Package: GSDA (via r-universe)

September 6, 2024

Type Package

Title Gene Set Distance Analysis (GSDA)

Version 1.0

Date 2021-01-014

Description The gene-set distance analysis of omic data is implemented by generalizing distance correlations to evaluate the association of a gene set with categorical and censored event-time variables.

Depends R (>= 3.5.0),msigdbr

License GPL (>= 2)

biocViews Microarray, Bioinformatics, Gene expression

Suggests knitr, rmarkdown

VignetteBuilder knitr

LazyLoad yes

Repository https://xueyuancao.r-universe.dev

RemoteUrl https://github.com/xueyuancao/gsda

RemoteRef HEAD

RemoteSha d7f71cb8b75caa0e7aae2c65795ee87326f5da94

Contents

GSDA-package *Gene Set Distance Analysis (GSDA)*

Description

The gene-set distance analysis of omic data is implemented by generalizing distance correlations to evaluate the association of a gene set with categorical and censored event-time variables.

Details

The DESCRIPTION file: This package was not yet installed at build time.

Index: This package was not yet installed at build time.

Author(s)

Xueyuan Cao [aut, cre], Stanley Pounds [aut]

Maintainer: Xueyuan Cao <xcao12@uthsc.edu>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

Examples

```
data(target.aml.clin)
data(target.aml.expr)
data(kegg.ml.gsets)
res=gsda(target.aml.expr,
         target.aml.clin,
         kegg.ml.gsets,
         "Chloroma","oe","ct")
```

`best.dist.corr` *Best Distance Correlattion*

Description

Use a backward elimination procedure to identify a subset of variables in X most strongly associated with Y according to the distance correlation p-value.

Usage

```
best.dist.corr(X, Y, x.dist = "oe", y.dist = "oe")
```

Arguments

`X`	The omic numeric data matrix with subjects in rows and variables in columns. Note that this is the TRANSPOSE of the omic data matrix for some other omic data analysis packages and for the gsda function of this package.
`Y`	Numeric data matrix, vector, or data.frame with each row representing a subject. The function assumes the same set of subjects are represent in the same order in X and Y.
`x.dist`	The distance metric for omic data (X), may be "oe" (overall Euclidean), "me" (marginal Euclidean), "om" (overall Manhattan), or "mm" (marginal Manhattan).
`y.dist`	The distance metric for clinical data, may be "oe" (overall Euclidean), "me" (marginal Euclidean), "om" (overall Manhattan), or "mm" (marginal Manhattan), same options as for X

Details

This function computes dist.corr for X and Y. It then determines which column of X may be dropped to give the smallest p-value in dist.corr. This process is repeated until X has been reduced to only one variable. In this way, a dist.corr p-value is obtained after dropping each X variable. The subset of X variables giving the smallest p-value in this series of analyses is returned with additional result details.

Value

A list with the following components:

`rX`	reduced X matrix
`best.res`	best result by backward elimination
`all.res`	all backward elimination results: the first column has the index of the column of X that was dropped; the second column has the negative log10 p-value of the resulting X matrix
`X`	echoes input X
`Y`	echoes input Y

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

See Also

dist.corr

Examples

```
data(target.aml.clin)
data(target.aml.expr)
target.aml.expr=sqrt(target.aml.expr)
target.aml.expr=t(target.aml.expr)

bdc.chl=best.dist.corr(target.aml.expr,
                       target.aml.clin$Chloroma,
                       "oe","ct")
```

---

cat.dist *Distance for a Categorical Variable*

Description

A function to calculate the distance for a categorical variable.

Usage

```
cat.dist(X)
```

Arguments

X	vector of category designations

Details

This function calculates distance function for a categorical variable. The result is a square n by n matrix in which entry (i,j) has value 1 if entry i and entry j of the input vector X are not equal and entry (i,j) of the result matrix has value 0 if entry i and entry j of the input vector are equal. The distance between subject i and subject j is zero if the two subjects have the same categorical designation. The distance between subject i and subject j is one if the two subjects do not have the same categorical designation.

Value

A square matrix with each dimension equal to the length of X.

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

See Also

surv.dist

Examples

```
data(target.aml.clin)
cd=cat.dist(target.aml.clin$Chloroma)
cd[1:5,1:5]
```

---

dist.corr *Distance Correlattion*

Description

Calculate the distance correlation for a gene set's omic data matrix with another variable.

Usage

```
dist.corr(X, Y, x.dist = "me", y.dist = "me")
```

Arguments

X	The omic numeric data matrix with subjects as rows and variables as columns. Note this is the TRANSPOSE of how some omic data analysis packages represent omic data and how the omic data is represented in the gsda function of this package.
Y	Numeric data matrix, vector, or data.frame. The rows of X and rows of Y must represent the same set of subjects in the same order.
x.dist	The distance metric for omic data (X), may be "oe" (overall Euclidean), "me" (marginal Euclidean), "om" (overall Manhattan), or "mm" (marginal Manhattan).
y.dist	The distance metric for clinical data, may be "oe" (overall Euclidean), "me" (marginal Euclidean), "om" (overall Manhattan), or "mm" (marginal Manhattan), same options as for X

Details

The function calculates distance matrix for X and Y using one of the four methods "oe" (overall Euclidean), "me" (marginal Euclidean), "om" (overall Manhattan), or "mm" (marginal Manhattan). Then, the distance matrices are centered by U-centering and distance correlation is calculated as the inner product of the two U-centered distance matrices over the squared of inner product of each of the two U-centered distance matrices. The distance correlation t-statistics follows a t-distribution with n*(n-3)/2 degree of freedom according to Zhu et al.(2020).

Value

A list with the following components:

`odCor`	overall distance correlation statistic
`t.odCor`	t-stat for overall distance correlation statistic
`p.odCor`	p-value for overall distance correlation statistic
`dCor`	distance-based correlation matrix for each pair of variables.
`t.dCor`	t-stat for distance-based correlation matrix
`p.dCor`	p-value for distance-based correlation matrix
`X`	echo input data matrice X
`Y`	echo input data matrice Y
`x.dist`	echo input distance metric for X
`y.dist`	echo input distance metric for Y

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

Zhu C, Yao S, Zhang X and Shao X (2020) Distance-based and RKHS-based Dependence Metrics in High Dimension. arXiv:1902.03291

See Also

`best.dist.corr`

Examples

```
data(target.aml.clin)
data(target.aml.expr)
target.aml.expr=sqrt(target.aml.expr)
target.aml.expr=t(target.aml.expr)

dc.chl=dist.corr(target.aml.expr,
                 target.aml.clin$Chloroma,
                 "oe","ct")
```

gsda *Gene-Set Distance Analsysis (GSDA)*

Description

This function impletements the gene-set distance analysis (GSDA) of omic data by generalizing distance correlations to evaluate the association of each of a series gene sets with numeric, categorical, and censored event-time variables.

Usage

```
gsda(omic.data, clin.data, vset.data, clin.vars, omic.dist, clin.dist)
```

Arguments

omic.data	The genomic data matrix with features as rows and subjects as columns. The column names of omic.data are assumed to be observation identifiers. The gsda function calls the function prep.gsda to merge omic.data (by column name) and clin.data (by the column named "ID") before performing the GSDA procedure.
clin.data	A data frame of clinical data. Each row is a subject and each column is a variable. The "ID" column of clin.data includes observation identifiers. The gsda function calls the function prep.gsda to merge omic.data (by column name) and clin.data (by the column named "ID") before performing the GSDA procedure.
vset.data	Variable set data. Each row assigns a variable (column named vID) to a variable set (column named vset).
clin.vars	Column name(s) of clinical variable(s) to be associated with the gene-sets.
omic.dist	The distance metric for omic data, may be "oe" (overall Euclidean), "me" (marginal Euclidean), "om" (overall Manhattan), or "mm" (marginal Manhattan)
clin.dist	The distance metric for clinical data, may be "oe" (overall Euclidean), "me" (marginal Euclidean), "om" (overall Manhattan), or "mm" (marginal Manhattan)

Details

This function performs the GSDA method described by Cao and Pounds (2020) through generalizing distance correlations to evaluate the association of a gene set with categorical and censored event-time variables. The distance matrices are centered by U-centering and distance correlation is the inner product of the two U-centered distance matrices over the squared of inner product of each of the two U-centered distance matrices. The distance correlation t-statistics asymptotically follows a t-distribution with n*(n-3)/2 degree of freedom according to Zhu et al. (2020).

Value

A data.frame with the following columns:

vset	The name of variable set (gene-set).
vIDs	The list of variables in the variable set (gene-set).

dCor	The distance association statistics for the variable set.
p.vset	The p-value.
comp.time	Computation time for each set.

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

Zhu C, Yao S, Zhang X and Shao X (2020) Distance-based and RKHS-based Dependence Metrics in High Dimension. arXiv:1902.03291

See Also

GSDA

Examples

```
data(target.aml.clin)
data(target.aml.expr)
data(kegg.ml.gsets)
res=gsda(target.aml.expr,
         target.aml.clin,
         kegg.ml.gsets,
         "Chloroma","oe","ct")
```

---

kegg.ml.gsets *KEGG gene set data for the AML and CML pathways*

Description

A data set with the list of ensemble gene identifiers for the acute myeloid leukemia (AML) and chronic myeloid leukemia pathways as defined in the KEGG pathway database

Usage

```
data("kegg.ml.gsets")
```

Format

A data frame with 128 rows describing the pairings between the following 2 variables.

vset KEGG pathway name

vID Ensemble gene (ENSG) identifier

Details

A dataset with assignments of ensemble gene identifiers (ENSG) to KEGG pathway names

Source

http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_CHRONIC_MYELOID_LEUKEMIA

http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_ACUTE_MYELOID_LEUKEMIA

Merged with information from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz to translate gene symbols to Ensemble gene (ENSG) identifiers

Examples

```
data(kegg.ml.gsets)
```

---

prep.gsda *GSDA Data Preparation*

Description

A function to prepare omic data, clinical data and variable set into an ordered matched format for GSDA analysis.

Usage

```
prep.gsda(data.mtx, clin.data, vset.data = NULL)
```

Arguments

data.mtx	Numeric data matrix with column names giving subject identifiers.
clin.data	Data.frame with column named "ID" with subject identifiers matching column names of data.mtx.
vset.data	data.frame of variable-set assignments with columns named "vID" for variable identifier and "vset" for name or identifier of a variable set (gene-set).

Details

The gsda function uses prep.gsda to prepare the omic data matrix, clinical dataframe and variable set (gene set) into ordered and matched format, which is then used for GSDA analysis.

Value

A list with the following components:

omic.data	data matrix with columns in the same order as clin.data$ID.
clin.data	data.frame with ID column in same order as columns of omic.data.
vset.data	variable set ordered by name of variable set.
vset.index	simple data.frame showing first and last row of vset.data for each variable set.

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

See Also

gsda

Examples

```
data(target.aml.clin)
data(target.aml.expr)
data(kegg.ml.gsets)
gsdaprep=prep.gsda(target.aml.expr,
         target.aml.clin,
         kegg.ml.gsets)
```

---

prep.msigdb *Preparation of MSigDB for GSDA*

Description

This function prepares the gene sets of a species in MsigDB for gene-set distance analysis.

Usage

```
prep.msigdb(species = "Homo sapiens", vset = "gs_name", vID = "gene_symbol")
```

Arguments

species	Name of species in MSigDB.
vset	Name of MSigDB column to use as vset in gsda, default is "gs_name".
vID	Name of MSigDB column to use as vID in gsda, default is "gene_symbol".

Details

Take a species from MsigDB (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp), extract gene set definiton and prepare a data frame with gene sets and genes to be used as vset.data in the gsda function.

Value

A two-column data.frame with the columns vset and vID

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

Examples

```
gsets=prep.msigdb()
head(gsets)
```

---

print.bdc *Print Method for Best Distance Correlation*

Description

Print the result of the best distance correlation (best.dist.corr)

Usage

```
## S3 method for class 'bdc'
print(x,...)
```

Arguments

x	a class of bdc
...	further arguments passed to or from other methods

Details

Print the summary of result of best distance correlation to stdout.

Value

No return value, called for side effects

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

See Also

print.dcor, print.gsda.result

---

print.dcor *Print Method for Distance Correlation*

Description

Print the summary of result of distance correlation (dist.corr function).

Usage

```
## S3 method for class 'dcor'
print(x,...)
```

Arguments

x	result of dist.corr, class dcor
...	further arguments passed to or from other methods

Details

Print the summary of result of distance correlation to stdout.

Value

No return value, called for side effects

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

See Also

print.bdc, print.gsda.result

`print.gsda.result` *Print Method for GSDA Result*

Description

Print the of result of gene-set distance analysis (gsda function).

Usage

```
## S3 method for class 'gsda.result'
print(x,...)
```

Arguments

`x`	result of gene-set distance analysis (gsda function)
`...`	further arguments passed to or from other methods

Details

Print the result of gene-set distance analysis to stdout.

Value

No return value, called for side effects

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

See Also

`print.dcor`, `print.bdc`

surv.dist *Distance of a Survival Endpoint*

Description

A function to calculate the distance for a survivial endpoint.

Usage

```
surv.dist(stime.evnt)
```

Arguments

stime.evnt	A data frame with time in first column and censor in second column.

Details

This function calculates the distance matrix for a censored event-time variable. The calculation is based on the formula in Section 2.4 of Cao and Pounds (2021). The distance metric for censored event-time data is based on the rank-based association statistic for this type of data proposed by Jung et al (2005).

Value

A square matrix with nrow and ncol equal to the nrow of stime.evnt. Entry (i,j) of the result matrix gives the survival distance between subjects represented in rows i and j of stime.evnt.

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

Jung SH, Owzar K, and George SL (2005) A mutiple testing procedure to associate gene expression levels with survival. Statistics in Medicine 24: 3077-88.

See Also

cat.dist

Examples

```
data(target.aml.clin)
srv.dist=surv.dist(target.aml.clin[,c("efs.time","efs.evnt")])
```

target.aml.clin *Clinical outcomes for AML TARGET Project*

Description

A dataset with subject identifier, survival time, and death indicator for 123 pediatric AML patients

Usage

```
data("target.aml.clin")
```

Format

A data frame with 123 observations of the following 5 variables.

`ID` subject identifier, a character vector

`Chloroma` a character vector

`logWBC` a numeric vector

`efs.time` event-free survival time, a numeric vector

`efs.evnt` event indicator (0 = censored, 1 = event) for efs.time, a numeric vector

Details

A dataset with clinical data for each of 123 pediatric AML patients

Source

obtained from https://target-data.nci.nih.gov/Public/AML/clinical/harmonized/

Examples

```
data(target.aml.clin)
```

target.aml.expr *RNA-seq expression from the AML TARGET project*

Description

A matrix of RNA-seq gene expression values for 123 pediatric AML patients from the TARGET project for genes in the KEGG AML and CML pathways

Usage

```
data("target.aml.expr")
```

Format

Each row contains the expression values of one 94 Ensemble genes for all 123 patients. Each column contains the expression values of all 94 Ensemble genes for one patient. The rownames give the Ensemble identifiers for the genes. The columns give the patient identifiers.

Details

A RNA-seq dataset with expression levels in 94 ensemble gene identifiers for 123 pediatric AML patients

Source

https://target-data.nci.nih.gov/Public/AML/mRNA-seq/L3/expression/

Examples

```
data(target.aml.expr)
```

U.center *U Centering*

Description

U-center the distance matrix in preparation of computing distance correlations.

Usage

```
U.center(d)
```

Arguments

d	A square numeric data matrix

Details

This funtion centers the distance matrix according to U-centering formula on page 6 of arXiv 1902.03291 paper

Value

A centered data matrix

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

Zhu C, Yao S, Zhang X and Shao X. Distance-based and RKHS-based Dependence Metrics in High Dimension. arXiv:1902.03291

See Also

uc.dist

Examples

```
data(target.aml.clin)
cd=cat.dist(target.aml.clin$Chloroma)
ud=U.center(cd)
ud[1:5,1:5]
```

---

uc.dist *U-centered Distance Matrix*

Description

The function calculates the U-centered distance matrix for a variable.

Usage

```
uc.dist(X, dmeth = "me")
```

Arguments

X	vector, matrix, or data.frame to compute a distance matrix
dmeth	Distance method to use, options include "oe" for overall Euclidean, "me" for marginal Euclidean, "om" for overall Manhattan, "mm" for marginal Manhattan, "ct" for categorical, and "st" for censored survival time.

Details

A distance matrix is first calculated for a scale or data frame of a variable. The distance matrix is then centered according to U-centering formula on page 6 of arXiv 1902.03291 paper.

Value

For distance methods "oe", "om", "ct", and "st", one matrix of overall distances computed using data from all variables. For distance methods "me" and "mm", an array of distance matrices, one distance matrix per variable.

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

References

Cao X and Pounds S (2021) Gene-Set Distance Associations (GSDA): A Powerful Tool for Gene-Set Association Analysis.

Zhu C, Yao S, Zhang X and Shao X. Distance-based and RKHS-based Dependence Metrics in High Dimension. arXiv:1902.03291

See Also

U.center, prep.gsda

Examples

```
data(target.aml.expr)
target.aml.expr=sqrt(target.aml.expr)
target.aml.expr=t(target.aml.expr)
oe.dist=uc.dist(target.aml.expr,"oe") # overall Euclidean
```

---

write.gsda.csv.file *Write GSDA Result to a Comma Delimited File*

Description

Write a gene-set distance analysis result to a comma delimited file (.csv)

Usage

```
write.gsda.csv.file(gsda.result, out.file)
```

Arguments

gsda.result	A class of gene-set distance analysis result
out.file	A .csv file name with directory

Value

A saved .csv file.

Author(s)

Xueyuan Cao <xcao12@uthsc.edu> and Stanley Pounds <stanley.pounds@stjude.org>

Index