Package 'GeneExpressionSignature'

September 11, 2025

Title Gene Expression Signature based Similarity Metric

Version 1.54.0

Date 2012-12-16

Description This package gives the implementations of the gene expression signature and its distance to each. Gene expression signature is represented as a list of genes whose expression is correlated with a biological state of interest. And its distance is defined using a nonparametric, rank-based pattern-matching strategy based on the Kolmogorov-Smirnov statistic. Gene expression signature and its distance can be used to detect similarities among the signatures of drugs, diseases, and biological states of interest.

Depends R (>= 4.0)

License GPL-2

URL https://github.com/yiluheihei/GeneExpressionSignature

BugReports https://github.com/yiluheihei/GeneExpressionSignature/issues/

LazyLoad yes

biocViews GeneExpression

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.1

Imports Biobase, stats, methods

Suggests apcluster, GEOquery, knitr, rmarkdown, BiocStyle

VignetteBuilder knitr

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

git_branch RELEASE_3_21

git_last_commit 3281b56

git_last_commit_date 2025-04-15

Repository Bioconductor 3.21

Date/Publication 2025-09-10

Author Yang Cao [aut, cre],
Fei Li [ctb],
Lu Han [ctb]

Maintainer Yang Cao <yiluheihei@gmail.com>

Contents

exampleSet 2
GeneExpressionSignature 3
getRLs 3
RankMerging 4
ScoreGSEA 5
ScorePGSEA 7
SignatureDistance 8

Index **10**

`exampleSet` *sample data, a subset of the C-MAP*

Description

sample data, a subset of the C-MAP as , which is a collection of 50 genome-wide transcriptional expression data from cultured human cells treated with 15 different small molecules

Format

A ExpressionSet: assay data represents the 50 genome-wide transcriptional expression data, phenotypic data describes 15 different small molecules corresponds to the expression data (assay data).

References

http://www.sciencemag.org/content/313/5795/1929.short Lamb et al., The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease, science 2006

GeneExpressionSignature *Gene Expression Signature based Similarity Metric*

Description

This package gives the implementations of the gene expression signature and its distance to each. Gene expression signature is represented as a list of genes whose expression is correlated with a biological state of interest. And its distance is defined using a nonparametric, rank-based pattern-matching strategy based on the Kolmogorov-Smirnov statistic. Gene expression signature and its distance can be used to detect similarities among the signatures of drugs, diseases, and biological states of interest.

getRLs *Convert gene expression profiles to a ranked list*

Description

Sorting the micro-array probe-set identifiers according to the differential expression values with respect to the untreated hybridization to obtain a ranked list. Gene-expression profiles in are represented in a nonparametric fashion.

Usage

```
getRLs(control, treatment)
```

Arguments

control	a matrix, including the vehicle control gene expression profiles corresponding to the treatment gene expression profiles.
treatment	a matrix, is composed of gene expression profiles.

Details

The genes on the array are rank-ordered according to their differential expression relative to the control. First, control and treatment values less than a primary threshold value (quartile) were set to that threshold value. Finally, probe sets were ranked in descending order of d, where d is the ratio of the corresponding treatment-to-control values. For probe sets where d=1, a lower threshold was applied to the original difference values and a new treatment to control ratio (d') calculated. These probe sets were then sub-sorted in descending order of d.

Value

A matrix is composed of ranked lists, a ranked list represents the corresponding gene expression profiles.

Examples

```
if (require(GEOquery)){
  # treatment gene-expression profiles
  file1 <- system.file(
    "extdata/GSM118720.soft",
    package = "GeneExpressionSignature"
  )
  GSM118720 <- getGEO(filename = file1)

  # control gene-expression profiles
  file2 <- system.file(
    "extdata/GSM118721.soft",
    package = "GeneExpressionSignature"
  )
  GSM118721 <- getGEO(filename = file2)

  # data ranking according to the different expression values
  control <- as.matrix(as.numeric(Table(GSM118721)[, 2]))
  treatment <- as.matrix(as.numeric(Table(GSM118720)[, 2]))
  ranked_list <- getRLs(control, treatment)
}
```

RankMerging *Merging the ranker lists with the same labels of the biological states into a single list with the Iorio's method*

Description

Merging the assay data according to phenotypic data of the input ExpressionSet. Each group of the ranked lists with the same phenotypic data is aggregated into a single list, return it as an ExpressionSet object.

Usage

```
RankMerging(
  exprSet,
  MergingDistance = c("Spearman", "Kendall"),
  weighted = TRUE
)
```

Arguments

exprSet	an ExpressionSet object, each column of assay data represents a ranked list obtained by preprocessing the corresponding gene expression profile, and phenotypic data represents the short description (characteristics of gene expression profile, such as the drug type, the disease state) about the assay data.
MergingDistance	distance to be used which "measures" the similarity of ordered lists, the default is "Spearman"

`weighted` there are tow rank merging approaches for two cases: if `weighted = FALSE`, all ranked list with the same biological state are treated equally important, a simple but useful method average ranking technique is selected; otherwise, `weighted = TRUE`, each individual ranked lists has its own ranked weights, this takes the iterative rank-aggregating algorithm, default is `TRUE`.

Details

The krubor function is used in the aggregating procedure. And the following methods are used in the implementation: a measure of the distance between two ranked lists (Spearman's Footrule), a method to merge two or more ranked lists the (Borda Merging Method), and a algorithm to obtain a single ranked list from a set of them in a hierarchical way (the Kruskal Algorithm). If choose Kendall as distance, the effectiveness of this function is certainly limited by the size of the merging problem.

Value

a `Biobase::ExpressionSet` object.

See Also

`SignatureDistance()`

Examples

```
# load the sample expressionSet
data(exampleSet)

# Merging each group of the ranked lists in the exampleSet with the same
# phenotypic data into a single PRL
MergingSet <- RankMerging(exampleSet, "Spearman", weighted = TRUE)
```

`ScoreGSEA` *Compute pairwise distances between samples with method in package GSEA*

Description

Compute pairwise distances between sample according to their (Prototype Ranked List) PRL, a `N x N` distance matrix is generated by calling this function, `N` is the length of PRL.

Usage

```
ScoreGSEA(
  MergingSet,
  SignatureLength,
  ScoringDistance = c("avg", "max"),
  p.value = FALSE
)
```

Arguments

MergingSet	an Biobase::ExpressionSet object. The assay data represents the PRLs of the samples, each column represents one PRL. The number of sample of this argument must be greater than 1, otherwise, this function is not meaningful.
SignatureLength	the length of "gene signature". In order to compute pairwise distances among samples, genes lists are ranked according to the gene expression ratio (fold change). And the "gene signature" includes the most up-regulated genes (near the top of the list) and the most down-regulated genes (near the bottom of the list).
ScoringDistance	the distance measurements between PRLs: the Average Enrichment Score Distance (:avg"), and the Maximum Enrichment Score Distance ("max").
p.value	logical, if TRUE return a matrix of p.values of the distance matrix, default FALSE.

Details

Once the PRL obtained for each sample, the distances between samples are calculated base on gene signature, including the expression of genes that seemed to consistently vary in response to the across different experimental conditions (e.g., different cell lines and different dosages). We take two distance measurements between PRLs: the Average Enrichment-Score Distance `Davg = (TES{x,y} + TES{y,x}) / 2`, and the Maximum Enrichment-Score Distance `Dmax = Min(TES{x,y}, TES{y,x}) / 2`.The avg is more stringent than max, where max is more sensitive to weak similarities, with lower precision but large recall.

Value

an distance-matrix, the max distance is more sensitive to weak similarities, providing a lower precision but a larger recall.If `p.value` is set to TRUE, then a list is returned that consists of the distance matrix as well as their p.values, otherwise, without p.values in the result.

See Also

ScorePGSEA(),SignatureDistance()

Examples

```
# load the sample expressionSet
data(exampleSet)
# Merging each group of the ranked lists in the exampleSet with the same
# phenotypic data into a single PRL
MergingSet <- RankMerging(exampleSet,"Spearman")
# get the distance matrix
ds <- ScoreGSEA(MergingSet, 250, "avg")
```

ScorePGSEA *Compute pairwise distances between samples with method in package PGSEA*

Description

Compute pairwise distances between sample according to their (Prototype Ranked List) PRL, get a `N x N` distance matrix is generated by calling this function , `N` is the length of PRL.

Usage

```
ScorePGSEA(
  MergingSet,
  SignatureLength,
  ScoringDistance = c("avg", "max"),
  p.value = FALSE
)
```

Arguments

`MergingSet`	an `Biobase::ExpressionSet` object. The assay data represents the PRLs of the samples, each column represents one PRL. The number of sample must be greater than 1, otherwise, this function is not meaningful.
`SignatureLength`	the length of "gene signature". In order to compute pairwise distances among samples, genes lists are ranked according to the gene expression ratio (fold change). And the "gene signature" includes the most up-regulated genes (near the top of the list) and the most down-regulated genes (near the bottom of the list).
`ScoringDistance`	the distance measurements between PRLs: the Average Enrichment Score Distance ("avg"), or the Maximum Enrichment Score Distance ("max").
`p.value`	logical, if `TRUE` return a matrix of `p.values` of the distance matrix, default `FALSE`.

Value

an distance-matrix, the max distance is more sensitive to weak similarities, providing a lower precision but a larger recall.If `p.value` is set to `TRUE`, then a list is returned that consists of the distance matrix as well as their p.values, otherwise, without p.values in the result.

See Also

`ScoreGSEA()`,`SignatureDistance()`

Examples

```
# load the sample expressionSet
data(exampleSet)
# Merging each group of the ranked lists in the exampleSet with the same
# phenotypic data into a single PRL
MergingSet <- RankMerging(exampleSet,"Spearman")
# get the distance matrix
ds <- ScorePGSEA(MergingSet,250, ScoringDistance="avg")
```

`SignatureDistance` *Compute pairwise distances comprehensively*

Description

This function integrated the function for rank merging and distance scoring, we can do the rank merging and distance scoring simply with it.

Usage

```
SignatureDistance(
  exprSet,
  SignatureLength,
  MergingDistance = c("Spearman", "Kendall"),
  ScoringMethod = c("GSEA", "PGSEA"),
  ScoringDistance = c("avg", "max"),
  weighted = TRUE,
  ...
)
```

Arguments

`exprSet`	an `ExpressionSet` object, each column of assay data represents a ranked list obtained by preprocessing the corresponding gene expression profile, and phenotypic data represents the short description (characteristics of gene expression profile, such as the drug type, the disease state) about the assay data.
`SignatureLength`	the length of "gene signature". In order to compute pairwise distances among samples, genes lists are ranked according to the gene expression ratio (fold change). And the "gene signature" includes the most up-regulated genes (near the top of the list) and the most down-regulated genes (near the bottom of the list).
`MergingDistance`	distance to be used which "measures" the similarity of ordered lists, "Spearman" or "Kendall".
`ScoringMethod`	method to be used to perform distance scoring, "GSEA" or "PGSEA".

ScoringDistance
: the distance measurements between PRLs: the Average Enrichment Score Distance ("avg"), or the Maximum Enrichment Score Distance ("max").

weighted
: there are tow rank merging approaches for two cases: if weighted = FALSE, all ranked list with the same biological state are treated equally important, a simple but useful method average ranking technique is selected; otherwise, weighted = TRUE, each individual ranked lists has its own ranked weights, this takes the iterative rank-aggregating algorithm, default is TRUE.

...
: additional arguments can be passed to ScoreGSEA()(while ScoringMethod = "GSEA") or to ScorePGSEA()(while ScoringMethod = "PGSEA".

Value

the result from ScoreGSEA() or ScorePGSEA().

See Also

RankMerging(),ScoreGSEA(),ScorePGSEA()

Examples

```
#load the sample expressionSet
data(exampleSet)

# distance scoring
SignatureDistance(
  exampleSet,
  SignatureLength = 250,
  MergingDistance = "Spearman",
  ScoringMethod = "GSEA",
  ScoringDistance = "avg",
  weighted = TRUE
)
```

Index

Biobase::ExpressionSet, *5–7*

exampleSet, 2

GeneExpressionSignature, 3
getRLs, 3

RankMerging, 4
RankMerging(), *9*

ScoreGSEA, 5
ScoreGSEA(), *7*, *9*
ScorePGSEA, 7
ScorePGSEA(), *6*, *9*
SignatureDistance, 8
SignatureDistance(), *5–7*