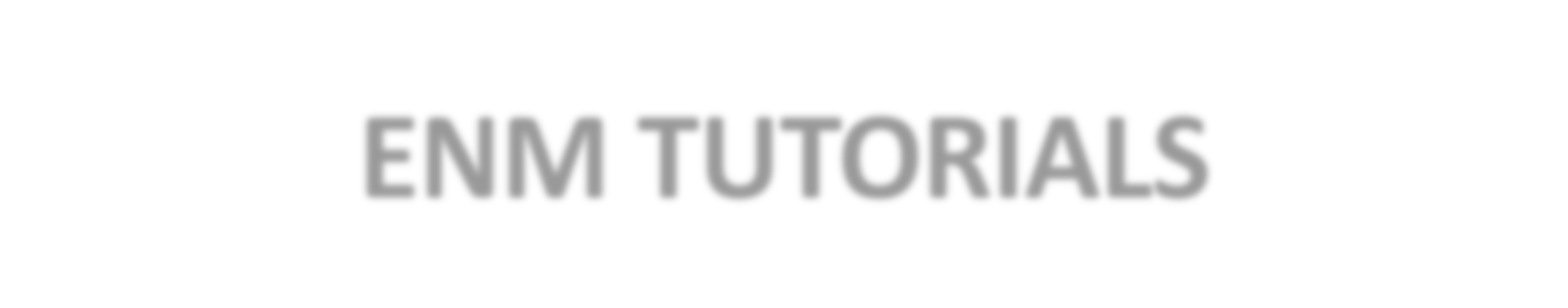
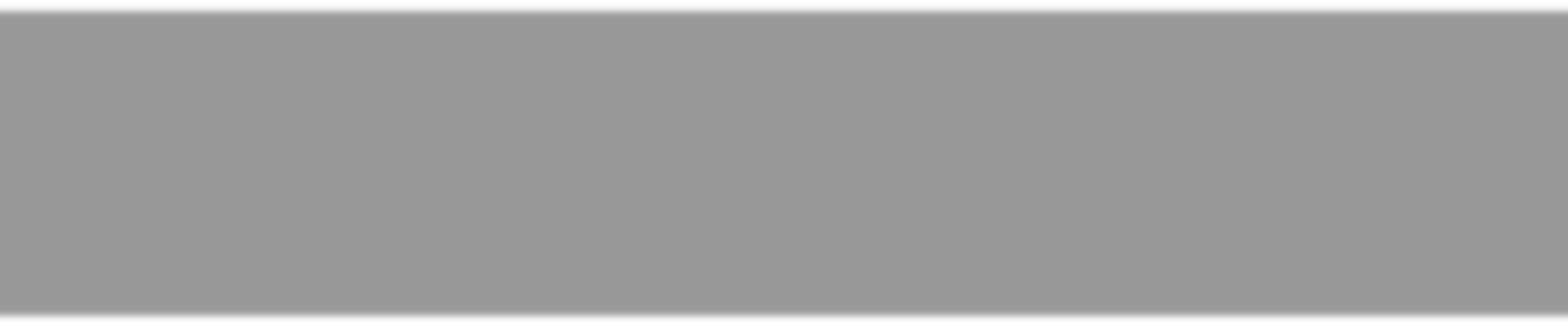
#### ENM TUTORIALS



**ENM TUTORIALS**



Omics descriptors calculation R package

|  |  |
| --- | --- |
| RELEASE DATE: | 31.1.2016 |
| USE: | Integrate omics data with pathway information to derive a new set of omics descriptors |
| VERSION: | V.0.9 |
| MAIN AUTHOR: | Georgia Tsiliki |
| PARTNER: | NTUA |
| CONTACT DETAILS: | [gtsiliki@central.ntua.gr](mailto:gtsiliki@central.ntua.gr)  +30 210 772 3236 |
| AUTHORS: | Georgia Tsiliki, Haralambos Sarimveis |





**TABLE OF CONTENTS**

1. [INTRODUCTION 3](#_bookmark0)
2. [APPLICATION DETAILS 3](#_bookmark1)
3. [ACKNOWLEDGMENTS 5](#_bookmark2)
4. [REFERENCES 5](#_bookmark3)
5. [ANNEXES 6](#_bookmark4)

[ANNEX I: GODESCRPRED 7](#_bookmark5)

[ANNEX II: GODESCRCALCULUS 8](#_bookmark6)

[6. KEYWORDS 9](#_bookmark7)

eNanoMapper 604134 31 January 2016 ENM TUTORIALS Page **2** of **9**



# INTRODUCTION

This document demonstrates the specifications of a novel methodology aiming to generate a new set of descriptors, referred to as GO descriptors, which would efficiently summarize omics data and also integrate Gene Ontology (GO) data. Our goal is to enrich the data using gene set information whilst emphasizing the importance of -omics data in modelling ENM toxicity. GO was selected as the golden standard for annotation in three ontology branches, namely Cellular Components (CC), Molecular Functions (MF), and Biological Processes (BP) containing 41,694 classes.

GO descriptors aim to summarize the omics data by their GO classes, and our particular use case was protein corona proteomics data (Walkey et al., 2014) integrated with pathway information from GO database. Summarization is performed by clustering the GO classes of the data. The number of the GO descriptors equals the number of clusters of those GO groups found to be highly enriched in the supplied data set. Clustering is performed using hierarchical clustering or bi-clustering algorithms (Parmeet et al., 2014) (Aggarwal and Reddy 2013) where number of clusters is a user-specified parameter for both clustering algorithms. Thus by defining the number of clusters the user can decide to what extent the data should be summarized, but also to what depth should the interaction relationships be detailed between the significant GO ids.

GO descriptors can be used in nanoQSAR models, possibly together with physicochemical descriptors, and be further exploited for their biological relevance and functional similarity.

# APPLICATION DETAILS

Two separate R packages have been created, called GOdescrPred and GOdescrCalculus. Although they both follow the above procedure, we have made some structural changes between the two, so that the user can either:

* + select to save the suggested clustering model (GOdescrPred), or
  + generate the new data set of GO descriptors (GOdescrCalculus).

In particular, GOdescrPred aims to save the cluster memberships of the proteins for a particular data set, and based on them to produce GO descriptors for a new set of data. In this case, it is a prerequisite that the new data given by the user should include the same protein/gene/ etc. ids as the original. Alternatively, GOdescrCalculus can be employed when the user wants to calculate GO descriptors for a particular set of omics data and does not have the ‘clustering model’, i.e. those set of proteins have not been previously used to create a set of GO descriptors. If the objective is to just create a set of new descriptors, then GOdescrCalculus should be used, where its output will be a data matrix (or array) with different columns corresponding to different GO descriptors.

GOdescrPred includes the following functions:

* + generate.biclust.model to generate a model using biclustering algorithm, given the supplied data set

eNanoMapper 604134 31 January 2016 ENM TUTORIALS Page **3** of **9**

* + generate.hierar.model to generate a model using hierarchical clustering, given the supplied data set



* + pred.descr to produce a new set of descriptors based on the ‘model’, in this case the cluster memberships supplied by either generate.biclust.model or generate.hierar.model

Further details can be found in Annex I and <http://147.102.82.122/ocpu/library/GOdescrPred/info>.

GOdescrCalculus includes the following functions:

* + generate.param.model to generate a serialized raw R model which in this case only includes a list with the parameters needed for the clustering of the omics data
  + generate.descr.biclust to generate the new set of GO descriptors using bi-clustering
  + generate.descr.hierar to generate the new set of descriptors using hierarchical clustering. Further details can be seen in Annex II and <http://147.102.82.122/ocpu/library/GOdescrCalculus/info>.

eNanoMapper 604134 31 January 2016 ENM TUTORIALS Page **4** of **9**



# ACKNOWLEDGMENTS

The eNanoMapper project is funded by the European Union's Seventh Framework Programme for research, technological development and demonstration (FP7-NMP-2013-SMALL-7) under grant agreement no. 604134.

# REFERENCES

* 1. Walkey, C. D.; Olsen, J. B.; Song, F.; Liu, R.; Guo, H.; Olsen, D. W. H.; Cohen, Y.; Emili, A.; Chan, W. C. W. Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. *ACS Nano* **2014**, *8*, 2439–2455 DOI: 10.1021/nn406018q.
  2. Parmeet, B.; Iovleff, S.; Govaert, G. [blockcluster: An R Package for Model Based Co-Clustering.](http://scholar.google.com/scholar_lookup?title=blockcluster%3A%20An%20R%20Package%20for%20Model%20Based%20Co-Clustering&amp;author=P.%20Bhatia&amp;author=S.%20Iovleff&amp;author=G.%20Govaert&amp;publication_year=2014) **2014.**
  3. Aggarwal, C. C.; Reddy, C.K. eds. *Data clustering: algorithms and applications*. CRC Press, **2013**.

eNanoMapper 604134 31 January 2016 ENM TUTORIALS Page **5** of **9**



# ANNEXES

eNanoMapper 604134 31 January 2016 ENM TUTORIALS Page **6** of **9**

## ANNEX I: GODESCRPRED



eNanoMapper 604134 31 January 2016 ENM TUTORIALS Page **7** of **9**

**Package ‘GOdescrPred’**

###### January 28, 2016

**Type** Package

**Title** Creates GOdescr model

**Version** 1.0

**Date** 2015-04-15

**Author** Georgia Tsiliki

**Maintainer** Georgia Tsiliki [<gtsiliki@central.ntua.gr>](mailto:gtsiliki@central.ntua.gr)

Depends

jsonlite, RCurl, pmml, org.Hs.eg.db, GSEABase, GOstats, Rgraphviz, blockcluster, Matrix, vegan

**Description** This package produces the model for GO descriptors based on GO ontology and pro- teomics data supplied. The resuting model are class memberships for proteins.

**License** GPL-2

### R topics documented:

GOdescrPred-package 2

dat1 3

dat1i 3

dat1m 4

dat1p 5

dat4h 5

generate.biclust.model 6

generate.hierar.model 8

pred.descr 9

read.in.json.for.pred 10

Index 12

1

2 *GOdescrPred-package*

GOdescrPred-package *Generate GO descriptors given an omics data*

Description

This package expects an omics data (e.g. proteomics/genomics), creates an R raw model to store the cluster memberships based on the data and the algorirthm selected; bi-clustering and hierarchical clustering algorithms are currently implemented. The cluster memberships can be stored as an R raw model and used for ’prediction’, i.e. to construct new decsriptors for similar data. By similar, we mean that the new data needs to include the genes/proteins/etc used for the initial clustering.

Details

Package: GOdescrPred Type: Package

Version: 1.0

Date: 2015-04-15

License: GPL-2

Generate GO descriptors given an omics data. Important functions are generate.biclust.model and generate.hierar.model.

Author(s)

Georgia Tsiliki

Maintainer: Georgia Tsiliki [<gtsiliki@central.ntua.gr>](mailto:gtsiliki@central.ntua.gr)

References

1. [blockcluster](https://cran.r-project.org/web/packages/blockcluster/index.html) 2.[vegan](https://cran.r-project.org/web/packages/vegan/index.html) 3.[GSEABase](http://bioconductor.org/packages/release/bioc/html/GSEABase.html) 4.[GOstats](https://www.bioconductor.org/packages/3.3/bioc/html/GOstats.html)

Examples

data("dat1p")

data("dat1i")

data("dat1m")

data.file<- read.in.json.for.pred(dat1p, dat1m, dat1i)

*dat1* 3

dat1 *A sample data object*

Description

The dataset for this test is a data frame

Usage

data("dat1")

Format

A list of two objects

**datasetURI** a character vector- ambit data set uri

**dataEntry** a data frame containing two columns: compound and values. Compound is a character vector with all compound anbit uris, and values is a data frame with all numberic values of the proteomics data set (compounds by features). Only dependent features are included.

Details

Please find more about the data set in $dataEntry at <http://pubs.acs.org/doi/abs/10.1021/nn406018q?journalCode=ancac3>

References

[Walkey et al., 2014](http://pubs.acs.org/doi/abs/10.1021/nn406018q?journalCode=ancac3)

Examples

data(dat1)

## maybe str(dat1) ; plot(dat1) ...

dat1i *Additional information needed for GO descriptors model.*

Description

Here we specify the function with which we summarize GO descriptor cluster memberships from dat1m given dat1 data set. For example, mean, sd, var are some typical functions that can be provided.

Usage

data("dat1i")

4 *dat1m*

Format

A data frame with one observation.

X.mean. a character vector, giving the name of the function used for summary

Details

Example function to summarize dat1 using dat1m cluster memberships

References

There are no references

Examples

data(dat1i)

## maybe str(dat1i) ; plot(dat1i) ...

dat1m *Serialized GO descriptors model file*

Description

A character string for a serialized GO descriptors model, i.e. provides clustering memberships for all proteins in proteomics data dat1.

Usage

data("dat1m")

Format

A character string

Details

Example GO descriptors model based on dat1

References

There are no references

Examples

data(dat1m)

## maybe str(dat1m) ; plot(dat1m) ...

*dat1p* 5

dat1p *A sample data object*

Description

The dataset for this test is a data frame

Usage

data("dat1p")

Format

A list of two objects

**datasetURI** a character vector- ambit data set uri

**dataEntry** a data frame containing two columns: compound and values. Compound is a character vector with all compound anbit uris, and values is a proteomics data frame with all numberic values of the data set (compounds by features)

Details

Data set for prediction with dat1m

References

[Walkey et al., 2014](http://pubs.acs.org/doi/abs/10.1021/nn406018q?journalCode=ancac3)

Examples

data(dat1p)

## maybe str(dat1p) ; plot(dat1p) ...

dat4h *A sample data object*

Description

The dataset for this test is a data frame

Usage

data("dat1p")

6 *generate.biclust.model*

Format

A list of two objects

**datasetURI** a character vector- ambit data set uri

**dataEntry** a data frame containing two columns: compound and values. Compound is a character vector with all compound anbit uris, and values is a proteomics data frame with all numberic values of the data set (compounds by features)

Details

Data set for prediction with dat1m

References

[Walkey et al., 2014](http://pubs.acs.org/doi/abs/10.1021/nn406018q?journalCode=ancac3)

Examples

data(dat4h)

## maybe str(dat4h) ; plot(dat4h) ...

generate.biclust.model

*GO descriptors model using bi-clustering algorithm*

Description

This function is used to estimate the cluster memberships for omics data, based on GO ontology. Data are clustered based on bi-clustering algorithm from the blockcluster R package using default values. The user needs to specify the number of clusters for both axes.

Usage

generate.biclust.model(dataset, predictionFeature, parameters)

Arguments

dataset list of 2 objects, datasetURI:= character sring, code name of dataset, dataEntry:= data frame with 2 columns

predictionFeature

character string specifying which is the prediction feature in dataEntry

parameters list with parameter values for ontology and biclustering. 5 objects should be included, i.e. ’key’a character sring for gene/protein/etc names id (for dat1 ’UNIPROT’ is the right key), ’onto’ a character vector showing the ontology and sub.ontologies used ( c(’GO’,’MF’)), ’pvalCutoff’ a numeric value for hy- pergeometric p-values cutoff (e.g. 0.05), ’nclust’ a numeric vector indicating the number of clusters for GOs(x axis) and genes/proteins (y axis) (e.g. c(5,4)), ’FUN’ a string, R function to summarize vector’s groups (e.g. mean).

*generate.biclust.model* 7

Details

More details can be found in https://cran.r-project.org/web/packages/blockcluster/index.html.

Value

A List

rawModel A serialized GO descriptors object (class raw) giving the cluster memberships of proteins/genes in the data.

pmmlModel A pmml GO descriptors object - now empty

predictedFeatures

A character vector with names for the new descriptors

independentFeatures

A list with Ambit names for all genes/ proteins features included in the model

additionalInfo A data frame with all independent features included in the model and their dummy name in the model - here empty

Author(s)

Georgia Tsiliki

References

1. [blockcluster](https://cran.r-project.org/web/packages/blockcluster/index.html) 2.[GSEABase](http://bioconductor.org/packages/release/bioc/html/GSEABase.html) 3.[GOstats](https://www.bioconductor.org/packages/3.3/bioc/html/GOstats.html)

See Also

generate.hierar.model

Examples

##---- Should be DIRECTLY executable !! ---- ##-- ==> Define data, use random,

##--or do help(data=index) for the standard data sets. data("dat1")

predF<- list()

required.param<- list(key=¹UNIPROT¹,onto=c(¹GO¹,¹MF¹),pvalCutoff=0.05,nclust=c(3,2),FUN=¹mean¹) clust.memb<- generate.biclust.model(dat1,predF,required.param)

8 *generate.hierar.model*

generate.hierar.model *GO descriptors model using hierarchical clustering algorithm*

Description

This function is used to estimate the cluster memberships for omics data, based on GO ontology. Data are clustered based on hierarchical clustering algorithm from the vegan R package using de- fault values. The user needs to specify the number of clusters or the height of the dendrogram.

Usage

generate.hierar.model(dataset, predictionFeature, parameters)

Arguments

dataset list of 2 objects, datasetURI:= character sring, code name of dataset, dataEntry:= data frame with 2 columns

predictionFeature

character string specifying which is the prediction feature in dataEntry

parameters list with parameter values for ontology and hierarrchical clustering. 7 objects should be included, i.e. ’key’a character sring for gene/protein/etc names id (for dat1 ’UNIPROT’ is the right key), ’onto’ a character vector showing the ontology and sub.ontologies used ( c(’GO’,’MF’)), ’pvalCutoff’ a numeric value for hypergeometric p-values cutoff (e.g. 0.05), ’distMethod’ distance method (could be one of those provided via vegan R package), ’hclustMethod’ (could be one of those provided via vegan R package) , ’nORh’ either a numeric value or character giving number of clusters or a function to define height respectively, ’FUN’ a string, R function to summarize vector’s groups (e.g. mean).

Details

Hierarchical clustering algorithm implementation was taken from the vegan cran pckage https://cran.r- project.org/web/packages/vegan/index.html

Value

A List

rawModel A serialized GO descriptors object (class raw) giving the cluster memberships of proteins/genes in the data.

pmmlModel A pmml GO descriptors object - now empty

predictedFeatures

A character vector with names for the new descriptors

independentFeatures

A list with Ambit names for all genes/ proteins features included in the model

additionalInfo A data frame with all independent features included in the model and their dummy name in the model - here empty

*pred.descr* 9

Author(s)

Georgia Tsiliki

References

1. [vegan](https://cran.r-project.org/web/packages/vegan/index.html) 2.[GSEABase](http://bioconductor.org/packages/release/bioc/html/GSEABase.html) 3.[GOstats](https://www.bioconductor.org/packages/3.3/bioc/html/GOstats.html)

See Also

generate.biclust.model

Examples

##---- Should be DIRECTLY executable !! ---- ##-- ==> Define data, use random,

##-- or do help(data=index) for the standard data sets. data("dat4h")

predF<- list()

required.param<- list(key=¹UNIPROT¹,onto=c(¹GO¹,¹MF¹),pvalCutoff=0.05, distMethod=¹euclidean¹,hclustMethod=¹ward.D2¹,nORh=¹mean¹,FUN=¹mean¹)

clust.memb<- generate.hierar.model(dat4h,predF,required.param)

pred.descr *Predict GO descriptors*

Description

Produces GO descriptors based on the data and model supplied. Also a summary statistic needs to be specified.

Usage

pred.descr(dataset, rawModel, additionalInfo)

10 *read.in.json.for.pred*

Arguments

dataset Data for prediction. A list of two objects: datasetURI (a character string ), dataEntry (a data frame).

rawModel R model serialized (GO cluster memberships for proteins/genes).

additionalInfo Any additional information needed for rawModel. Here a data frame with one cell giving the function to summarize dataset based on rawModel (e.g. ’mean’).

Details

This function reads in the information provided by either generate.biclust.model() or generate.hierar.model() functions to convert their output in suitable form that can be easily read in JSON format.

Value

A data.frame with the new GO descriptors given the function’s parameters. Number of columns is the number of new GO descriptors, number of rows is the number of features used.

Author(s)

Georgia Tsiliki

References

[jsonlite](https://cran.r-project.org/web/packages/jsonlite/index.html)

Examples

##---- Should be DIRECTLY executable !! ---- ##-- ==> Define data, use random,

##--or do help(data=index) for the standard data sets.

data("dat1p")

data("dat1i")

data("dat1m")

pred.res<- pred.descr(dat1p, dat1m, dat1i)

read.in.json.for.pred *Read in function for json files for prediction*

Description

This function reads in a json data file and produces a list with independent features, GO clustering memberships raw model.

*read.in.json.for.pred* 11

Usage

read.in.json.for.pred(dataset, rawModel, additionalInfo)

Arguments

dataset Data for prediction. A list of two objects: datasetURI (a character string ), dataEntry (a data frame).

rawModel R model serialized (GO cluster memberships for proteins/genes).

additionalInfo Any additional information needed for rawModel. Here a data frame with one cell giving the function to summarize dataset based on rawModel (e.g. ’mean’).

Details

No further details required

Value

A List including:

x.mat data frame with independent variables (proteins/genes/ etc)

model R model (GO cluster memberships for proteins/genes)

additionalInfo Any additional information needed for rawModel. Here a data frame with one cell giving the function to summarize dataset based on rawModel (e.g. ’mean’).

Author(s)

Georgia Tsiliki

References

[jsonlite](https://cran.r-project.org/web/packages/jsonlite/index.html)

Examples

##---- Should be DIRECTLY executable !! ---- ##-- ==> Define data, use random,

##--or do help(data=index) for the standard data sets.

data("dat1p")

data("dat1i")

data("dat1m")

data.file<- read.in.json.for.pred(dat1p, dat1m, dat1i)

**Index**

Topic **datasetsh**

⇤

dat4h, 5 Topic **datasetsi**

⇤

dat1i, 3

Topic **datasetsm**

⇤

dat1m, 4 Topic **datasetsp**

⇤

dat1p, 5 Topic **datasets**

⇤

dat1, 3

##### Topic generateBiclustGOdescrModel

⇤

generate.biclust.model, 6

##### Topic generateHierarClustGOdescr-

⇤

##### Model

generate.hierar.model, 8 Topic **package**

⇤

GOdescrPred-package, 2 Topic **predlmjson**

⇤

pred.descr, 9 Topic **readjsonp**

⇤

read.in.json.for.pred, 10

dat1, 3 dat1i, 3 dat1m, 4 dat1p, 5 dat4h, 5

generate.biclust.model, 6, *9* generate.hierar.model, *7*, 8 GOdescrPred *(*GOdescrPred-package*)*, 2 GOdescrPred-package, 2

pred.descr, 9 read.in.json.for.pred, 10

12



## ANNEX II: GODESCRCALCULUS

eNanoMapper 604134 31 January 2016 ENM TUTORIALS Page **8** of **9**

**Type** Package

**Package ‘GOdescrCalculus’**

###### January 28, 2016

**Title** Creates GO descriptors

**Version** 1.0

**Date** 2015-04-16

**Author** Georgia Tsiliki

**Maintainer** Georgia Tsiliki [<gtsiliki@central.ntua.gr>](mailto:gtsiliki@central.ntua.gr)

Depends

RCurl, pmml, org.Hs.eg.db, GSEABase, GOstats, Rgraphviz, blockcluster, Matrix, vegan, jsonlite

Description

This package produces GO descriptors based on GO ontology and proteomics data supplied.

**License** GPL-2

### R topics documented:

GOdescrCalculus-package 2

dat1 3

dat1i1 3

dat1i2 4

dat1m1 5

dat1m2 5

dat1p 6

generate.descr.biclust 7

generate.descr.hierar 8

generate.param.model 9

read.in.json.for.pred 11

Index 13

1

2 *GOdescrCalculus-package*

GOdescrCalculus-package

*Produce GO descriptors given an omics data*

Description

This package expects an omics data (e.g. proteomics/genomics), and produces a set of GO de- scriptors given the data and the parameters supplied based on the data and the algorirthm selected; bi-clustering and hierarchical clustering algorithms are currently implemented. The clustering al- gorithm parameters can be stored as an R raw model and used for ’prediction’, i.e. to construct GO decsriptors for the same data.

Details

Package: GOdescrCalculus Type: Package

Version: 1.0

Date: 2015-04-16

License: GPL-2

Generate GO descriptors given an omics data. Important functions are generate.descr.biclust and generate.descr.hierar.

Author(s)

Georgia Tsiliki

Maintainer: Georgia Tsiliki [<gtsiliki@central.ntua.gr>](mailto:gtsiliki@central.ntua.gr)

References

1. [blockcluster](https://cran.r-project.org/web/packages/blockcluster/index.html) 2.[vegan](https://cran.r-project.org/web/packages/vegan/index.html) 3.[GSEABase](http://bioconductor.org/packages/release/bioc/html/GSEABase.html) 4.[GOstats](https://www.bioconductor.org/packages/3.3/bioc/html/GOstats.html)

Examples

data("dat1p") data("dat1m1") data("dat1i1")

res1<- read.in.json.for.pred(dat1p, dat1m1, dat1i1)

*dat1* 3

dat1 *A sample data object*

Description

The dataset for this test is a data frame

Usage

data("dat1")

Format

A list of two objects

**datasetURI** a character vector- ambit data set uri

**dataEntry** a data frame containing two columns: compound and values. Compound is a character vector with all compound anbit uris, and values is a data frame with all numberic values of the proteomics data set (compounds by features). Only dependent features are included.

Details

Please find more about the data set in $dataEntry at <http://pubs.acs.org/doi/abs/10.1021/nn406018q?journalCode=ancac3>

References

[Walkey et al., 2014](http://pubs.acs.org/doi/abs/10.1021/nn406018q?journalCode=ancac3)

Examples

data(dat1)

## maybe str(dat1) ; plot(dat1) ...

dat1i1 *Additional Information needed for biclustering algorithm*

Description

Additionali information, i.e. the names of the predicted new features as generated by biclustering algorithm

Usage

data("dat1i1")

4 *dat1i2*

Format

A list with one object ’predictedFeatures’including the names of the new descriptors

Details

Example of new descriptor names as generated by generate.param.model

References

There are no references

Examples

data(dat1i1)

## maybe str(dat1i1) ; plot(dat1i1) ...

dat1i2 *Additional Information needed for hierarchical clustering algorithm*

Description

Additionali information, i.e. the names of the predicted new features as generated by hierarchical clustering algorithm

Usage

data("dat1i2")

Format

A list with one object ’predictedFeatures’including the names of the new descriptors

Details

Example of new descriptor names as generated by generate.param.model

References

There are no references

Examples

data(dat1i2)

## maybe str(dat1i2) ; plot(dat1i2) ...

*dat1m1* 5

dat1m1 *Serialized list of parameters for biclustering algorithm*

Description

A character string for a serialized parameters list, i.e. a set of values needed from the biclusteruing algorithm in order to produce the GO descriptors. The list includes ’key’ (e.g. ’UNIPROT’), ’onto’ (e.g. c(’GO’,’MF’)), ’pvalCutoff’ for hypergeometric test (e.g. 0.05), ’nclust’ for the number of clusters in the x and y axis respectively (e.g c(4,2)), and ’FUN’ to specify the function used to summarize the omics data (e.g. ’mean’).

Usage

data("dat1m1")

Format

A character string

Details

Example set of parameters needed for generate.descr.biclust function

References

There are no references

Examples

data(dat1m1)

## maybe str(dat1m1) ; plot(dat1m1) ...

dat1m2 *Serialized list of parameters for hierarchical clustering algorithm*

Description

A character string for a serialized parameters list, i.e. a set of values needed from the biclusteruing algorithm in order to produce the GO descriptors. The list includes ’key’ (e.g. ’UNIPROT’), ’onto’ (e.g. c(’GO’,’MF’)), ’pvalCutoff’ for hypergeometric test (e.g. 0.05), ’distMethod’ (e.g. ’euclidean’ or other alternatives from the vegan package), ’hclustMethod’ (e.g. ’ward.D2’ or other alternatives from the vegan package), ’nORh’ th enumber of clusters in the data (e.g. 10), and ’FUN’ to specify the function used to summarize the omics data (e.g. ’mean’).

Usage

data("dat1m2")

6 *dat1p*

Format

A character string

Details

Example set of parameters needed for generate.descr.hierar function

References

There are no references

Examples

data(dat1m2)

## maybe str(dat1m2) ; plot(dat1m2) ...

dat1p *A sample data object*

Description

The dataset for this test is a data frame

Usage

data("dat1p")

Format

A list of two objects

**datasetURI** a character vector- ambit data set uri

**dataEntry** a data frame containing two columns: compound and values. Compound is a character vector with all compound anbit uris, and values is a proteomics data frame with all numberic values of the data set (compounds by features)

Details

Data set for prediction with dat1m

References

[Walkey et al., 2014](http://pubs.acs.org/doi/abs/10.1021/nn406018q?journalCode=ancac3)

Examples

data(dat1p)

## maybe str(dat1p) ; plot(dat1p) ...

*generate.descr.biclust* 7

generate.descr.biclust

*Produce GO descriptors using bi-clustering algorithm*

Description

This function is used to estimate GO descriptors for omics data, based on GO ontology. Data are clustered based on bi-clustering algorithm from the blockcluster R package using default values. The user needs to specify the number of clusters for each axes.

Usage

generate.descr.biclust(dataset, rawModel, additionalInfo)

Arguments

dataset Data for prediction. A list of two objects: datasetURI (a character string ), dataEntry (a data frame).

rawModel A serialized list of parameters for biclustering, as produced by generate.param.model

additionalInfo Any additional information needed for rawModel. A list with one objects called predictedFeatures which are character names of the new descriptors.

Details

More details can be found in https://cran.r-project.org/web/packages/blockcluster/index.html.

Value

A list giving the new GO descriptors given the function’s parameters. Number of columns are the number of new GO descriptors, number of rows is the number of features used. The new descriptors are given as data frames per feature.

Author(s)

Georgia Tsiliki

References

1. [blockcluster](https://cran.r-project.org/web/packages/blockcluster/index.html) 2.[GSEABase](http://bioconductor.org/packages/release/bioc/html/GSEABase.html) 3.[GOstats](https://www.bioconductor.org/packages/3.3/bioc/html/GOstats.html)

8 *generate.descr.hierar*

Examples

##---- Should be DIRECTLY executable !! ---- ##-- ==> Define data, use random,

##--or do help(data=index) for the standard data sets.

data("dat1p") data("dat1m1") data("dat1i1")

pred.res<- generate.descr.biclust(dat1p, dat1m1, dat1i1)

generate.descr.hierar *Produce GO descriptors using hierarchical clustering algorithm*

Description

This function is used to estimate GO descriptors for omics data, based on GO ontology. Data are clustered based on hierarchical clustering algorithm from vegan R package using default values. The user needs to specify the number of clusters, the distance matrix method, and the hierarchical clust method.

Usage

generate.descr.hierar(dataset, rawModel, additionalInfo)

Arguments

dataset Data for prediction. A list of two objects: datasetURI (a character string ), dataEntry (a data frame).

rawModel A serialized list of parameters for hierarchical clustering, as produced by gener- ate.param.model

additionalInfo Any additional information needed for rawModel. A list with one objects called predictedFeatures which are character names of the new descriptors.

Details

Hierarchical clustering algorithm implementation was taken from the vegan cran pckage https://cran.r- project.org/web/packages/vegan/index.html

Value

A list giving the new GO descriptors given the function’s parameters. Number of columns are the number of new GO descriptors, number of rows is the number of features used. The new descriptors are given as data frames per feature.

*generate.param.model* 9

Author(s)

Georgia Tsiliki

References

1. [vegan](https://cran.r-project.org/web/packages/vegan/index.html) 2.[GSEABase](http://bioconductor.org/packages/release/bioc/html/GSEABase.html) 3.[GOstats](https://www.bioconductor.org/packages/3.3/bioc/html/GOstats.html)

Examples

##---- Should be DIRECTLY executable !! ---- ##-- ==> Define data, use random,

##--or do help(data=index) for the standard data sets.

data("dat1p") data("dat1m2") data("dat1i2")

pred.res<- generate.descr.hierar(dat1p, dat1m2, dat1i2)

generate.param.model *A ’training’ function to produce a serialized list for the parameters needed to produce GO descriptors*

Description

This function is used to pass on all the neccessary information to generate.descr.biclust and gener- ate.descr.hierar functions.

Usage

generate.param.model(dataset, predictionFeature, parameters)

Arguments

dataset list of 2 objects, datasetURI:= character sring, code name of dataset, dataEntry:= data frame with 2 columns

predictionFeature

character string specifying which is the prediction feature in dataEntry, here empty

parameters list with parameter values for ontology and biclustering. 5 or 7 objects should be included depending on whether we then intend to use generate.descr.biclust or generate.descr.hierar function, respectively. In the first case: ’key’a character sring for gene/protein/etc names id (for dat1 ’UNIPROT’ is the right key), ’onto’ a character vector showing the ontology and sub.ontologies used ( c(’GO’,’MF’)),

10 *generate.param.model*

’pvalCutoff’ a numeric value for hypergeometric p-values cutoff (e.g. 0.05), ’nclust’ a numeric vector indicating the number of clusters for GOs(x axis) and genes/proteins (y axis) (e.g. c(5,4)), ’FUN’ a string, R function to sum- marize vector’s groups (e.g. mean). In the second case: ’key’a character sring for gene/protein/etc names id (for dat1 ’UNIPROT’ is the right key), ’onto’ a character vector showing the ontology and sub.ontologies used ( c(’GO’,’MF’)), ’pvalCutoff’ a numeric value for hypergeometric p-values cutoff (e.g. 0.05), ’distMethod’ distance method (could be one of those provided via vegan R package), ’hclustMethod’ (could be one of those provided via vegan R pack- age) , ’nORh’ either a numeric value or character giving number of clusters or a function to define height respectively, ’FUN’ a string, R function to summarize vector’s groups (e.g. mean).

Details

Parameters are structured in a suitable format so that they can be read in by other functions of the package.

Value

rawModel A serialized object of the parameters list supplied. pmmlModel A pmml GO descriptors object - now empty independentFeatures

A list with Ambit names for all genes/ proteins features included in the model

predictedFeatures

A character vector with dummy names for the GO descriptors that will be pro- duced by functions generate.descr.biclust or generate.descr.hierar

additionalInfo A list with one objects called predictedFeatures which are character names of the new descriptors

Author(s)

Georgia Tsiliki

References

No references for this function.

Examples

##---- Should be DIRECTLY executable !! ---- ##-- ==> Define data, use random,

##--or do help(data=index) for the standard data sets. data("dat1")

predF<- list()

required.param<- list(key=¹UNIPROT¹,onto=c(¹GO¹,¹MF¹),pvalCutoff=0.05,nclust=c(3,2),FUN=¹mean¹)

*read.in.json.for.pred* 11

params1<- generate.param.model(dat1,predF,required.param)

read.in.json.for.pred *Read in function for json files for prediction*

Description

This function reads in a json data file and produces a list with independent features, parameters list for GO clustering saved as raw model

Usage

read.in.json.for.pred(dataset, rawModel, additionalInfo)

Arguments

dataset Data for prediction. A list of two objects: datasetURI (a character string ), dataEntry (a data frame) which should include an omics data set (e.g. pro- teomics/genomics/etc).

rawModel Raw model for prediction. Here a seerilized list of parameters to be used by genrate.descr.biclust or generate.descr.hierar functions.

additionalInfo A list with one objects called predictedFeatures which are character names of the new descriptors.

Details

No further details required

Value

A List including:

x.mat data frame with independent variables values (proteins/genes/ etc)

model R model (a list of parameters to be used by genrate.descr.biclust or generate.descr.hierar functions)

additionalInfo Any additional information needed for rawModel. Here a list with two ob- jects, data frame giving the Ambit names of the independent features included in x.mat, and the predictedFeatures as given in the input.

Author(s)

Georgia Tsiliki

12 *read.in.json.for.pred*

References

[jsonlite](https://cran.r-project.org/web/packages/jsonlite/index.html)

Examples

##---- Should be DIRECTLY executable !! ---- ##-- ==> Define data, use random,

##--or do help(data=index) for the standard data sets.

data("dat1p") data("dat1m1") data("dat1i1")

res1<- read.in.json.for.pred(dat1p, dat1m1, dat1i1)

**Index**

Topic **datasetsi1**

⇤

dat1i1, 3 Topic **datasetsi2**

⇤

dat1i2, 4

Topic **datasetsm1**

⇤

dat1m1, 5

Topic **datasetsm2**

⇤

dat1m2, 5 Topic **datasetsp**

⇤

dat1p, 6 Topic **datasets**

⇤

dat1, 3

Topic **generateBiclust**

⇤

generate.descr.biclust, 7 Topic **generateHierar**

⇤

generate.descr.hierar, 8 Topic **generateParamModel**

⇤

generate.param.model, 9 Topic **package**

⇤

GOdescrCalculus-package, 2 Topic **readjsonp**

⇤

read.in.json.for.pred, 11

dat1, 3 dat1i1, 3 dat1i2, 4 dat1m1, 5 dat1m2, 5 dat1p, 6

generate.descr.biclust, 7 generate.descr.hierar, 8 generate.param.model, 9 GOdescrCalculus

*(*GOdescrCalculus-package*)*, 2

GOdescrCalculus-package, 2 read.in.json.for.pred, 11

13



# 6. KEYWORDS

Omics data, Gene Ontology, hierarchical clustering, bi-clustering, integration

eNanoMapper 604134 31 January 2016 ENM TUTORIALS Page **9** of **9**