Package 'SemDist'

January 20, 2025

Version 1.40.0

Date 2014-09-04

Title Information Accretion-based Function Predictor Evaluation

Author Ian Gonzalez and Wyatt Clark

Maintainer Ian Gonzalez <gonzalez.isv@gmail.com>

Depends R (>= 3.1), AnnotationDbi, GO.db, annotate

Suggests GOSemSim

Description This package implements methods to calculate information accretion for a given version of the gene ontology and uses this data to calculate remaining uncertainty, misinformation, and semantic similarity for given sets of predicted annotations and true annotations from a protein function predictor.

biocViews Classification, Annotation, GO, Software

License GPL (>= 2)

URL http://github.com/iangonzalez/SemDist

git_url https://git.bioconductor.org/packages/SemDist

git_branch RELEASE_3_20

git_last_commit 4eead33

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2025-01-19

Contents

computeIA *Compute information accretion for an ontology*

Description

Calculates information accretion for each term in the specified ontology using either user -specified data or the sequence annotations for the organisms specified (note that organism-specific pacakges must be downloaded separately. See "note" section).

Usage

```
computeIA(ont, organism, evcodes = NULL, specify.ont = FALSE,
          myont = NULL, specify.annotations = FALSE,
          annotfile = NULL)
```

Arguments

ont	Character representation of ontology version to use. One of "CC", "MF", or "BP" , corresponding to Cellular Component, Molecular Function, and Biological Process.
organism	A character vector indicating which organism's annotation data to use.
evcodes	A character vector specifying which evidence codes to use in the ontology data. Default NULL value causes all codes to be used.
specify.ont	A boolean indicating whether the user wants to specify their own version of the ontology.
myont	Character object indicating what file to read in the specified ontology from. The ontology should be specified as a tab-delimited file with 2 columns (no header). Each row in the file should indicate a parent-child relationship between two GO accessions (e.g. "GO:0003674 GO:0004000")
specify.annotations	Boolean indicating whether the user wants to specify sequence annotations from a file. Should only be TRUE if specify.ont is TRUE.
annotfile	Character object indicating which file to read sequence annotations from. Should be a tab-delimited file with 2 columns. The first column is a list of sequences, the second is a list of GO accessions in the same rows as the sequences they annotate.

Value

Does not return a specific value. Saves the information accretion values for each term in the ontology in a .rda file that specifies the organism and the ont version. Parent count and term count objects are also saved in similarly formatted files so that IA calculations from multiple organisms can be combined.

Note

In order to compute IA for an organism, the specific annotation data set for that organism must be installed by the user. Here is a list of supported organisms (names in the format that should be passed to computeIA) and the correspoinding packages needed:

anopheles = org.Ag.eg.db

arabidopsis = org.At.tair.db

bovine = org.Bt.eg.db

canine = org.Cf.eg.db

chicken = org.Gg.eg.db

chimp = org.Pt.eg.db

ecolik12 = org.EcK12.eg.db

fly = org.Dm.eg.db

human = org.Hs.eg.db

malaria = org.Pf.plasmo.db

mouse = org.Mm.eg.db

pig = org.Ss.eg.db

rat = org.Rn.eg.db

rhesus = org.Mmu.eg.db

worm = org.Ce.eg.db

xenopus = org.Xl.eg.db

yeast = org.Sc.sgd.db

zebrafish = org.Dr.eg.db

Author(s)

Ian Gonzalez and Wyatt Clark

See Also

RUMIcurve findRUMI

Examples

```
# Calculate IA, specify ontology and annotations
ontfile <- system.file("extdata", "mfo_ontology.txt", package="SemDist")
annotations <- system.file("extdata", "MFO_LABELS_TEST.txt", package="SemDist")
computeIA("my", "values", specify.ont=TRUE,
          myont=ontfile, specify.annotations=TRUE,
          annotfile=annotations)
```

---

findRUMI *Information accretion based predictor assessment*

Description

Reads in a file containing the true terms annotating a set of sequences and a file containing the predicted terms and scores for a set of sequences and outputs a data frame containing the remaining uncertainty and misinformation values for the predictions made for each sequence.

Usage

```
findRUMI(ont, organism, threshold = 0.05, truefile="",
          predfile = "", IAccr = NULL)
```

Arguments

ont	Character representation of ontology version to use. One of "CC", "MF", or "BP" , corresponding to Cellular Component, Molecular Function, and Biological Process.
organism	A character vector indicating which organism(s) annotation data to use.
threshold	Score above which a predicted annotation should be included in the calculation. Must be a numeric value between 0 and 1, or else findRUMI throws an error.
truefile	Text file from which to read true annotations of sequences. Should be a tab-delineated file with 2 columns: Sequences and GO terms (accessions).
predfile	Text file from which to read predicted annotations of sequences. Should be a tab-delineated file with 3 columns: Sequences, GO terms (accessions), and probability score from 0 to 1 for each prediction.
IAccr	A variable containing a named numeric vector of IA values for all the GO terms being used that will be used for calculations instead of R packages. This argument is optional.

Value

A data frame containing the RU and MI values for each sequence in the file.

Author(s)

Ian Gonzalez and Wyatt Clark

See Also

computeIA RUMIcurve

Examples

```
# Using test data sets from SemDist, calculate RU and MI:
truefile <- system.file("extdata", "MFO_LABELS_TEST.txt", package="SemDist")
predfile <- system.file("extdata", "MFO_PREDS_TEST.txt", package="SemDist")
rumiTable <- findRUMI("MF", "human", 0.75, truefile, predfile)
avgRU <- mean(rumiTable$RU)
avgMI <- mean(rumiTable$MI)
```

IAccr *Information Accretion Data*

Description

This data set contains the information accretion values for each term in the requested ontology/species.

Usage

```
IAccr
```

Format

A named numeric vector with one value corresponding to each GO accession in the ontology.

Source

The gene ontology data was obtained from the GO.db package and the annotation data was obtained from the following packages for each organism:

anopheles = org.Ag.eg.db

arabidopsis = org.At.tair.db

bovine = org.Bt.eg.db

canine = org.Cf.eg.db

chicken = org.Gg.eg.db

chimp = org.Pt.eg.db

ecolik12 = org.EcK12.eg.db

fly = org.Dm.eg.db

human = org.Hs.eg.db

malaria = org.Pf.plasmo.db

mouse = org.Mm.eg.db

pig = org.Ss.eg.db

rat = org.Rn.eg.db

rhesus = org.Mmu.eg.db

worm = org.Ce.eg.db

xenopus = org.Xl.eg.db

yeast = org.Sc.sgd.db

zebrafish = org.Dr.eg.db

Examples

```
data("Info_Accretion_mouse_CC", package = "SemDist")
str(IAccr)
```

parentcnt *Parent Count Data*

Description

This data set contains the parent count values for each term in the requested ontology/species (the number of times that the term's parents annotate a protein). This can be used along with the term count to calculate information accretion.

Usage

```
parentcnt
```

Format

A named numeric vector with one value corresponding to each GO accession in the ontology.

Source

The gene ontology data was obtained from the GO.db package and the annotation data was obtained from the following packages for each organism:

anopheles = org.Ag.eg.db

arabidopsis = org.At.tair.db

bovine = org.Bt.eg.db

canine = org.Cf.eg.db

chicken = org.Gg.eg.db

chimp = org.Pt.eg.db

ecolik12 = org.EcK12.eg.db

fly = org.Dm.eg.db

human = org.Hs.eg.db

malaria = org.Pf.plasmo.db

mouse = org.Mm.eg.db

pig = org.Ss.eg.db

rat = org.Rn.eg.db

rhesus = org.Mmu.eg.db

worm = org.Ce.eg.db

xenopus = org.Xl.eg.db

yeast = org.Sc.sgd.db

zebrafish = org.Dr.eg.db

Examples

```
data("Parent_Count_mouse_CC", package = "SemDist")
str(parentcnt)
```

RUMIcurve *Information accretion based predictor assessment (across many thresholds)*

Description

Reads in a (tab-delimited) file containing the true annotations for a set of sequences, a (tab-delimited) file containing the predicted annotations and corresponding scores for the same sequences. Calculates and outputs the average remaining uncertainty, misinformation, and semantic similarity at a series of user-specified thresholds.

Usage

```
RUMIcurve(ont, organism, increment = 0.05, truefile, predfiles,
          IAccr = NULL, add.weighted = FALSE,
          add.prec.rec = FALSE)
```

Arguments

ont	Character representation of ontology version to use. One of "CC", "MF", or "BP" , corresponding to Cellular Component, Molecular Function, and Biological Process.
organism	A character vector indicating which organism(s) annotation data to use.
increment	A numeric value between 0 and 1 indicating the distance between each threshold that should be calculated. Note that the iteration starts from a threshold of 1, so an increment value of 0.08 will result in the thresholds 0.92, 0.84, 0.76 ... being used.
truefile	A character vector indicating the file from which to read the true annotations for the given sequences. Should be tab-delimited, with the first column containing the sequence ids and the second containing GO accessions.
predfiles	A character vector containing which files to read in as the predicted annotations. Should be tab-delimited, with the first column containing sequences, the second column containing GO accessions, and the third column containing the predictors 0-1 score for that prediction.
IAccr	A variable containing a named numeric vector of IA values for all the GO terms being used that will be used for calculations instead of R packages. This argument is optional.
add.weighted	A boolean indicating whether or not to add calculation of information content weighted versions of RU, MI, and SS to the output.
add.prec.rec	A boolean indicating whether or not to calculate precision, recall and specificity values for the prediction at each threshold and add to the output.

Value

Returns a named list with the same number of elements as the input "predfiles". Each element is a data frame containing all of the user-requested values for the data at each threshold.

Author(s)

Ian Gonzalez and Wyatt Clark

See Also

computeIA findRUMI

Examples

```
# Using test data sets from SemDist, plot a RUMI curve:
truefile <- system.file("extdata", "MFO_LABELS_TEST.txt", package="SemDist")
predfile <- system.file("extdata", "MFO_PREDS_TEST.txt", package="SemDist")
avgRUMIvals <- RUMIcurve("MF", "human", 0.05, truefile, predfile)
firstset <- avgRUMIvals[[1]]
plot(firstset$RU, firstset$MI)
```

---

termcnt *Term Count Data*

Description

This data set contains the term count values for each term in the requested ontology/species (the number of times that the term annotates a protein). This can be used along with the parent count to calculate information accretion.

Usage

```
termcnt
```

Format

A named numeric vector with one value corresponding to each GO accession in the ontology.

Source

The gene ontology data was obtained from the GO.db package and the annotation data was obtained from the following packages for each organism:

anopheles = org.Ag.eg.db

arabidopsis = org.At.tair.db

bovine = org.Bt.eg.db

canine = org.Cf.eg.db

chicken = org.Gg.eg.db

chimp = org.Pt.eg.db

ecolik12 = org.EcK12.eg.db

fly = org.Dm.eg.db

human = org.Hs.eg.db

malaria = org.Pf.plasmo.db

mouse = org.Mm.eg.db

pig = org.Ss.eg.db

rat = org.Rn.eg.db

rhesus = org.Mmu.eg.db

worm = org.Ce.eg.db

xenopus = org.Xl.eg.db

yeast = org.Sc.sgd.db

zebrafish = org.Dr.eg.db

Examples

```
data("Term_Count_mouse_CC", package = "SemDist")
str(termcnt)
```

Index