Package 'GeoDE'

February 19, 2015

Type Package

Title A geometrical Approach to Differential expression and gene-set enrichment

Version 1.0

Date 2014-06-06

Author Neil R. Clark and Avi Ma'ayan

Maintainer Neil Clark <neil.clark@mssm.edu>

Description Given expression data this package calculate a multivariate geometrical characterization of the differential expression and can also perform gene-set enrichment.

License GPL-2

Depends R (>= 2.10), Matrix, MASS

NeedsCompilation no

Repository CRAN

Date/Publication 2014-07-31 07:17:53

R topics documented:

GeoDE-package 2
AllGMTfiles 5
chdirAnalysis 6
example_expression_data 8
example_gammas 9
example_sampleclass 10
gmt 10
multigmtPAEAAnalysis 11
PAEAAnalysis 13

Index **16**

GeoDE-package *Differential Expression and Enrichment Analysis with (Geo)metrical(D)ifferential(E)expression.*

Description

This package contains functions for performing multivariate analysis of geneome-wide expression data and also enrichment analysis.

Details

Package:	GeoDE
Type:	Package
Version:	1.0
Date:	2014-06-06
License:	GPL-2

Given gene expression data from two classes (e.g. controll verses perturbed samples) with biological replicates in each class, this package can be used to extract the most significant genes and gene-sets.

Differential expression is characterised with a single direction in expression space, which can be interpreted to extract the most signiicant genes: this is achieved with the `chdirAnalysis` function.

Once the characeristic direction has been calculated gene-set enrichmnet can be evaluated using the `PAEAAnalysis` function. The user is free to use any library of gene-sets, however, included in this package is a broad range of gene-set libraries listed below:

BioCarta_pathways.gmt

Cancer_Cell_Line_Encyclopedia.gmt

ChEA.gmt

Chromosome_location.gmt

CORUM.gmt.gmt

GeneOntology_BP.gmt

GeneOntology_CC.gmt

GeneOntology_MF.gmt

GeneSigDB.gmt

Genome_Browser_PWMs.gmt

HMDB_Metabolites.gmt

Human_Gene_Atlas.gmt

KEA.gmt

KEGG_pathways.gmt

MGI_MP_top3.gmt

MGI_MP_top4.gmt

microRNA.gmt

Mouse_Gene_Atlas.gmt

NCI60.gmt

NURSA-IPMS.gmt

OMIM_disease_genes.gmt

OMIM_Expanded.gmt

Pfam-InterPro-domains.gmt

PPI_Hub_Proteins.gmt

Reactome_pathways.gmt

TF_PPIs.gmt

VirusMINT.gmt

WikiPathways_pathways.gmt

Author(s)

Author: Neil Clark and Avi Ma'ayan

Maintainer: Neil R. Clark <neil.clark@mssm.edu>

References

Clark, Neil R., et al. "The characteristic direction: a geometrical approach to identify differentially expressed genes." BMC bioinformatics 15.1 (2014): 79.

Examples

```
###################################
#
# An example characteristic direction analysis
#
###################################

# Load the example data

data(example_expression_data)
data(example_sampleclass)
data(example_gammas)

# Examine the expression data
head(example_expression_data)

# Examine the corresponding sample class factor
example_sampleclass
```

```
# Run the analysis
chdir_analysis_example <- chdirAnalysis(example_expression_data,example_sampleclass,example_gammas
,CalculateSig=TRUE,nnull=10)

# Examine the results with the first value of the shrinkage parameter (gamma)

# show the first few of the most important genes.

lapply(chdir_analysis_example$results, function(x) x[1:10])

# We can also extract the results of the \code{chdirSig} function
# for example chdir_analysis_example$chdirprops[[1]] gives the whole
# characteristic direction vector for each value of gamma:

lapply(chdir_analysis_example$chdirprops[[1]],head)

# and the estimated number of significant genes can be recovered with

chdir_analysis_example$chdirprops$number_sig_genes

##################################
#
# An example PAEA analysis
#
##################################
# Load the expression data

data(example_expression_data)
data(example_sampleclass)
data(example_gammas)

#load a gmt file
data(GeneOntology_BP.gmt)

# Run the characteristic direction analysis
chdir_analysis_example <- chdirAnalysis(example_expression_data,example_sampleclass,example_gammas
,CalculateSig=FALSE)

# Run the PAEA analysis

PAEAtest <- PAEAAnalysis(chdir_analysis_example$chdirprops, gmt[1:100], example_gammas)

# Examine the p values

PAEAtest$p_values

# Examine the principal angles

PAEAtest$principal_angles

##################################
#
# An example multigmtPAEA analysis
```

```
#
##################################
# Load the expression data

data(example_expression_data)
data(example_sampleclass)
data(example_gammas)

#load GMT file names
data(AllGMTfiles)

# Run the characteristic direction analysis
chdir_analysis_example <- chdirAnalysis(example_expression_data,example_sampleclass,example_gammas
,CalculateSig=FALSE)

# Run the PAEA analysis over the first two GMT files in the library

multiPAEAtest <- multigmtPAEAAnalysis(chdir_analysis_example$chdirprops, AllGMTfiles[2:3],
example_gammas)

# To run on all the gmt files

#multiPAEAtestAll <- multigmtPAEAAnalysis(chdir_analysis_example$chdirprops, gammas=example_gammas)
```

`AllGMTfiles` *The gmt files included in GeoDE*

Description

This list contains the names of all the GMT files included in GeoDE.

Usage

```
data(AllGMTfiles)
```

Format

The format is: chr "AllGMTfiles"

Source

Chen, Edward Y., et al. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool." BMC bioinformatics 14.1 (2013): 128.

Examples

```
# load the GMT file data

data(AllGMTfiles)

# load the first gmt file in the list

data(list=AllGMTfiles[[1]])


## maybe str(AllGMTfiles) ; plot(AllGMTfiles) ...
```

---

chdirAnalysis *A Function to Perform Characteristic Direction Analysis.*

Description

This function takes genome-wide expression data as input and returns the characteristic direction - a unit vector in expression space which characterizes the differential expression. Also produced are 2D projections of the data and the characteristic direction. Optionaly this function will produce an evaluation of the signifcance of the result.

Usage

```
chdirAnalysis(datain, sampleclass, gammas = list(1), nnull = 10, CalculateSig = FALSE)
```

Arguments

datain	A data frame containing the common gene names (first) and the expression profiles.
sampleclass	A factor with levels "\'1\'" and "\'2\'" indicating the class of the samples in the data. For each column of the data frame (excluding the gene names) this factor should contain an entry indicating the class from which the sample derives (e.g. controll sample,"\'1\'"", or perturbed sample, "\'2\'".)
gammas	A set of values for the shrinkage parameter. The default value is gammas=c(1.0).
nnull	If a significance estimate is to be made the number of random directions used is set with this value.
CalculateSig	A logical value which determines whether a significance estimate is to be calculated.

Value

chdirprops	This is a list of properties of the characteristic direction. The first element is chdirprops$chdir, the vector in expresion space whose direction characerises the differential expression. The second element, chdirprops$pca2d, is the 2D PCA projection of the data. The third element, chdirprops$chdir_pca2d is the 2D projection of the characteristic direction vector into PCA space.

results	A list with an element corresponding to each of the shrinkage parameter vlues giing the sorted list of genes and their characteristic direction coefficients. If a significance estimate has been made then only the significant genes are returned here.
plots	for each value of the shrinkage parameter a 2D PCA projection of the data and the characteristic direction is generated. If a significance estimate is made then the significance curve is also produced. A positive peak indicates that the two classes of samples are significantly different.

Author(s)

Neil R Clark and Avi Ma'ayan

References

Clark, Neil R., et al. "The characteristic direction: a geometrical approach to identify differentially expressed genes." BMC bioinformatics 15.1 (2014): 79.

See Also

chdirSig

Examples

```
##---- Should be DIRECTLY executable !! ----
##-- ==>  Define data, use random,
##--or do  help(data=index)  for the standard data sets.

###################################
#
# An example characteristic direction analysis
#
###################################

# Load the example data

data(example_expression_data)
data(example_sampleclass)
data(example_gammas)

# Examine the expression data
head(example_expression_data)

# Examine the corresponding sample class factor
example_sampleclass

# Run the analysis
chdir_analysis_example <- chdirAnalysis(example_expression_data,example_sampleclass,example_gammas
,CalculateSig=TRUE,nnull=10)

# Examine the results with the first value of the shrinkage parameter (gamma)
```

```
# show the first few of the most important genes.

lapply(chdir_analysis_example$results, function(x) x[1:10])

# We can also extract the results of the \code{chdirSig} function
# for example chdir_analysis_example$chdirprops[[1]] gives the whole
# characteristic direction vector for each value of gamma:

lapply(chdir_analysis_example$chdirprops[[1]],head)

# and the estimated number of significant genes can be recovered with

chdir_analysis_example$chdirprops$number_sig_genes


## The function is currently defined as
function (datain, sampleclass, gammas = list(1), nnull = 3, CalculateSig = FALSE)
{
    if (length(sampleclass) != (length(datain) - 1))
        stop("number of elements in sampleclass is inconsistent with input data")
    if (!is.data.frame(datain))
        stop("Input data is not in the form of a data frame")
    if (FALSE %in% (c("1", "2") %in% levels(sampleclass)))
        stop("sample class does not include '1' and '2'")
    if (length(datain[sampleclass == 1]) < 2)
        stop("too few controll samples")
    if (length(datain[sampleclass == 2]) < 2)
        stop("too few samples")
    chdirresults <- chdirSig(datain, sampleclass, gammas, nnull = nnull,
        CalculateSig = CalculateSig)
    chdirplots(chdirresults, sampleclass, gammas, CalculateSig)
    outAll <- lapply(chdirresults[[1]], function(x) {
        x[sort.list(x^2, decreasing = TRUE), ]
    })
    if (CalculateSig) {
        outSig <- mapply(function(x, ns) {
            x[sort.list(x^2, decreasing = TRUE)[1:ns], ]
        }, chdirresults[[1]], chdirresults[[6]])
        list(chdirprops = chdirresults, results = outSig)
    }
    else {
        list(chdirprops = chdirresults, results = outAll)
    }
  }
```

---

example_expression_data

Example Expression Data

Description

This is the data used in the examples.

Usage

```
data(example_expression_data)
```

Format

The format is: chr "example_expression_data"

Examples

```
data(example_expression_data)
## maybe str(example_expression_data) ; plot(example_expression_data) ...
```

---

example_gammas *Example Gamma Value*

Description

This is the list of gamma values used in the examples.

Usage

```
data(example_gammas)
```

Format

The format is: chr "example_gammas"

Examples

```
data(example_gammas)
## maybe str(example_gammas) ; plot(example_gammas) ...
```

example_sampleclass *Example Sample Class Factor*

Description

This is the factor variable that defined the classes of the samples in the examples.

Usage

```
data(example_sampleclass)
```

Format

The format is: chr "example_sampleclass"

Examples

```
data(example_sampleclass)
## maybe str(example_sampleclass) ; plot(example_sampleclass) ...
```

gmt *The currently loaded gmtfile*

Description

When a GMT file is loaded e.g. by `data("GeneOntology_BP.gmt")` the gmt file is stored in the variable `gmt`

Usage

```
gmt
```

Format

The format is alist of character variables defining gene sets and their labels (first element).

Examples

```
data("GeneOntology_BP.gmt")
## maybe str(gmt) ; plot(gmt) ...
```

multigmtPAEAAnalysis *A Function to Evaluate Gene-Set Enrichment Using PAEA Over Multiple GMT Files.*

Description

This is a wrapper function for PAEAAnalysis which evaluates the enrichmnet of gene sets in expression data using the PAEA method. A characteristic direction (the result of the function chdirAnalysis and a Gene Matrix Transposed (GMT) file, which is a set of subsets of genes whose enrichmnet is evaluated, and returns a prioritized list of the gene sets.

Ths function takes multiple GMT files as input and output the results to Tab Separated Value files.

Usage

```
multigmtPAEAAnalysis(chdirresults, gmtfiles=AllGMTfiles, gammas = c(1),
casesensitive = FALSE, showprogress=TRUE)
```

Arguments

chdirresults	This input is the first part of the output from chdirAnalysis, which has the names $chdirprops. This contains the characteristic direction which forms the basis for the enrichmnet analysis.
gmtfiles	This is a list of names of GMT files(A data set composed of a list of lists of genes) over which the enrichment analysis is calculated. By default all included GMT files are used (not that it is neccessary to use data(AllGMTfiles) to load the names first.)
gammas	The list of shrinkage parameter values as used in the calculation of the characteristic direction.
casesensitive	A logical variable which determines whether the gene comparisons should be case sensitive.
showprogress	show a progress bar.

Value

p_values-GMTfile.txt
: The results of the enerichmnet for each GMT file is saved to file in the current working directory.

Author(s)

Neil R. Clark and Avi Ma'ayan

References

Clark, Neil R., et al. "The characteristic direction: a geometrical approach to identify differentially expressed genes." BMC bioinformatics 15.1 (2014): 79.

Examples

```
##---- Should be DIRECTLY executable !! ----
##-- ==>  Define data, use random,
##--or do  help(data=index)  for the standard data sets.


###################################
#
# An example multigmtPAEA analysis
#
###################################
# Load the expression data

data(example_expression_data)
data(example_sampleclass)
data(example_gammas)

#load GMT file names
data(AllGMTfiles)

# Run the characteristic direction analysis
chdir_analysis_example <- chdirAnalysis(example_expression_data,example_sampleclass,
example_gammas,CalculateSig=FALSE)

# Run the PAEA analysis over the first two GMT files in the library

multiPAEAtest <- multigmtPAEAAnalysis(chdir_analysis_example$chdirprops,
AllGMTfiles[2:3], example_gammas)

# To run on all the gmt files

#multiPAEAtestAll <- multigmtPAEAAnalysis(chdir_analysis_example$chdirprops, gammas=example_gammas)

## The function is currently defined as
function (chdirresults, gmtfile, gammas = c(1), casesensitive = FALSE)
{
    gmtlinenames <- lapply(gmtfile, function(x) x[[1]])
    gmtlines <- lapply(gmtfile, function(x) x[-1])
    PAEAresults <- lapply(gmtlines, function(x) PAEA(chdirresults[[1]],
        x, casesensitive = casesensitive))
    gammalabels <- unlist(lapply(gammas, function(x) paste("gamma=",
        x)))
    pvalues <- lapply(PAEAresults, function(x) x[[2]])
    pvalues <- matrix(unlist(pvalues), ncol = length(gmtlines),
        dimnames = list(gammalabels, gmtlinenames))
    pavalues <- lapply(PAEAresults, function(x) x[[1]])
    pavalues <- matrix(unlist(pvalues), ncol = length(gmtlines),
        dimnames = list(gammalabels, gmtlinenames))
    gmtp <- sort.list(pvalues[1, ])
    list(t(pvalues[, gmtp]), t(pavalues[, gmtp]))
  }
```

PAEAAnalysis *A Function to Evaluate Gene-Set Enrichment Using PAEA.*

Description

This function evaluates the enrichmnet of gene sets in expression data using the PAEA method. A characteristic direction (the result of the function `chdirAnalysis` and a Gene Matrix Transposed (GMT) file, which is a set of subsets of genes whose enrichmnet is evaluated, and returns a prioritized list of the gene sets.

Usage

```
PAEAAnalysis(chdirresults, gmtfile, gammas = c(1), casesensitive = FALSE,
showprogress=TRUE)
```

Arguments

`chdirresults`	This input is the first part of the output from `chdirAnalysis`, which has the names `$chdirprops`. This contains the characteristic direction which forms the basis for the enrichmnet analysis.
`gmtfile`	A data set composed of a list of lists of genes (a GMT file); each list of genes has, as its first element, a label for the gene list that follows in that line. For example, if the genes are members of a pathway then the list will be of the form: pathway_name Gene1 Gene2 ... GeneN.
`gammas`	The list of shrinkage parameter values as used in the calculation of the characteristic direction.
`casesensitive`	A logical variable which determines whether the gene comparisons should be case sensitive.
`showprogress`	show a progress bar.

Value

`p_values`	A matrix of p values with gene sets down the rows and values of the shrinkage parameter values across the columns.
`principal_angles`	A matrix with rows corresponding to gene sets, columns corresponding to values of the shrinkage parameter and elements equal to the principal angle between the gene set subspace and the characteristic direction.

Author(s)

Neil R. Clark and Avi Ma'ayan

References

Clark, Neil R., et al. "The characteristic direction: a geometrical approach to identify differentially expressed genes." BMC bioinformatics 15.1 (2014): 79.

Examples

```
##---- Should be DIRECTLY executable !! ----
##-- ==>  Define data, use random,
##--or do  help(data=index)  for the standard data sets.


###################################
#
# An example PAEA analysis
#
###################################
# Load the expression data

data(example_expression_data)
data(example_sampleclass)
data(example_gammas)

#load a gmt file
data(GeneOntology_BP.gmt)

# Run the characteristic direction analysis
chdir_analysis_example <- chdirAnalysis(example_expression_data,example_sampleclass,
example_gammas,CalculateSig=FALSE)

# Run the PAEA analysis

PAEAtest <- PAEAAnalysis(chdir_analysis_example$chdirprops, gmt[1:100], example_gammas)

# Examine the p values

PAEAtest$p_values

# Examine the principal angles

PAEAtest$principal_angles


## The function is currently defined as
function (chdirresults, gmtfile, gammas = c(1), casesensitive = FALSE)
{
    gmtlinenames <- lapply(gmtfile, function(x) x[[1]])
    gmtlines <- lapply(gmtfile, function(x) x[-1])
    PAEAresults <- lapply(gmtlines, function(x) PAEA(chdirresults[[1]],
        x, casesensitive = casesensitive))
    gammalabels <- unlist(lapply(gammas, function(x) paste("gamma=",
        x)))
    pvalues <- lapply(PAEAresults, function(x) x[[2]])
    pvalues <- matrix(unlist(pvalues), ncol = length(gmtlines),
        dimnames = list(gammalabels, gmtlinenames))
    pavalues <- lapply(PAEAresults, function(x) x[[1]])
    pavalues <- matrix(unlist(pvalues), ncol = length(gmtlines),
        dimnames = list(gammalabels, gmtlinenames))
```

```
  gmtp <- sort.list(pvalues[1, ])
  list(t(pvalues[, gmtp]), t(pavalues[, gmtp]))
}
```

Index

∗Topic **DE**
chdirAnalysis, 6
∗Topic **characteristic**
GeoDE-package, 2
∗Topic **datasets**
AllGMTfiles, 5
example_expression_data, 8
example_gammas, 9
example_sampleclass, 10
gmt, 10
∗Topic **differential**
chdirAnalysis, 6
GeoDE-package, 2
∗Topic **direction**
GeoDE-package, 2
∗Topic **enrichment**
multigmtPAEAAnalysis, 11
PAEAAnalysis, 13
∗Topic **enrichmnet**
GeoDE-package, 2
∗Topic **expression**
chdirAnalysis, 6
GeoDE-package, 2
∗Topic **genes**
chdirAnalysis, 6
∗Topic **gene**
GeoDE-package, 2
∗Topic **multivariate**
GeoDE-package, 2
multigmtPAEAAnalysis, 11
PAEAAnalysis, 13
∗Topic **package**
GeoDE-package, 2

AllGMTfiles, 5

chdirAnalysis, 6

example_expression_data, 8
example_gammas, 9
example_sampleclass, 10

GeoDE *(*GeoDE-package*)*, 2
GeoDE-package, 2
gmt, 10

multigmtPAEAAnalysis, 11

PAEAAnalysis, 13