Package ‘PhylteR’

July 11, 2017

Type Package

Title Detection of Outliers in a List of Gene Trees

Version 1.0.0

Description Detection of outliers in a list of genes trees. Detected outliers could be complete outliers (genes and/or species) different from all the others. PhylteR can also detect gene/specie outliers couples
(a leaf outlier in a particular tree). Phylter also use a missMDA method (imputePCA) to impute the position of
missing species in gene trees. A web application of this package exists online at \{ }url{lbbe-shiny.univ-lyon1.fr/PhylteR}.

Encoding UTF-8

LazyData true

License GPL-3

Depends R (>= 3.3.0)

Imports FactoMineR, ape, DistatisR, phangorn, ggplot2, stats, utils

RoxygenNote 6.0.1

NeedsCompilation no

Author Damien de Vienne [aut, ctb, cre],
Stephane Dray [ctb],
Aurore Comte [aut, ctb]

Maintainer Damien de Vienne <damien.de-vienne@univ-lyon1.fr>

R topics documented:

`detect.cell.outliers` *detect.cell.outliers*

Description

Function to detect cell outliers (species and genes)

Usage

```
detect.cell.outliers(mat2WR, k = 3)
```

Arguments

`mat2WR`	the 2WR matrix obtained with the Dist2WR function.
`k`	the strength of outlier assignement. the Higher this value the more stringent the detection (less outliers detected).

Details

This function must be used after all complete outliers (species and genes) have been removed from the data. detect.cell.outliers is a function taken from the method phylo-MCOA (de Vienne M.D., Ollier S. et Aguileta G. (2012) Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis. Molecular Biology and Evolution 29 : 1587 – 1598)

Value

"outcell" All cell-by-cell outliers as a matrix with two columns. Each line represents a cell-by-cell outliers

detect.complete.outliers *detect.complete.outliers*

Description

Function to detect complete outliers (species and genes)

Usage

```
detect.complete.outliers(mat2WR, k = 1.5, thres = 0.5)
```

Arguments

mat2WR	the 2WR matrix obtained with the Dist2WR function.
k	the strength of outlier assignement. the Higher this value the more stringent the detection (less outliers detected).
thres	threshold above which genes or species are considered as complete outliers. 0.5 means that a gene or a species is a complete outlier if it is detected as outlier for more than 50% of the species or genes respectively.

Details

Must be runed before the detection of cell outliers detect.complete.outliers is a function taken from the method phylo-MCOA (de Vienne M.D., Ollier S. et Aguileta G. (2012) Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis. Molecular Biology and Evolution 29 : 1587 – 1598).

Value

"mat2WR" The 2WR matrix used to detect outliers. "outgn" Array containing all the complete outlier genes detected. "outsp" Array containing all the complete outlier species detected.

Dist2WR *Dist2WR*

Description

This function creates the two-way reference matrix (2WR) from distatis results.

Usage

```
Dist2WR(Distatis)
```

Arguments

`Distatis` is the output of the fonction mat2Dist (or of distatis from the Distatis R package (Beaton D., Chin Fatt C., Abdi H. (2013) DistatisR: DISTATIS Three Way Metric Multidimensional Scaling. Package R))

Value

2WR matrix is a gene x specie matrix. Each cell corresponds to the distance of a specie from a gene tree to the reference position of this specie for every gene trees.

`Fungi` *Fungi*

Description

246 genes trees of a set of 21 fungi species.

Usage

```
data(Fungi)
```

Format

Multiphylo.

Source

https://www.ncbi.nlm.nih.gov/pubmed/18709599

`impMean` *impMean*

Description

Imputing missing data in matrices. A missing species for a gene is imputed by the mean of the values of this species for every others genes.

Usage

```
impMean(matrices)
```

Arguments

`matrices` A list of distance matrices containing missing data. Each matrices should be named (use the rename.genes function if it is not the case)

Value

Return a list of matrices without missing data.

---

impPCA.multi *impPCA.multi*

Description

Imputing missing data in matrices

Usage

```
impPCA.multi(matrices, ncp = 3, center = FALSE, scale = FALSE,
  maxiter = 1000)
```

Arguments

matrices	A list of distance matrices with missing data. Each matrices should be named (use the rename.genes function if it is not the case)
ncp	integer corresponding to the number of components used to to predict the missing entries.
center	boolean. By default FALSE leading to data not centered.
scale	boolean. By default FALSE leading to not a same weight for each variable.
maxiter	integer, maximum number of iteration for the algorithm.

Value

Return a list of matrices without missing data.

See Also

imputePCA2

imputePCA2 *imputePCA2*

Description

Impute the missing values of a dataset with the Principal Components Analysis model.

Usage

```
imputePCA2(X, ncp = 2, center = FALSE, scale = FALSE,
  method = c("Regularized", "EM"), row.w = NULL, coeff.ridge = 1,
  threshold = 1e-06, seed = NULL, nb.init = 1, maxiter = 1000)
```

Arguments

X	a data.frame with continuous variables containing missing values
ncp	integer corresponding to the number of components used to to predict the missing entries
center	boolean. By default FALSE leading to data not centered
scale	boolean. By default FALSE leading to not a same weight for each variable
method	"Regularized" by default or "EM"
row.w	row weights (by default, a vector of 1 for uniform row weights)
coeff.ridge	1 by default to perform the regularized imputePCA2 algorithm; useful only if method="Regularized". Other regularization terms can be implemented by setting the value to less than 1 in order to regularized less (to get closer to the results of the EM method) or more than 1 to regularized more (to get closer to the results of the mean imputation)
threshold	the threshold for assessing convergence
seed	integer, by default seed = NULL implies that missing values are initially imputed by the mean of each variable. Other values leads to a random initialization
nb.init	integer corresponding to the number of random initializations; the first initialization is the initialization with the mean imputation
maxiter	integer, maximum number of iteration for the algorithm

Details

imputePCA function from missMDA package (Josse J. et Husson F. (2012) Handling missing values in exploratory multivariate data analysis method. Journal de la Société Française de Statistique vol. 153 (2): 79-99.) with some ajustements to fit trees data.

see also ?missMDA::imputePCA

mat2Dist *mat2Dist*

Description

mat2Dist applies distatis on a list of distance matrices.

Usage

```
mat2Dist(matrices, Norm = "NONE")
```

Arguments

`matrices`	A list of distance matrices
`Norm`	Norm = "none" (defaut) if we dont want to normalize data. Norm = "mfa" to normalize data.

Details

This function uses distatis from the DistatisR package (Beaton D., Chin Fatt C., Abdi H. (2013) DistatisR : DISTATIS Three Way Metric Multidimensional Scaling.).

See Also

`distatis`

normalize *normalize*

Description

This function normalizes the 2WR matrix (or any matrix) according to the species (rows) or to the genes (columns). normalize is a function taken from the method phylo-MCOA (de Vienne M.D., Ollier S. et Aguileta G. (2012) Phylo-MCOA: A Fast and Efficient Method to Detect Outlier Genes and Species in Phylogenomics Using Multiple Co-inertia Analysis. Molecular Biology and Evolution 29 : 1587 – 1598)

Usage

```
normalize(mat, what = "none")
```

Arguments

mat — A matrix

what — Character string indicating whether the matrix should be normalized and how. If what="none", the matrix is not normalized (the default), if what="species", the matrix is normalized so that the difference between species is increased, and if what="genes", the matrix is normalized so that the difference between genes is increased.

Value

A normalized matrix

PhylteR *PhylteR*

Description

This function finds complete and cell outliers inside a list of gene trees.

Usage

```
PhylteR(trees, distance = "patristic", bvalue = 0, method.imp = "IPCA",
ncp = 3, center = FALSE, scale = FALSE, maxiter = 1000, k = 1.5,
thres = 0.5, gene.names = NULL, Norm = "NONE")
```

Arguments

trees — The list of gene trees

distance — parameter from the function trees2matrices to transform trees into distance matrices. Distance could be "nodal" or "patristic" (default).

bvalue — This argument is only used if trees contain bootstrap values. It determines under what bootstrap values the nodes should be collapsed. Value 0 (the default) means that no nodes are collapsed.

method.imp — The method used for missing data imputation. "IPCA" for imputation with iteractive PCA (Slower but more accurate) ."MEAN" for imputation by means (faster but less accurate).

ncp — only used if method.imp = "IPCA". integer corresponding to the number of components used to to predict the missing entries.

center — only used if method.imp = "IPCA". boolean. By default FALSE leading to data not centered.

scale — only used if method.imp = "IPCA". boolean. By default FALSE leading to not a same weight for each variable.

maxiter — only used if method.imp = "IPCA". integer, maximum number of iteration for the algorithm.

`k` the strength of outlier assignement. The higher this valu,e the more stringent the detection (less outliers detected).

`thres` For the detection of complete outlier. Threshold above which genes or species are considered as complete outliers. 0.5 means that a gene or a species is a complete outlier if it is detected as outlier for more than 50% of the species or genes respectively.

`gene.names` List of gene names if the user want to renames the list of trees. NULL by default.

`Norm` Type of normalization used for the function mat2dist. Current options are NONE (default) or MFA (that normalizes each matrix so that its first eigenvalue is equal to one).

Details

The detection is done in two steps. The first step is the detection of complete outliers. Complete outliers detected are then removed of the list of trees and the second step is the detection of cell outliers in this list.

Value

`$Complete$mat2WR`
The 2WR matrix used to detect complete outliers.

`$complete$outgn`
The list of complete outliers genes.

`$complete$outsp`
The list of complete outliers species.

`$CellByCell$outcell`
The list of cell outliers.

Examples

```
# Detecting outliers of the dataset Fungi using nodal distances.
# This data set doesn't contain any missing data.

data(Fungi)

Results <- PhylteR(Fungi, distance = "nodal", bvalue = 0, k = 3,
thres = 0.6, gene.names = NULL, Norm = "NONE")

# See results
# Complete outliers

outgn <- Results$complete$outgn
outsp <- Results$complete$outsp

# outliers cell

outcell <- Results$CellByCell$outcell

# you can visualize the 2WR matrices (genes x species) with the function plot2WR.
```

```
plot = plot2WR(Results$Complete$mat2WR)
```

plot2WR *plot2WR*

Description

This function permits to plot the 2WR matrix.

Usage

```
plot2WR(matrixWR2)
```

Arguments

`matrixWR2`	The two-way reference matrix (2WR) from the Dist2WR function.

Value

Return a level plot of the 2WR matrix. It can be informative to look at the complete 2WR-matrix before doing any further analysis. It gives a visual idea of the overall congruence or incongruence in the dataset.

Examples

```
# Detecting outliers of the dataset Fungi using nodal distances.
# This data set doesn't contain any missing data.

data(Fungi)

Results <- PhylteR(Fungi, distance = "nodal", bvalue = 0, k = 3,
thres = 0.6, gene.names = NULL, Norm = "NONE")

# you can visualize the 2WR matrices (genes x species) with the function plot2WR.

plot = plot2WR(Results$Complete$mat2WR)
```

plotDistatisPartial *plotDistatisPartial*

Description

plotDistatisPartial plots maps of the factor scores of the observations from a distatis analysis.

Usage

```
plotDistatisPartial(trees, distance = "patristic", bvalue = 0,
gene.names = NULL, method.imp = "IPCA", ncp = 3, center = FALSE,
scale = FALSE, maxiter = 1000, Norm = "none")
```

Arguments

trees	A list of gene trees in multiphylo format.
distance	A method to generate distance matrices. It could be "nodal" to establish that the distance between two species is the number of nodes that separate them. Or "patristic" (default) if the distance between two species is be the sum of branch lengths between them.
bvalue	This argument is only used if trees contain bootstrap values. It determines under what bootstrap values the nodes should be collapsed. Value 0 (the default) means that no nodes are collapsed.
gene.names	List of gene names if the user want to renames the list of trees. NULL by default.
method.imp	The method used for missing data imputation. "IPCA" for imputation with iteractive PCA (Slower but more accurate) ."MEAN" for imputation by means (faster but less accurate).
ncp	only used if method.imp = "IPCA". integer corresponding to the number of components used to to predict the missing entries.
center	only used if method.imp = "IPCA". boolean. By default FALSE leading to data not centered.
scale	only used if method.imp = "IPCA". boolean. By default FALSE leading to not a same weight for each variable.
maxiter	only used if method.imp = "IPCA". integer, maximum number of iteration for the algorithm.
Norm	Type of normalization used for the function mat2dist. Current options are NONE (default) or MFA (that normalizes each matrix so that its first eigenvalue is equal to one).

Details

Function GraphDistatisPartial from DistatisR package (DiSTATIS Three Way Metric Multidimensional Scaling by Derek Beaton (2015)).

Value

`constraints`	A set of plot constraints that are returned.
`item.colors`	A set of colors for the observations are returned.
`participant.colors`	A set of colors for the participants are returned.

`rename.genes` *rename.genes*

Description

This function permits the user to add names to the genes trees.

Usage

```
rename.genes(trees, gene.names = NULL)
```

Arguments

`trees`	A list of gene trees in multiphylo format
`gene.names`	List of genes names the user wants to give to the list of trees. It should be of the same lenght of the list of trees. If NULL, genes are numeroted from 1 to the number of genes.

Value

The list of renamed trees in multiphylo format.

`rm.gene.and.species` *rm.gene.and.species*

Description

Suppress species or genes in a list of gene trees.

Usage

```
rm.gene.and.species(trees, sp2rm, gn2rm)
```

Arguments

`trees`	list of gene trees (in multiphylo format) from which we want to remove species or genes
`sp2rm`	species to remove as a list
`gn2rm`	genes to remove as a list

Value

Return a list of gene trees without the species or genes removed.

`trees2matrices` *trees2matrices*

Description

trees2matrices changes a list of trees into a list of matrices.

Usage

```
trees2matrices(trees, distance = "patristic", bvalue = 0)
```

Arguments

`trees`	A list of gene trees in multiphylo format.
`distance`	A method to generate distance matrices. It could be "nodal" to establish that the distance between two species is the number of nodes that separate them. Or "patristic" (default) if the distance between two species is be the sum of branch lengths between them.
`bvalue`	This argument is only used if trees contain bootstrap values. It determines under what bootstrap values the nodes should be collapsed. Value 0 (the default) means that no nodes are collapsed.

Value

return a list of distance matrices

Examples

```
# transforming a lsit of trees into a list of distances matrices using patristic distances:
data(Fungi)
matrices = trees2matrices(Fungi, distance = "patristic", bvalue = 0)
```

VizualizeGene *VizualizeGene*

Description

VizualizeGene plots, for a given gene, the distance between each species (red lines) with every other species (red points)

Usage

```
VizualizeGene(trees, gene, distance = "patristic",
bvalue = 0, gene.names = NULL, method.imp = "IPCA",
ncp = 3, center = FALSE, scale = FALSE, maxiter = 1000)
```

Arguments

`trees`	A list of gene trees in multiphylo format.
`gene`	The gene to plot.
`distance`	A method to generate distance matrices. It could be "nodal" to establish that the distance between two species is the number of nodes that separate them. Or "patristic" (default) if the distance between two species is be the sum of branch lengths between them.
`bvalue`	This argument is only used if trees contain bootstrap values. It determines under what bootstrap values the nodes should be collapsed. Value 0 (the default) means that no nodes are collapsed.
`gene.names`	List of gene names if the user want to renames the list of trees. NULL by default.
`method.imp`	The method used for missing data imputation. "IPCA" for imputation with iteractive PCA (Slower but more accurate) ."MEAN" for imputation by means (faster but less accurate).
`ncp`	only used if method.imp = "IPCA". integer corresponding to the number of components used to to predict the missing entries.
`center`	only used if method.imp = "IPCA". boolean. By default FALSE leading to data not centered.
`scale`	only used if method.imp = "IPCA". boolean. By default FALSE leading to not a same weight for each variable.
`maxiter`	only used if method.imp = "IPCA". integer, maximum number of iteration for the algorithm.

VizualizeSpe *VizualizeSpe*

Description

VizualizeSpe plots the distance between a chosen species and every other species (grey cirle) for every genes (red lines).

Usage

```
VizualizeSpe(trees, species, distance = "patristic",
bvalue = 0, gene.names = NULL, method.imp = "IPCA",
ncp = 3, center = FALSE, scale = FALSE, maxiter = 1000)
```

Arguments

`trees`	A list of gene trees in multiphylo format.
`species`	The species to plot.
`distance`	A method to generate distance matrices. It could be "nodal" to establish that the distance between two species is the number of nodes that separate them. Or "patristic" (default) if the distance between two species is be the sum of branch lengths between them.
`bvalue`	This argument is only used if trees contain bootstrap values. It determines under what bootstrap values the nodes should be collapsed. Value 0 (the default) means that no nodes are collapsed.
`gene.names`	List of gene names if the user want to renames the list of trees. NULL by default.
`method.imp`	The method used for missing data imputation. "IPCA" for imputation with iteractive PCA (Slower but more accurate) ."MEAN" for imputation by means (faster but less accurate).
`ncp`	only used if method.imp = "IPCA". integer corresponding to the number of components used to to predict the missing entries.
`center`	only used if method.imp = "IPCA". boolean. By default FALSE leading to data not centered.
`scale`	only used if method.imp = "IPCA". boolean. By default FALSE leading to not a same weight for each variable.
`maxiter`	only used if method.imp = "IPCA". integer, maximum number of iteration for the algorithm.

Index