compared
common.expanded  Expanded profile made from the genes appearing in both lists of genes
relationType  Type of relation between gene lists compared through the expanded profiles. It can be INCLUSION, INTERSECTION or DISJOINT
method  The approximation method to the sampling distribution under the null hypothesis specifying that the samples pn and qm come from the same population. See the 'Details' section below
confidence  The confidence level of the confidence interval in the result
ab.approx  The approximation used for computing 'a' and 'b' coefficients (see details)
...  Other arguments for the methods 'basicProfile' or 'compareGoProfiles'

Value

The result of the comparison is a list with a variable number of arguments, depending for which ontologies has been performed the comparison. Each list member is an object of class 'htest' corresponding to the output of the function compareGOProfiles

Author(s)

Alex Sanchez

References


See Also

compareGeneLists, expandedProfile

Examples

require(goProfiles)
data(sampleProfiles)
comparedMF<- compareProfilesLists (expandedWelsh01, expandedSingh01, relationType="DISJOINT")
print(comparedMF)
print(compSummary(comparedMF))
conversionFunctions

Functions to transform convert objects between different types

Description

These functions transform data from one class type into another, or pack simple processes such as compute the profiles needed for one annotations package.

Usage

\[
as.GOTerms.frame(myGOTermsList, na.rm=TRUE) \\
as.GOTerms.list(genelist, probeType, orgPackage=NULL, anotPkg=NULL, onto="any", \\
BioCpack2EntrezIDS(anotPkg, na.rm=FALSE) \\
BioCpack2Profiles(anotPkg, orgPackage, level=2, na.rm=FALSE, expanded=FALSE) \\
BioCprobes2Entrez(probeslist , anotPkg, na.rm=TRUE) \\
GOTermsFrame2GOTermsList(myGOTermsFrame, evid=FALSE)
\]

Arguments

- **myGOTermsList**
  - GOTermsList object to transform
- **myGOTermsFrame**
  - GOTermsFrame object to transform
- **genelist**
  - List of genes (Entrez Ids) to transform
- **evid**
  - Type of evidence supporting the selected GO Terms
- **na.rm**
  - Flag indicating if those ids returning NA must be removed from the output
- **probeType**
  - Type of probes to transform into Entrez Ids
- **probeslist**
  - List of probes to transform into Entrez Ids
- **orgPackage**
  - Name of the organism (‘org.Xx.eg.db’) annotation package
- **anotPkg**
  - Name of the chip annotation package
- **level**
  - GO level at which the profile is built
- **onto**
  - ontology
- **expanded**
  - Flag to decide if an expanded profile has to be computed

Details

Not yet available

Value

Every function returns a transformed object or a list of computed profiles

Author(s)

Alex Sanchez
Examples

data(CD4Ids)
myGOTermsList <- GOTermsList(CD4LLIds[1:5], orgPkg="org.Hs.eg.db")
myGOTermsFrame<- as.GOTerms.frame(myGOTermsList, na.rm=TRUE)
GOTermsFrame2GOTermsList(myGOTermsFrame, evid=FALSE)

drosophila

Entrez identifiers for genes related with an eye mutation in drosophila

Description

Entrez identifiers for genes related with an eye mutation in drosophila.

- ostrinIds: List of genes in Entrez, generated by Ostrin et al.
- michaudIds: List of genes in Entrez, generated by Michaud et al.
- drosophilaIds: List of Drosophila genes in Entrez.

Usage

data(drosophila)

Format

Each dataset is a character vector with a different number of elements which (should) correspond to valid Entrez identifiers

Examples

data(drosophila)

equivSummary

This function returns a brief summary of the equivalence test between two profiles.

Description

Function to return a brief summary of the equivalence test between two profiles. If in its current version it is better that equivalentGOProfiles is called with option simplify set to FALSE before equivSummary can be used.

Usage

equivSummary(l, decs = 6)

Arguments

- l: A list of comparison results as returned by a call to compareGenelists
- decs: Number of decimal places to use in the output
equivalentGOProfiles

Value

A data frame with the summarized results of each comparison. The values contained are:
- Sqr.Eucl.Dist: The squared euclidean distance.
- Standard Err: The standard error estimate.
- pValue: p value of the equivalence test.
- up conf.int: Upper value for the desired confidence interval.
- d0: Threshold value for equivalence test.
- Equivalent: Numerical value set to 1 if profiles can be considered equivalent and to zero if they cannot.

Author(s)

Alex Sanchez

Examples

data(sampleProfiles)
comparedMF <- compareGOProfiles (pn=expandedWelsh01['MF'],
qm = expandedSingh01['MF'],
pqn0= commonExpandedWelsh01Singh01['MF'])
equivMF<- equivalentGOProfiles(comparedMF)
print(equivSummary(equivMF, decs=5))

---
equivalentGOProfiles

Are two lists of genes equivalent in terms of their Gene Ontology profiles?

Description

Performs an equivalence test based on the squared Euclidean distance between the Gene Ontology profiles of two lists of genes. Equivalence is declared if the upper limit d.sup of a one-sided confidence interval [0, d.sup] for the distance is lesser than the equivalence limit d0.

Usage

equivalentGOProfiles(goObject, ...)
## S3 method for class 'GOProfileHtest':
equivalentGOProfiles(goObject, equivEpsilon = 0.05, d0 = NULL, confidence = NULL)
## S3 method for class 'list':
equivalentGOProfiles(goObject, ...)
## S3 method for class 'ExpandedGOProfile':
equivalentGOProfiles(goObject, qm=NULL, pqn0=NULL,
n = ngenes(goObject), m = ngenes(qm), n0 = ngenes(pqn0),
confidence = 0.95,
equivEpsilon = 0.05, d0 = NULL,
simplify = FALSE, ...)
## Default S3 method:
equivalentGOProfiles(goObject, ...)

Arguments

g0Object  an object related to GO profiles or comparisons between them
qm        an expanded GO profile, i.e. object of class 'ExpandedGOProfile'
pqn0      an expanded GO profile, i.e. object of class 'ExpandedGOProfile'
n         a numeric vector with the number of genes profiled in each column of g0Object.
          This parameter is included to allow the possibility of exploring the consequences of varying sample sizes, other than the true sample size in g0Object.
m         a numeric vector with the number of genes profiled in each column of qm.
n0        a numeric vector with the number of genes profiled in each column of pqn0.
confidence the nominal confidence level of the one-sided confidence interval on the distance
d0        a positive value specifying the equivalence limit
equivEpsilon a positive value used to compute 'd0' if it is not directly available
simplify  should the result be simplified, if possible? See the 'Details' section
...       further arguments, typically the same than to 'compareGOProfiles'

Details

An object of S3 class "ExpandedGOProfile" is, essentially, a "data.frame" object with each column representing the relative frequencies in all observed node combinations, resulting from profiling a set of genes, for a given and fixed ontology. The 'row.names' attribute codifies the node combinations and each "data.frame" column (say, each profile) has an attribute, 'ngenes', indicating the number of profiled genes.

In the 'ExpandedGOProfile' interface, the arguments 'g0Object', 'qm' and 'pqn0' are compared in a column by column wise, recycling columns, if necessary, in order to perform max(ncol(g0Object),ncol(qm),ncol(pqn0)) equivalence tests (each test resulting in an object of class 'htest'). In order to be properly tested, these arguments are expanded by row, according to their row names. That is, the data arguments can have unequal row numbers. Then, they are expanded adding rows with zero frequencies, in order to make them comparable. In the i-th comparison (i from 1 to max(ncol(g0Object),ncol(qm),ncol(pqn0))), the parameters n, m and n0 are included to allow the possibility of exploring the consequences of varying sample sizes, other than the true sample sizes included as an attribute in goObject, qm and pqn0. When qm = NULL, the genes profiled in g0Object are compared with a subsample of them, those profiled in pqn0 (is there equivalence between a set of genes and a restricted subset, e.g. those overexpressed under a disease, in terms of their profiles?). When pqn0 = NULL, an equivalence test between two profiles with no genes in common is performed.

In the 'GOProfileHtest' interface, the one-sided confidence interval for the squared Euclidean distance is computed from the distance and its standard error stored in the corresponding fields of the argument goObject, itself typically an object of class 'GOProfileHtest' resulting from a call to 'compareGOProfiles' with simplify=T.

In the 'list' interface, the argument goObject is intended to be a list of objects of class 'GOProfileHtest' typically resulting from a call to 'compareGOProfiles' with simplify=F. The call to the method is iterated along the list, and the one-sided confidence interval for the squared Euclidean distance is computed from the distance and its standard error stored in the corresponding fields of all the members of the list.

In the default interface, the 'g0Object' argument is previously converted into an object of class 'ExpandedGOProfile' and then this interface is used.

If the argument 'd0' is not provided it is computed as $d0 < -s * equivEpsilon^2$, where 's' stands for the number of non empty GO nodes in any of the GO profiles being compared.
In the 'ExpandedGOProfile' and 'list' interfaces, the result is a list of objects of class "htest" or a single "htest" object if there is only one object in the list and simplify == T. In the other interfaces, the result is a single "htest" object. Each one of these "htest" objects has the following fields:

- **statistic**: test statistic, \((\text{distance} - d0) / se\)
- **parameter**: \(d0\) and the sample sizes (number of genes) \(n\) and \(m\)
- **p.value**: associated p-value to test the null hypothesis of profiles inequivalence
- **conf.int**: asymptotic one-sided confidence interval for the squared euclidean distance. Its attribute "conf.level" contains its nominal confidence level.
- **estimate**: squared euclidean distance between the contracted profiles. Its attribute "se" contains its standard error estimate
- **data.name**: a character string giving the names of the data
- **alternative**: a character string describing the alternative hypothesis (always 'Equivalence or similarity, true squared Euclidean distance between the contracted profiles is less than \(d0\)')

**Author(s)**

Jordi Ocana

**See Also**

'compareGOProfiles'

**Examples**

data(sampleProfiles)
comparedMF <- compareGOProfiles (pn=expandedWelsh01['MF']),
qm = expandedSingh01['MF'],
pq0 = commonExpandedWelsh01Singh01['MF']])
equivMF<- equivalentGOProfiles(comparedMF)
print(equivSummary(equivMF, decs=5))

---

**expandedLevel**

*Function to create expanded levels which can contain GO Terms at different GO levels*

**Description**

This function, combined with function expandTerm, allows to create mixed levels which can contain terms belonging to different GO levels. Specifically one can take one (or several, but one by one) term at a given GO level and expand it into its children terms using function expandTerm and then combine them into a new level using this function.

**Usage**

```r
expandedLevel(LevelTerms, Term2Expand, onto)
expandTerm(GOTerm, onto)
```
### expandedProfile

**Description**

Expanded profiles are used mainly for comparison of profiles based on the theory developed by Sanchez et al (2007) (see references)

**Usage**

```r
expandedProfile(genelist, idType = "Entrez", onto = "ANY", level = 2, orgPackage = NULL, multilevels = NULL, ord = TRUE, na.rm = TRUE, percentage = TRUE)
```

**Arguments**

- **genelist**: List of genes on which the Profile has to be based
- **idType**: Type of identifiers for the genes. Use 'Entrez' preferably
- **onto**: Ontology on which the profile has to be built
- **level**: Level of the ontology at which the profile has to be built
- **orgPackage**: Name of a Bioconductor’s organism annotations package ('org.Xx-eg-db'). This field must be provided if the gene list passed to the function is either a character vector of 'Entrez' (NCBI) identifiers or a character vector of probe names
fitGOProfile

anotPackage Name of Bioconductor annotations package. This field must be provided if the gene list passed to the function is a character vector of probe names

des

ord Set to 'TRUE' if the profile has to appear ordered by the category names

multilevels If it is not NULL it must be a vector of GO categories that defines the level at where the profile is built

na.rm Set to 'TRUE' if NAs should be removed

percentage Set to 'TRUE' if the profile must be built using percentages

Details

The function admits three types of entries: Entrez ('Entrez'), Bioconductor probe set names ('BioCprobes') or a special type of data frames ('GOTermsFrames'). If the identifier type are 'BioCprobes' then an annotation package name must be provided too.

Value

An object of class GOProfile containing an expanded profile

Author(s)

Alex Sanchez

References


See Also

basicProfile

Examples

data(CD4Ids)
CD4.Expanded <- expandedProfile(genelist=CD4LLids, onto='MF', level=2, orgPackage="org.Hs.eg.db")

fitGOProfile

Does a "sample" GO profile 'pn', observed in a sample of 'n' genes, fit a "population" or "model" p0?

Description

'fitGOProfile' implements some inferential procedures to solve the preceding question. These procedures are based on asymptotic properties of the squared euclidean distance between the contracted versions of pn and p0.

Usage

fitGOProfile(pn, p0, n = ngenes(pn), method = "lcombChisq", ab.approx = "asymptotic", nsims = 10000, nsims = 10000, simplify = T)
Arguments

- **pn**: an object of class `ExpandedGOProfile` representing one or more "sample" expanded GO profiles for a fixed ontology (see the 'Details' section)
- **p0**: an object of class `ExpandedGOProfile` representing one or more "population" or "theoretical" expanded GO profiles (see also the 'Details' section)
- **n**: a numeric vector with the number of genes profiled in each column of `pn`. This parameter is included to allow the possibility of exploring the consequences of varying sample sizes, other than the true sample size in `pn`
- **method**: the approximation method to the sampling distribution under the null hypothesis "p = p0", where p is the 'true' population profile originating each column of `pn`. See the 'Details' section below
- **ab.approx**: the method used to compute the constants 'a' and 'b' described in the paper. See the 'Details' section
- **confidence**: the confidence level of the confidence interval in the result
- **nsims**: some inferential methods require a simulation step; the number of simulation replicates is specified with this parameter
- **simplify**: should the result be simplified, if possible? See the 'Details' section

Details

An object of class 'ExpandedGOProfile' is, essentially, a 'data.frame' object with each column representing the relative frequencies in all observed node combinations, resulting from profiling a set of genes, for a given and fixed ontology. The row.names attribute codifies the node combinations and each data.frame column (say, each profile) has an attribute, 'ngenes', indicating the number of profiled genes. (Actually, the 'ngenes' attribute of each 'p0' column is ignored and is taken as if it were infinite, 'Inf'.) The arguments 'pn' and 'p0' are compared in a column by column wise, recycling columns, if necessary, in order to perform max(ncol(pn),ncol(p0)) comparisons (each comparison resulting in an object of class 'htest'). In order to be properly compared, 'pn' and 'p0' are expanded by row, according to their row names. That is, both arguments can have unequal row numbers. Then, they are expanded adding rows with zero frequencies, in order to make them comparable.

In the i-th comparison (i from 1 to max(ncol(pn),ncol(p0))), if p stands for the profile originating the sample profile `pn[,i]` and d(,) for the squared euclidean distance, if p != `p0[,i]`, the distribution of \( \sqrt{n} \left( d(pn[,i],p0[,i]) - d(p,p0[,i]) \right) / se \) is approximately standard normal, N(0,1). This provides the basis for the confidence interval in the result field `conf.int`. When `p==p0[,i]`, the asymptotic distribution of n d(pn[,i],p0[,i]) is the distribution of a linear combination of independent chi-square random variables, each one with one degree of freedom. This sampling distribution may be directly computed (approximating it by simulation, `method="lcombChisq"`) or approximated by a chi-square distribution, based on two correcting constants a and b (`method="chi-square"`). These constants are chosen to equate the first two moments of both distributions (the distribution of a linear combination of chi square variables and the approximating chi-square distribution). When `method="chi-square"`, the returned test statistic value is the chi-square approximation \( n \left( d(pn,p0) - b \right) / a \). Then, the result field 'parameter' is a vector containing the 'a' and 'b' values and the number of degrees of freedom, 'df'. Otherwise, the returned test statistic value is n d(pn,p0) and 'parameter' contains the coefficients of the linear combination of chi-squares

Value

A list containing max(ncol(pn),ncol(p0)) objects of class 'htest', or a single 'htest' object if ncol(pn)==1 and ncol(p0)==1 and simplify == T. Each 'htest' object has the following fields:
Description

Performs Gene Ontology based analysis for gene sets or other type of biological identifiers which can be annotated in the Gene Ontology.

Details

<table>
<thead>
<tr>
<th>Package</th>
<th>goProfiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Package</td>
</tr>
<tr>
<td>Version</td>
<td>1.0</td>
</tr>
<tr>
<td>Date</td>
<td>2007-06-12</td>
</tr>
<tr>
<td>License</td>
<td>GPL</td>
</tr>
</tbody>
</table>
Author(s)

Alex Sanchez and Jordi Ocana

References


See Also

goTools, GOstat, ontoTools, and other Bioconductor packages for GO based analysis

hugoIds

Entrez Identifiers obtained from the Human Genome Organization

Description

Entrez identifiers obtained from the Human Genome Organization. They correspond to the column named ‘Entrez Gene Id (mapped)’ in the ‘All data’ table in the Hugo Genome Nomenclature website (http://www.genenames.org/index.html)

Usage

data(hugoIds)

References

http://www.genenames.org/cgi-bin/hgnc_downloads.cgi

Examples

data(hugoIds)

mergeProfilesLists

Combines two lists of profiles into one

Description

Combines two lists of profiles, that is two lists with three components, 'MF', 'BP', 'CC' into a single one.

Usage

mergeProfilesLists(profilesList1, profilesList2, emptyCats = F, profNames = NULL)
omimIds

Arguments

profilesList1
  First list to combine
profilesList2
  Second list to combine
emptyCats
  Boolean. Set to TRUE if there are empty categories that should be accounted for in any of the profiles
profNames
  Names for the profiles (optional). If missing they are set to 'Frequency-1', 'Frequency-2', etc.

Value

A list of profiles with more than one column each.

Author(s)

Alex Sanchez

Examples

data(sampleProfiles)
merged<-mergeProfilesLists (basicSingh01, basicWelsh01, profNames=c('Welsh','Singh'))

Description

Entrez identifiers for several lists of genes related with human disease.

diseaseIds contains the Entrez identifiers corresponding to disease-related genes found in the OMIM database. This list has been manually curated by Nuria Lopez-Bigas et al. who kindly provided it to us.
morbidmapIds contains the Entrez identifiers for all the genes in the morbidmap table. This list would correspond to disease-related genes if there had been no manual curation, as in the previous list ('diseaseIds')
dominantIds contains the Entrez identifiers for dominant genes after manual curation by Nuria Lopez-Bigas who has kindly allowed us to include them in the package.
recessiveIds contains the Entrez identifiers for recessive genes after manual curation by Nuria Lopez-Bigas who has kindly allowed us to include them in the package.
dominantIdsEBI contains the Entrez identifiers for dominant genes in the EBI version of the OMIM database recovered using SRS with the term 'dominant' in the KEYWORDS field.
recessiveIdsEBI contains the Entrez identifiers for recessive genes in the EBI version of the OMIM database recovered using SRS with the term 'recessive' in the KEYWORDS field.
dominantIdsNCBI contains the Entrez identifiers for dominant genes in the NCBI version of the OMIM database recovered using ENTREZ with the term 'dominant' in the CLINICAL field.
recessiveIdsNCBI contains the Entrez identifiers for recessive genes in the NCBI version of the OMIM database recovered using ENTREZ with the term 'recessive' in the CLINICAL field.
Usage

data(omimIds)

Format

Each dataset is a character vector with a different number of elements which (should) correspond to valid Entrez identifiers.

Details

Lopez-Bigas et al. analyzed the distribution of functional categories in genes causing disease in human. They did several comparisons which can also be done using goProfiles. In order to perform these comparisons we first tried to obtain the same lists of genes using standard database browsers, such as 'SRS', at the European Bioinformatics Institute, or 'Entrez', at the National Center for Biotechnological Information. Curiously both approaches provided very different lists so we asked the authors for their data and they kindly provided them to us. In order to facilitate the use of functions included in goProfiles we have trimmed the list of recessive and dominant genes so that (i) They become exclusive (no gene belongs to both lists) (2) They are both included in the diseaseIds list. This eliminated 39 genes (out of 639) from the list of recessive genes and 52 genes (out of 414) from the list of dominant genes.

References


Examples

data(omimIds)

plotProfiles  

Plot functional profiles

Description

Plots basic functional profiles created with the `basicProfile` instruction. If several profiles have to be plot together they must be first merged using the `mergeProfiles` function. The labels of the Y-axis of the plots are the descriptions of the GO Terms. If the label is longer than 20 characters it is truncated and ended by three dots.

Usage

plotProfiles(aProf, aTitle = "Functional Profile", anOnto = NULL, percentage = FALSE, HORIZVERT = TRUE, legendText = NULL, colores = c("white", "red"), multiplePlots = FALSE)
printProfiles

**Arguments**

- `aProf` Functional profile to plot
- `aTitle` Title for the figures
- `anOnto` Ontology (to appear in the title)
- `percentage` Plot absolute or relative frequencies (not summing to 100)
- `HORIZVERT` Plot horizontal or vertical bars
- `legendText` Text of the legend for the plot
- `colores` Colors to be used
- `multiplePlots` Plot all profiles for a given dataset in one figure
- `multipleWindows` Open a new window after each plot
- `labelWidth` Width of Y axis labels (Names of GO categories) in the plot

**Value**

The plot

**Author(s)**

Alex Sanchez

**Examples**

```r
data(sampleProfiles)
plotProfiles(basicWelsh01[['MF']], 'Functional profiles for Welsh dataset', percentage=TRUE)
merged<-mergeProfilesLists (basicSingh01, basicWelsh01, profNames=c('Welsh','Singh'))
plotProfiles(merged, percentage=TRUE, multiplePlots=TRUE, labelWidth=30)
```

---

**printProfiles**  
*Print functional profiles*

**Description**

Prints basic functional profiles created with the 'basicProfile' instruction. Allows for several formatting operations such as truncating long labels, removing empty categories or choosing between absolute or relative frequencies. If several profiles have to be printed together they must be first merged using the 'mergeProfiles' function.

**Usage**

```r
printProfiles(aProf, aTitle = "Functional Profile", anOnto = NULL, percentage = FALSE, Width=25, emptyCats=FALSE)
```
**Arguments**

- **aProf**
  - Functional profile to plot
- **aTitle**
  - Title for the figures
- **anOnto**
  - Ontology (to appear in the title)
- **percentage**
  - Plot absolute or relative frequencies (not summing to 100)
- **Width**
  - Maximum width for the description of GO categories
- **emptyCats**
  - Set to 'TRUE' if empty categories should appear in the profile

**Value**

The printout

**Author(s)**

Alex Sanchez

**Examples**

data(sampleProfiles)
printProfiles(basicWelsh01[['MF']],'Functional profiles for Welsh dataset', anOnto='MF', percentage=TRUE)
printProfiles(basicWelsh01,'Functional profiles for Welsh dataset',anOnto='MF', percentage=TRUE, emptyCats=TRUE)
merged<-mergeProfileLists (basicSingh01, basicWelsh01, profNames=c('Welsh','Singh'))
printProfiles(merged, percentage=TRUE, emptyCats=TRUE)

---

**Description**

Entrez identifiers for genes related with Prostate Cancer selected from two datasets analyzed by Welsh et al. (2001) and Singh et al. (2002) respectively. The genes have been selected from freely available datasets in the internet using a standard workflow for selecting differentially expressed genes. The dataset contains 4 character vectors, each corresponding to the entrez identifiers of the genes selected at a 5% and 1% significance level from the Welsh and Singh dataset respectively.

- **welsh05EntrezIDs** List of genes selected from Welsh et al. study at a 0.05 significance level.
- **welsh01EntrezIDs** List of genes selected from Welsh et al. study at a 0.01 significance level.
- **singh05EntrezIDs** List of genes selected from Singh et al. study at a 0.05 significance level.
- **singh01EntrezIDs** List of genes selected from Singh et al. study at a 0.01 significance level.

**Usage**

data(prostateIds)

**Format**

Each dataset is a character vector with a different number of elements which (should) correspond to valid Entrez identifiers


Examples

data(prostateIds)

data(sampleProfiles)

Description

This dataset contains some basic and expanded sample profiles computed on some of the datasets in the package. They are intended to be used in examples of profile manipulation, without having to recompute the profiles each time the example is checked. The prefix of the object indicates if it is a 'basic' or an 'expanded' profile, whereas the name can be clearly associated to some of the datasets available in the package.

basicWelsh01 Basic (contracted) profile made from Welsh01EntrezIds dataset with default parameters.

basicSingh01 Basic (contracted) profile made from Singh01EntrezIds dataset with default parameters.

basicCD4 Basic (contracted) profile made from CD4LLids dataset with default parameters.

basicHugo Basic (contracted) profile made from hugoIds dataset with default parameters.

expandedWelsh01 Expanded profile made from Welsh01EntrezIds dataset with default parameters.

expandedSingh01 Expanded profile made from Singh01EntrezIds dataset with default parameters.

commonExpandedWelsh01Singh01 Expanded profile made from common genes in Welsh01EntrezIds and Singh01EntrezIds dataset with default parameters.

expandedCD4 Expanded profile made from CD4LLids dataset with default parameters.

expandedHugo Expanded profile made from hugoIds dataset with default parameters.

Usage

data(sampleProfiles)
Examples

data(sampleProfiles)
Index

*Topic* datasets
- CD4Ids, 1
- drosophila, 11
- hugoIds, 19
- omimIds, 20
- prostateIds, 23
- sampleProfiles, 24

*Topic* hplot
- plotProfiles, 21
- printProfiles, 22

*Topic* htest
- basicProfile, 3
- compareGeneLists, 7
- compareGOProfiles, 5
- compareProfilesLists, 8
- compSummary, 4
- equivalentGOProfiles, 12
- equivSummary, 11
- expandedProfile, 15
- fitGOProfile, 16

*Topic* manip
- conversionFunctions, 10
- expandedLevel, 14
- GOTermsList, 1

*Topic* package
- goProfiles-package, 18

*Topic* utilities
- mergeProfilesLists, 19

```
as.GOTerms.frame
   (conversionFunctions), 10
as.GOTerms.list
   (conversionFunctions), 10

basicCD4(sampleProfiles), 24
basicHugo(sampleProfiles), 24
basicProfile, 3, 8
basicSingh01(sampleProfiles), 24
basicWelsh01(sampleProfiles), 24
BioCpack2EntrezIDS
   (conversionFunctions), 10
BioCpack2Profiles
   (conversionFunctions), 10
   BioCprobes2Entrez
      (conversionFunctions), 10
   CD4GOTermsFrame(CD4Ids), 1
   CD4GOTermsList(CD4Ids), 1
   CD4Ids, 1
   CD4Llids(CD4Ids), 1
   commonExpandedWelsh01Singh01
      (sampleProfiles), 24
   compareGeneLists, 7, 9
   compareGOProfiles, 5, 8
   compareProfilesLists, 8
   compSummary, 4
   conversionFunctions, 10
   diseaseIds(omimIds), 20
   dominantIds(omimIds), 20
   dominantIdsEBI(omimIds), 20
   dominantIdsNCBI(omimIds), 20
   drosophila, 11
   drosophilaIds(drosophila), 11
   equivalentGOProfiles, 12
   equivSummary, 11
   expandedCD4(sampleProfiles), 24
   expandedHugo(sampleProfiles), 24
   expandedLevel, 14
   expandedProfile, 9, 15
   expandedSingh01(sampleProfiles), 24
   expandedWelsh01(sampleProfiles), 24
   expandTerm(expandedLevel), 14
   fitGOProfile, 16
   getAncestorsList(GOTermsList), 1
   getGOLevel(GOTermsList), 1
   goProfiles(goProfiles-package), 18
   goProfiles-package, 18
   GOTermsFrame2GOTermsList
      (conversionFunctions), 10
   GOTermsList, 1
```

26
hugoIds, 19
mergeProfilesLists, 19
michaudIds (drosophila), 11
morbidmapIds (omimIds), 20
omimIds, 20
ostrinIds (drosophila), 11
plotProfiles, 21
printProfiles, 22
prostateIds, 23
recessiveIds (omimIds), 20
recessiveIds EBI (omimIds), 20
recessiveIds NCBI (omimIds), 20
sampleProfiles, 24
singh01EntrezIDs (prostateIds), 23
singh05EntrezIDs (prostateIds), 23
welsh01EntrezIDs (prostateIds), 23
welsh05EntrezIDs (prostateIds), 23