Package 'Rtapas'

June 13, 2024

Title Random Tanglegram Partitions

Version 1.2

Description Applies a given global-fit method to random partial tanglegrams of a fixed size to identify the associations, terminals, and nodes that maximize phylogenetic (in)congruence. It also includes functions to compute more easily the confidence intervals of classification metrics and plot results, reducing computational time. See Llaberia-Robledillo et al., (2023) <doi:10.1093/sysbio/syad016>.

License MIT + file LICENSE

Encoding UTF-8

LazyData TRUE

Depends phytools, parallel, R (>= 3.5.0)

Imports ape, distory, paco, parallelly, stats, stringr, vegan

RoxygenNote 7.3.1

NeedsCompilation no

Suggests testthat (>= 3.2.1.1)

Config/testthat/edition 3

Author Mar Llaberia-Robledillo [aut, cre, cph] (<https://orcid.org/0000-0002-7989-796X>), Juan A. Balbuena [aut, cph] (<https://orcid.org/0000-0003-4006-1353>), José Ignacio Lucas-Lledó [aut, cph] (<https://orcid.org/0000-0001-6254-8942>), Oscar Alejandro Pérez-Escobar [aut, cph] (<https://orcid.org/0000-0001-9166-2410>)

Maintainer Mar Llaberia-Robledillo <mar.llaberia@uv.es>

Repository CRAN

Date/Publication 2024-06-13 12:00:01 UTC

Contents

`amph_trem` *amph_trem dataset*

Description

Data set of mitochondrial haplotypes of the trematode *Coitocaecum parvum* (Crowcroft, 1945) and those of its amphipod host, *Paracalliope fluviatilis* (Thomson, 1879) from several locations in South Island, New Zeland (Largue et al. 2016).

Usage

```
data(amph_trem)
```

Format

This data set compresses five objects:

`am_matrix` Associations between 17 haplotypes of *Coitocaecum parvum* and 59 haplotypes of *Paracalliope fluviatilis*. A binary matrix with 59 rows (amphipod) and 17 variables (trematode).

`amphipod` *Paracalliope fluviatilis* consensus tree. An object of class `"phylo"` containing a list with the details of the consensus phylogenetic tree (i.e. edges, edges length, nodes, and tips names).

`trematode` *Coitocaecum parvum* consensus tree. An object of class `"phylo"` containing a list with the details of the phylogenetic tree (i.e. edges, edges length, nodes and tips names).

`amphipod_1000tr` 1000 Bayesian posterior probability trees of *Paracalliope fluviatilis*. List of class `"multiphylo"` containing a 1000 phylogenetic trees with their respective details (i.e. edges, edges length, nodes, and tips names).

`trematode_1000tr` 1000 Bayesian posterior probability trees of *Coitocaecum parvum*. List of class `"multiphylo"` containing a 1000 phylogenetic trees with their respective details (i.e. edges, edges length, nodes, and tips names).

Source

Balbuena J.A., Perez-Escobar O.A., Llopis-Belenguer C., Blasco-Costa I. (2022). User's Guide Random Tanglegram Partitions V.1.0.0. Zenodo. doi:10.5281/zenodo.6327235

References

Lagrue C., Joannes A., Poulin R., Blasco-Costa I. (2016). Genetic structure and host–parasite co-divergence: evidence for trait-specific local adaptation. Biological Journal of the Linnean Society. 118:344–358.

Balbuena J.A., Perez-Escobar O.A., Llopis-Belenguer C., Blasco-Costa I. (2022). User's Guide Random Tanglegram Partitions V.1.0.0. Zenodo. doi:10.5281/zenodo.6327235

`assoc_mat` *Create an host-symbiont association matrix*

Description

Creates a binary host-symbiont association matrix from a two-columns matrix or data frame of host-symbiont associations.

Usage

```
assoc_mat(hs)
```

Arguments

`hs`	A two-columns matrix or data frame representing associations between hosts (column 1) and symbionts (column 2) species.

Value

An association binary matrix, with hosts in rows and symbionts in columns, sorted alphabetically.

Examples

```
data(nuc_cp)
NTaxa <- sort(NUCtr$tip.label)
CPTaxa <- sort(CPtr$tip.label)

NC <- assoc_mat(data.frame(NTaxa, CPTaxa))
```

geo_D *Geodesic distance between trees*

Description

For any trimmed matrix produced with `trimHS_maxC()` it prunes the host-symbiont phylogenies to conform with the trimmed matrix and computes geodesic distance between the pruned trees. NOTE: This function can only be used with strictly bifurcating trees.

Usage

```
geo_D(ths, treeH, treeS, strat = "sequential", cl = 1)
```

Arguments

`ths`	A trimmed matrix.
`treeH`	Host phylogeny. An object of class `"phylo"`.
`treeS`	Symbiont phylogeny. An object of class `"phylo"`.
`strat`	Flag indicating whether execution is to be `"sequential"` or `"parallel"`. Default is `"sequential"`, resolves R expressions sequentially in the current R process. If `"parallel"` resolves R expressions in parallel in separate R sessions running in the background.
`cl`	Number of cluster to be used for parallel computing. `parallelly::availableCores()` returns the number of clusters available. Default is `cl = 1` resulting in `"sequential"` execution.

Value

Geodesic distance

NOTE

The `node.label` object in both trees can not contain NAs or null values (i.e. no numeric value). All nodes should have a value. Else remove node labels within the `"phylo"` class tree with `tree$node.label <- NULL`. For more details, see `distory::dist.multiPhylo()`.

```
This function can not be used with the trimmed matrices produced
with \code{\link[=trimHS_maxI]{trimHS_maxI()}} or with the algorithm
\code{\link[=max_incong]{max_incong()}} in datasets with
multiple host-symbiont associations.
```

Source

Balbuena J.A., Perez-Escobar O.A., Llopis-Belenguer C., Blasco-Costa I. (2022). User's Guide Random Tanglegram Partitions V.1.0.0. Zenodo.

References

Schardl C.L., Craven K.D., Speakman S., Stromberg A., Lindstrom A., Yoshida R. (2008). A Novel Test for Host-Symbiont Codivergence Indicates Ancient Origin of Fungal Endophytes in Grasses. Systematic Biology. 57:483–498.

Balbuena J.A., Perez-Escobar Ó.A., Llopis-Belenguer C., Blasco-Costa I. (2020). Random Tanglegram Partitions (Random TaPas): An Alexandrian Approach to the Cophylogenetic Gordian Knot. Systematic Biology. 69:1212–1230.

Examples

```
data(amph_trem)
N = 10 #for the example, we recommend 1e+4 value
n = 8

TAM <- trimHS_maxC(N, am_matrix, n, check.unique = TRUE)
GD <- geo_D(TAM, amphipod, trematode, strat = "sequential", cl = 1)
```

`gini_ci` *Plot the confidence intervals of Gini coefficient*

Description

Computes and displays in a boxplot the Gini coefficient and their confidence intervals of the frequency (or residual/corrected frequencies) distributions of the estimated (in)congruence metric (with any of the three global-fit methods) of the individual host-symbiont associations.

Usage

```
gini_ci(LF_1, M01, ylab = "Gini coefficient", plot = TRUE, ...)
```

Arguments

`LF_1`	Vector of statistics produced with `max_cong()` or `max_incong()` for `"geoD"`, `"paco"` or `"paraF"`.
`M01`	Matrix produced with `prob_statistic()` for `"geoD"`, `"paco"` or `"paraF"` using `LF_1`.
`ylab`	Title of the y label.
`plot`	Default is `"TRUE"`, plots the Gini coefficient and its confidence intervals in a boxplot.
`...`	Any optional argument admissible in `boxplot()`

Value

The Gini values obtained and their representation in a boxplot, with their confidence intervals.

NOTE

It produces a conventional Gini coefficient (G) (Ultsch and Lötsch 2017) if all output values are positive, or a normalized Gini coefficient (G*) (Raffinetti et al. 2015) if negative values are produced due to corrected frequencies (if `res.fq = TRUE` or `diff.fq = TRUE`). For more details see Raffinetti et al. (2015).

References

Ultsch A., Lötsch J. (2017). A data science based standardized Gini index as a Lorenz dominance preserving measure of the inequality of distributions. PLOS ONE. 12:e0181572. doi:10.1371/journal.pone.0181572

Raffinetti E., Siletti E., Vernizzi A. (2015). On the Gini coefficient normalization when attributes with negative values are considered. Stat Methods Appl. 24:507–521. doi:10.1007/s10260014-02934

Examples

```
data(nuc_cp)
N = 1 #for the example, we recommend 1e+4 value
n = 15
# Maximizing congruence
NPc_PACo <- max_cong(np_matrix, NUCtr, CPtr, n, N, method = "paco",
                symmetric = FALSE, ei.correct = "sqrt.D",
                percentile = 0.01, res.fq = FALSE)

# Loaded directly from dataset
# THSC <- trimHS_maxC(N, np_matrix, n)
# pp_treesPACo_cong <- prob_statistic(ths = THSc, np_matrix, NUC_500tr[1:10],
#                          CP_500tr[1:10], freqfun = "paco", NPc_PACo,
#                          symmetric = FALSE, ei.correct = "sqrt.D",
#                          percentile = 0.01, correction = "none")

gini_ci(LF_1 = NPc_PACo, M01 = pp_treesPACo_cong,
         ylab = "Gini Coefficient (G)",
         plot = TRUE, ylim = c(0.3, 0.8))
abline(h = 1/3) # because res.fq = TRUE
```

`gini_RSV` *The Gini coefficient adjusted for negative attributes (Raffinetti, Siletti, & Vernizzi, 2015)*

Description

Computes the Gini coefficient adjusted for negative (even weighted) data.

Usage

```
gini_RSV(y)
```

Arguments

y	a vector of attributes containing even negative elements

Value

The value of the Gini coefficient adjusted for negative attributes.

NOTE

It produces a conventional Gini coefficient (G) (Ultsch and Lötsch 2017) if all output values are positive, or a normalized Gini coefficient (G*) (Raffinetti et al. 2015) if negative values are produced due to corrected frequencies (if `res.fq = TRUE` or `diff.fq = TRUE`). For more details see Raffinetti et al. (2015).

References

Ultsch A., Lötsch J. (2017). A data science based standardized Gini index as a Lorenz dominance preserving measure of the inequality of distributions. PLOS ONE. 12:e0181572. doi:10.1371/journal.pone.0181572

Raffinetti E., Siletti E., Vernizzi A. (2015). On the Gini coefficient normalization when attributes with negative values are considered. Stat Methods Appl. 24:507–521. doi:10.1007/s10260014-02934

Examples

```
data(nuc_cp)
N = 10 #for the example, we recommend 1e+4 value
n = 15
# Maximizing congruence
NPc_PACo <- max_cong(np_matrix, NUCtr, CPtr, n, N, method = "paco",
                symmetric = FALSE, ei.correct = "sqrt.D",
                percentile = 0.01, res.fq = FALSE)
gini_RSV(y = NPc_PACo)
```

---

linkf_CI *Confidence intervals for the frequency of host-symbiont association*

Description

From the matrix obtained in `prob_statistic()`, compute the confidence intervals for the frequencies (or residual/corrected frequencies) of the host-symbiont associations using a set of pairs of posterior probability trees of host and symbiont.

Usage

```
linkf_CI(
  freqfun = "paco",
  x,
  fx,
  c.level = 95,
  barplot = TRUE,
  col.bar = "lightblue",
  col.ci = "darkblue",
  y.lim = NULL,
  ...
)
```

Arguments

freqfun	Global-fit method. Options are "geoD" (Geodesic Distances), "paco" (PACo) or "paraF" (ParaFit). It should be the same method used to obtain "fx".
x	Matrix produced with prob_statistic() for"geoD" (Geodesic Distances), "paco" (PACo) or "paraF" (ParaFit).
fx	Vector of statistics produced with max_cong() or max_incong for"geoD" (Geodesic Distances), "paco" (PACo) or "paraF" (ParaFit).
c.level	Confidence interval level. Default is 95 (95\%).
barplot	Default is "TRUE", plots the distribution and confidence intervals of the frequencies.
col.bar	A vector of colors for the bars or bar components. By default, "lightblue" is used.
col.ci	A vector of colors for the confidence intervals arrows. By default, "darkblue" is used.
y.lim	Limits for the y axis.
...	Any graphical option admissible in barplot()

Value

A dataframe with associations information (columns 1 and 2), the observed value of the frequencies for these associations (column 3), the mean, the minimum and the maximum value of the frequencies (columns 4, 5 and 6) obtained with the sets of posterior probability trees.

Examples

```
data(nuc_cp)
N = 10 #for the example, we recommend 1e+4 value
n = 8
# Maximizing incongruence
NPi <- max_incong(np_matrix, NUCtr, CPtr, n, N, method = "paco",
                  symmetric = FALSE, ei.correct = "sqrt.D",
                  percentile = 0.99, diff.fq = TRUE,
                  strat = "parallel", cl = 8)
```

```
# Loaded directly from dataset
# THSi <- trimHS_maxI(N, np_matrix, n)
# pp_treesPACo_incong <- prob_statistic(ths = THSi, np_matrix,
#                        NUC_500tr[1:5], CP_500tr[1:5], freqfun = "paco",
#                        NPi, symmetric = FALSE, ei.correct = "sqrt.D",
#                        percentile = 0.99, diff.fq = TRUE, res.fq = FALSE,
#                        below.p = FALSE, strat = "parallel", cl = 8)
LFci <- linkf_CI (freqfun = "paco", x = pp_treesPACo_incong, fx = NPi,
                  c.level = 95, ylab = "Observed - Expected frequency")
```

`link_freq` *Frequency of host-symbiont association*

Description

Determines the frequency (or residual/corrected frequency) of each host-symbiont association in a given percentile of cases that maximize phylogenetic (in)congruence.

Usage

```
link_freq(
  x,
  fx,
  HS,
  percentile = 0.01,
  sep = "-",
  below.p = TRUE,
  res.fq = TRUE
)
```

Arguments

`x`	List of trimmed matrices produced by `trimHS_maxC()` or `trimHS_maxI()`.
`fx`	Vector of statistics produced with `geo_D()`, `paco_ss()` or `paraF()`
`HS`	Host-symbiont association matrix.
`percentile`	Percentile to evaluate (p). Default is `0.01` (1\%).
`sep`	Character that separates host and symbiont labels.
`below.p`	Determines whether frequencies are to be computed below or above the percentile set. Default is `TRUE`.
`res.fq`	Determines whether a correction to avoid one-to-one associations being over-represented in the percentile evaluated. If `TRUE` (default) a residual frequency value (observed - expected frequency) is computed for each host-symbiont association.

Value

A dataframe with host-symbiont associations in rows. The first and second columns display the names of the host and symbiont terminals, respectively. The third column designates the host-symbiont association by pasting the names of the terminals, and the fourth column displays the frequency of occurrence of each host-symbiont association. If `res.fq = TRUE`, column 5 displays the corrected frequencies as a residual.

NOTE

The `res.fq = TRUE` correction is recommended in tanglegrams with large portion of multiple (as opposed to one-to-one) host-symbiont associations. For future usage, frequencies of host-symbiont associations above a given percentile values can also be computed setting `below.p = FALSE`.

Examples

```
data(amph_trem)
N = 10 #for the example, we recommend 1e+4 value
n = 8

TAM <- trimHS_maxC(N, am_matrix, n, check.unique = TRUE)
PACO <- paco_ss(TAM, amphipod, trematode, symmetric = TRUE,
                ei.correct = "sqrt.D", strat = "parallel", cl = 8)
LFPACO <- link_freq(TAM, PACO, am_matrix, percentile = 0.01,
                    below.p = TRUE, res.fq = TRUE)
```

---

max_cong *Algortihm for maximizing congruence between two phylogenies*

Description

Prunes the host (H) and symbiont (S) phylogenies to conform with trimmed matrices and computes the given global fit method, Geodesic distances (GD), Procrustes Approach to Cophylogeny (PACo) or ParaFit (Legendre et al. 2002) between the pruned trees. Then, determines the frequency or corrected residual of each host-symbiont association occurring in a given percentile of cases that maximize phylogenetic congruence.

Usage

```
max_cong(
  HS,
  treeH,
  treeS,
  n,
  N,
  method = "paco",
  symmetric = FALSE,
```

```
  ei.correct = "none",
  percentile = 0.01,
  res.fq = TRUE,
  strat = "sequential",
  cl = 1
)
```

Arguments

HS	Host-Symbiont association matrix.
treeH	Host phyolgeny. An object of class "phylo".
treeS	Symbiont phylogeny. An object of class "phylo".
n	Number of unique associations.
N	Number of runs.
method	Specifies the desired global-fit method (GD, PACo or ParaFit). The default is PACo. Options are "geoD" (Geodesic Distances), "paco" (PACo) or "paraF" (ParaFit).
symmetric	Specifies the type of Procrustes superimposition. Default is FALSE, indicates that the superposition is applied asymmetrically (S depends on H). If TRUE, PACo is applied symmetrically (dependency between S and H is reciprocal).
ei.correct	Specifies how to correct potential negative eigenvalues from the conversion of phylogenetic distances into Principal Coordinates: "none" (the default) indicates that no correction is applied, particularly if H and S are ultrametric; "sqrt.D" takes the element-wise square-root of the phylogenetic distances; "lingoes" and "cailliez" apply the classical Lingoes and Cailliez corrections, respectively.
percentile	Percentile to evaluate (p). Default is 0.01 (1\%).
res.fq	Determines whether a correction to avoid one-to-one associations being over-represented in the percentile evaluated. If TRUE (default) a residual frequency value (observed - expected frequency) is computed for each host-symbiont association.
strat	Flag indicating whether execution is to be "sequential" or "parallel". Default is "sequential", resolves R expressions sequentially in the current R process. If "parallel" resolves R expressions in parallel in separate R sessions running in the background.
cl	Number of cluster to be used for parallel computing. parallelly::availableCores() returns the number of clusters available. Default is cl = 1 resulting in "sequential" execution.

Value

A dataframe with host-symbiont associations in rows. The first and second columns display the names of the host and symbiont terminals, respectively. The third column designates the host-symbiont association by pasting the names of the terminals, and the fourth column displays the frequency of occurrence of each host-symbiont association in p. If res.fq = TRUE, column 5 displays the corrected frequencies as a residual.

NOTE

If the node.label object in both trees contains NAs or empty values (i.e. no numeric value). All nodes should have a value. Else remove node labels within the "phylo" class tree with tree$node.label <- NULL. For more details, see distory::dist.multiPhylo()

Examples

```
data(nuc_pc)
N = 1 #for the example, we recommend 1e+4 value
n = 15
NPc <- max_cong(np_matrix, NUCtr, CPtr, n, N, method = "paco",
                symmetric = FALSE, ei.correct = "sqrt.D",
                percentile = 0.01, res.fq = FALSE)
```

---

max_incong *Algortihm for maximizing incongruence between two phylogenies*

Description

Prunes the host (H) and symbiont (S) phylogenies to conform with the trimmed matrix and computes the given global-fit method (PACo or ParaFit) between the pruned trees. Then, determines the frequency of each host-symbiont association occurring in a given percentile of cases that maximize phylogenetic incongruence.

Usage

```
max_incong(
  HS,
  treeH,
  treeS,
  n,
  N,
  method = "paco",
  symmetric = FALSE,
  ei.correct = "none",
  percentile = 0.99,
  diff.fq = FALSE,
  strat = "sequential",
  cl = 1
)
```

Arguments

HS	Host-Symbiont association matrix.
treeH	Host phyolgeny. An object of class "phylo".
treeS	Symbiont phylogeny. An object of class "phylo".

n	Number of associations.
N	Number of runs.
method	Specifies the desired global-fit method (PACo or ParaFit). The default is PACo. Options are "paco" (PACo) or "paraF" (ParaFit).
symmetric	Specifies the type of Procrustes superimposition. Default is FALSE, indicates that the superposition is applied asymmetrically (S depends on H). If TRUE, PACo is applied symmetrically (dependency between S and H is reciprocal).
ei.correct	Specifies how to correct potential negative eigenvalues from the conversion of phylogenetic distances into Principal Coordinates: "none" (the default) indicates that no correction is applied, particularly if H and S are ultrametric; "sqrt.D" takes the element-wise square-root of the phylogenetic distances; "lingoes" and "cailliez" apply the classical Lingoes and Cailliez corrections, respectively.
percentile	Percentile to evaluate (p). Default is 0.99 (99\%).
diff.fq	Determines whether a correction to detect those associations that present a similar contribution to (in)congruence and occur with some frequency at the 0.01 and 0.99 percentiles. These correction avoid multiple associations being overrepresented. If TRUE a corrected frequency value (observed in p - observed in (p-1)) is computed for each host-symbiont association.
strat	Flag indicating whether execution is to be "sequential" or "parallel". Default is "sequential", resolves R expressions sequentially in the current R process. If "parallel" resolves R expressions in parallel in separate R sessions running in the background.
cl	Number of cluster to be used for parallel computing. parallelly::availableCores() returns the number of clusters available. Default is cl = 1 resulting in "sequential" execution.y.

Value

A dataframe with host-symbiont associations in rows. The first and second columns display the names of the host and symbiont terminals, respectively. The third column designates the host-symbiont association by pasting the names of the terminals, and the fourth column displays the frequency of occurrence of each host-symbiont association in p. If diff.fq = TRUE, column 5 displays the corrected frequencies.

NOTE

The node.label object in both trees can not contain NAs or null values (i.e. no numeric value). All nodes should have a value. Else remove node labels within the "phylo" class tree with tree$node.label <- NULL. For more details, see distory::dist.multiPhylo().

```
\code{GD} method can not be used with the trimmed matrices produced
with \code{\link[=trimHS_maxI]{trimHS_maxI()}} or with the algorithm
\code{\link[=max_incong]{max_incong()}} for those datasets with
multiple associations.
```

Examples

```
data(nuc_pc)
N = 1 #for the example, we recommend 1e+4 value
n = 15
NPi <- max_incong(np_matrix, NUCtr, CPtr, n, N, method = "paco",
                  symmetric = FALSE, ei.correct = "sqrt.D",
                  percentile = 0.99, diff.fq = TRUE)
```

---

nuc_cp *Nuclear and chloroplast dataset of orchids*

Description

Data set of nuclear and chloroplast loci of 52 orchid taxa from Kew DNA and Tissue Collection, https://dnabank.science.kew.org/homepage.html (Perez-Escobar et al. 2021).

Usage

```
data(nuc_cp)
```

Format

This data set consists of seven objects:

`np_matrix` Associations one-to-one between the 52 orchid taxa. A binary matrix with 52 rows (nuclear) and 52 columns (chloroplast).

`NUCtr` Phylogeny constructed by sequence data of nuclear loci of orchids (Perez-Escobar et al. 2021). An object of class `"phylo"` containing the details of the phylogenetic tree (i.e. edge, edge length, nodes and tips names).

`CPtr` Phylogeny constructed by sequence data of chloroplast loci of orchids (Perez-Escobar et al. 2021). An object of class `"phylo"` containing the details of the phylogenetic tree (i.e. edge, edge length, nodes and tips names).

`NUC_500tr` 500 bootstrap replicates trees from Perez-Escobar et al. (2021). Object of class `"multiphylo"` containing a 500 phylogenetic trees with their respective details (i.e. edges, edges length, nodes, and tips names).

`CP_500tr` 500 bootstrap replicates trees from Perez-Escobar et al. (2021). Object of class `"multiphylo"` containing a 500 phylogenetic trees with their respective details (i.e. edges, edges length, nodes, and tips names).

`pp_treesPACo_cong` Matrix with the value of the PACo statistics generated for each pair (H and S) of posterior probability trees maximizing congruence between them.

`pp_treesPACo_incong` Matrix with the value of the PACo statistics generated for each pair (H and S) of posterior probability trees maximizing incongruence between them.

Source

Perez-Escobar O.A., Dodsworth S., Bogarin D., Bellot S., Balbuena J.A., Schley R., Kikuchi I., Morris S.K., Epitawalage N., Cowan R., Maurin O., Zuntini A., Arias T., Serna A., Gravendeel B., Torres M.F., Nargar K., Chomicki G., Chase M.W., Leitch I.J., Forest F., Baker W.J. (2021). Hundreds of nuclear and plastid loci yield novel insights into orchid relationships. American Journal of Botany, 108(7), 1166-1180.

References

Perez-Escobar O.A., Dodsworth S., Bogarin D., Bellot S., Balbuena J.A., Schley R., Kikuchi I., Morris S.K., Epitawalage N., Cowan R., Maurin O., Zuntini A., Arias T., Serna A., Gravendeel B., Torres M.F., Nargar K., Chomicki G., Chase M.W., Leitch I.J., Forest F., Baker W.J. (2021). Hundreds of nuclear and plastid loci yield novel insights into orchid relationships. American Journal of Botany, 108(7), 1166-1180.

`one2one_f` *Maximum number of unique one-to-one association over a number of runs*

Description

For a binary matrix of host-symbiont associations, it finds the maximum number of host-symbiont pairs, `n`, for which one-to-one unique associations can be chosen.

Usage

```
one2one_f(
  HS,
  reps = 10000,
  interval = NULL,
  strat = "sequential",
  cl = 1,
  plot = TRUE
)
```

Arguments

`HS`	Host-symbiont association matrix.
`reps`	Number of runs to evaluate.
`interval`	Vector with the minimum and maximum `n` that the user wants to test. Default is `"NULL"`, where a minimum `n` (10% of the total associations) and a maximum `n` (20% of the total associations) are automatically assigned.
`strat`	Flag indicating whether execution is to be `"sequential"` or `"parallel"`. Default is `"sequential"`, resolves R expressions sequentially in the current R process. If `"parallel"` resolves R expressions in parallel in separate R sessions running in the background.

cl	Number of cluster to be used for parallel computing. parallelly::availableCores() returns the number of clusters available. Default is cl = 1 resulting in "sequential" execution.
plot	Default is "TRUE", plots the number of unique host- symbiont associations in the "interval" range against the number of runs that could be completed.

Value

The maximum number of unique one-to-one associations (n).

NOTE

It can be used to decide the best n prior to application of max_cong().

Examples

```
N = 10  #for the example, we recommend 1e+4 value
data(amph_trem)
n <- one2one_f(am_matrix, reps = N, interval = c(2, 10), plot = TRUE)
```

---

paco_ss *Procrustes Approach to Cophylogeny (PACo) of the host and symbiont configurations*

Description

For any trimmed matrix produced with trimHS_maxC() or trimHS_maxI(), it prunes the host (H) and symbiont (S) phylogenies to conform with the trimmed matrix and runs Procruste Approach to Cophylogeny (PACo) to produce the squared sum of residuals of the Procrustes superimposition of the host and symbiont configurations in Euclidean space.

Usage

```
paco_ss(
  ths,
  treeH,
  treeS,
  symmetric = FALSE,
  proc.warns = FALSE,
  ei.correct = "none",
  strat = "sequential",
  cl = 1
)
```

Arguments

ths	Trimmed matrix.
treeH	Host phylogeny. An object of class "phylo".
treeS	Symbiont phylogeny. An object of class "phylo".
symmetric	Specifies the type of Procrustes superimposition. Default is FALSE, indicates that the superposition is applied asymmetrically (S depends on H). If TRUE, PACo is applied symmetrically (dependency between S and H is reciprocal).
proc.warns	Switches on/off trivial warnings returned when treeH and treeS differ in size (number of tips). Default is FALSE.
ei.correct	Specifies how to correct potential negative eigenvalues from the conversion of phylogenetic distances into Principal Coordinates: "none" (the default) indicates that no correction is applied, particularly if H and S are ultrametric; "sqrt.D" takes the element-wise square-root of the phylogenetic distances; "lingoes" and "cailliez" apply the classical Lingoes and Cailliez corrections, respectively.
strat	Flag indicating whether execution is to be "sequential" or "parallel". Default is "sequential", resolves R expressions sequentially in the current R process. If "parallel" resolves R expressions in parallel in separate R sessions running in the background.
cl	Number of cluster to be used for parallel computing. parallelly::availableCores() returns the number of clusters available. Default is cl = 1 resulting in "sequential" execution.

Value

A sum of squared residuals.

Source

Balbuena J.A., Perez-Escobar O.A., Llopis-Belenguer C., Blasco-Costa I. (2022). User's Guide Random Tanglegram Partitions V.1.0.0. Zenodo.

References

Balbuena J.A., Miguez-Lozano R., Blasco-Costa I. (2013). PACo: A Novel Procrustes Application to Cophylogenetic Analysis. PLOS ONE. 8:e61048.

Balbuena J.A., Perez-Escobar Ó.A., Llopis-Belenguer C., Blasco-Costa I. (2020). Random Tanglegram Partitions (Random TaPas): An Alexandrian Approach to the Cophylogenetic Gordian Knot. Systematic Biology. 69:1212–1230.

Examples

```
data(amph_trem)
N = 10 #for the example, we recommend 1e+4 value
n = 8

TAM <- trimHS_maxC(N, am_matrix, n, check.unique = TRUE)
```

```
PACO <- paco_ss(TAM, amphipod, trematode, symmetric = TRUE,
                ei.correct = "sqrt.D", strat = "parallel", cl = 8)
```

---

paraF *Test of host-symbiont coevolution*

Description

For any trimmed matrix produced with `trimHS_maxC()` or `trimHS_maxI()`, it prunes the host (H) and symbiont (S) phylogenies to conform with the trimmed matrix and runs `ape::parafit()` (Legendre et al. 2002) to calculate the ParaFitGlobal Statistic.

Usage

```
paraF(ths, treeH, treeS, ei.correct = "none", strat = "sequential", cl = 1)
```

Arguments

`ths`	Trimmed matrix.
`treeH`	Host phylogeny. An object of class `"phylo"`.
`treeS`	Symbiont phylogeny. An object of class `"phylo"`.
`ei.correct`	Specifies how to correct potential negative eigenvalues from the conversion of phylogenetic distances into Principal Coordinates: `"none"` (the default) indicates that no correction is applied, particularly if H and S are ultrametric; `"sqrt.D"` takes the element-wise square-root of the phylogenetic distances; `"lingoes"` and `"cailliez"` apply the classical Lingoes and Cailliez corrections, respectively.
`strat`	Flag indicating whether execution is to be `"sequential"` or `"parallel"`. Default is `"sequential"`, resolves R expressions sequentially in the current R process. If `"parallel"` resolves R expressions in parallel in separate R sessions running in the background.
`cl`	Number of cluster to be used for parallel computing. `parallelly::availableCores()` returns the number of clusters available. Default is `cl = 1` resulting in `"sequential"` execution.

Value

A number object with the ParaFitGlobal Statistic of host-symbiont test for the N trimmed matrix.

References

Legendre P., Desdevises Y., Bazin E. (2002). A Statistical Test for Host–Parasite Coevolution. Systematic Biology. 51:217–234.

Balbuena J.A., Perez-Escobar O.A., Llopis-Belenguer C., Blasco-Costa I. (2020). Random Tanglegram Partitions (Random TaPas): An Alexandrian Approach to the Cophylogenetic Gordian Knot. Systematic Biology. 69:1212–1230.

Examples

```
data(amph_trem)
N = 10 #for the example, we recommend 1e+4 value
n = 8

TAM <- trimHS_maxC(N, am_matrix, n, check.unique = TRUE)
PF <- paraF(TAM, amphipod, trematode, ei.correct = "sqrt.D",
           strat = "parallel", cl = 8)
```

`prob_statistic` *Frequencies of the associations for the posterior probability trees*

Description

Computes frequencies (or residual/corrected frequencies) of the host-symbiont associations for pairs (H and S) of posterior probability trees from the statistics generatedwith `GD` (Geodesic Distances), `PACo` (PACo) or `ParaFit`(ParaFit).

Usage

```
prob_statistic(
  ths,
  HS,
  mTreeH,
  mTreeS,
  freqfun = "paco",
  fx,
  percentile = 0.01,
  correction = "none",
  symmetric = FALSE,
  ei.correct = "none",
  algm = "maxcong",
  proc.warns = FALSE,
  strat = "sequential",
  cl = 1
)
```

Arguments

`ths`	List of trimmed matrices produced by `trimHS_maxC()` or `trimHS_maxI()`.
`HS`	Host-Symbiont association matrix.
`mTreeH`	Number of posterior-probability trees of host.
`mTreeS`	Number of posterior-probability trees of symbiont.

freqfun	The global-fit method to compute using the posterior probability trees. Options are "geoD" (Geodesic Distances), "paco" (PACo) or "paraF" (ParaFit). It should be the same method used to obtain "fx".
fx	Vector of statistics produced with max_cong() or max_incong() for GD, PACo or ParaFit.
percentile	Percentile to evaluate (p). Default is 0.01 (1\%).
correction	Correction to be assumed. The default value is "none". If = "res.fq", a residual frequency value (observed - expected frequency) is computed for each host-symbiont association that maximizes phylogenetic congruence. If = "diff.fq", a corrected frequency value (observed in p - observed in (p-1)) is computed for each host-symbiont association. It should be the same correction used to obtain "fx".
symmetric	Specifies the type of Procrustes superimposition. Default is FALSE, indicates that the superposition is applied asymmetrically (S depends on H). If TRUE, PACo is applied symmetrically (dependency between S and H is reciprocal).
ei.correct	Specifies how to correct potential negative eigenvalues from the conversion of phylogenetic distances into Principal Coordinates: "none" (the default) indicates that no correction is applied, particularly if H and S are ultrametric; "sqrt.D" takes the element-wise square-root of the phylogenetic distances; "lingoes" and "cailliez" apply the classical Lingoes and Cailliez corrections, respectively.
algm	Only required if correction = "none". Specifies the algorithm that has been used to obtain "fx" without correction. Use = "maxcong" for max_cong(), and = "maxincong" for max_incong().
proc.warns	Switches on/off trivial warnings returned when treeH and treeS differ in size (number of tips). Default is FALSE.
strat	Flag indicating whether execution is to be "sequential" or "parallel". Default is "sequential", resolves R expressions sequentially in the current R process. If "parallel" resolves R expressions in parallel in separate R sessions running in the background.
cl	Number of cluster to be used for parallel computing. parallelly::availableCores() returns the number of clusters available. Default is cl = 1 resulting in "sequential" execution.

Value

A matrix with the value of the statistics for each of the probability trees.

Examples

```
data("nuc_cp")
N = 10 #for the example, we recommend 1e+4 value
n = 15
# Maximizing congruence (not run)
NPc <- max_cong(np_matrix, NUCtr, CPtr, n, N, method = "paco",
                symmetric = FALSE, ei.correct = "sqrt.D",
                percentile = 0.01, strat = "parallel", cl = 8)
```

```
THSc <- trimHS_maxC(N, np_matrix, n)
pp_treesPACOo_cong <- prob_statistic(THSc, np_matrix, NUC_500tr[1:10],
                         CP_500tr[1:10], freqfun = "paco", NPc,
                         percentile = 0.01, correction = "none",
                         algm = "maxcong", symmetric = FALSE,
                         ei.correct = "sqrt.D",
                         strat = "parallel", cl = 8)

# Maximizing incongruence
NPi <- max_incong(np_matrix, NUCtr, CPtr, n, N, method = "paco",
                  symmetric = FALSE, ei.correct = "sqrt.D",
                  percentile = 0.99, diff.fq = TRUE)
THSi <- trimHS_maxI(N, np_matrix, n)
pp_treesPACOo_incong <- prob_statistic(THSi, np_matrix, NUC_500tr[1:5],
                                CP_500tr[1:5], freqfun = "paco", NPi,
                                percentile = 0.99, correction = "diff.fq",
                                symmetric = FALSE, ei.correct = "sqrt.D",
                                strat = "parallel", cl = 8)
```

---

tangle_gram *Tanglegram of the host-symbiont frequencies*

Description

Maps the estimated (in)congruence metrics of the individual host-symbiont associations as heatmap on a tanglegram. It also plots the average frequency (or residual/corrected frequency) of occurrence of each terminal and optionally, the fast maximum likelihood estimators of ancestral states of each node.

Usage

```
tangle_gram(
  treeH,
  treeS,
  HS,
  fqtab,
  colscale = "diverging",
  colgrad,
  nbreaks = 50,
  node.tag = TRUE,
  cexpt = 1,
  link.lwd = 1,
  link.lty = 1,
  fsize = 0.5,
  pts = FALSE,
  link.type = "straight",
  ftype = "i",
```

```
  ...
)
```

Arguments

treeH	Host phylogeny. An object of class "phylo".
treeS	Symbiont phylogeny. An object of class "phylo".
HS	Host-symbiont association matrix.
fqtab	Dataframe produced with max_cong() or max_incong().
colscale	Choose between "diverging", color reflects distance from 0 (centered at 0, recommended if "res.fq = TRUE") or "sequential", color reflects distance from minimum value (spanning from the min to max frequencies observed).
colgrad	Vector of R specified colors defining the color gradient of the heatmap.
nbreaks	Number of discrete values along "colgrad".
node.tag	Specifies whether maximum likelihood estimators of ancestral states are to be computed. Default is TRUE.
cexpt	Size of color points at terminals and nodes.
link.lwd	Line width for plotting, default to 1.
link.lty	Line type. Coded as lty in par().
fsize	Relative font size for tip labels.
pts	Logical value indicating whether or not to plot filled circles at each vertex of the tree, as well as at transition points between mapped states. Default is FALSE.
link.type	If curved linking lines are desired, set to "curved". Default is "straight".
ftype	Font type. Options are "reg", "i" (italics), "b" (bold) or "bi" (bold-italics).
...	Any graphical option admissible in phytools::plot.cophylo()

Value

A tanglegram with quantitative information displayed as heatmap.

NOTE

In order to calculate the ancestral states in the phylogenies, all nodes of the trees (node.label) must have a value (NA or empty values are not allowed). In addition, the trees must be time-calibrated and preferably rooted. If one of these elements is missing, an error will be generated and nodes and points of terminals will be displayed as black.

Examples

```
data(nuc_cp)
N = 10 #for the example, we recommend 1e+4 value
n = 8
NPc <- max_cong(np_matrix, NUCtr, CPtr, n, N, method = "paco",
                symmetric = TRUE, ei.correct = "sqrt.D",
                percentile = 0.01, res.fq = FALSE,
```

```
                strat = "parallel", cl = 4)
col = c("darkorchid4", "gold")
tangle_gram(NUCtr, CPtr, np_matrix, NPc, colscale = "sequential",
            colgrad = col, nbreaks = 50, node.tag = TRUE)
```

`trimHS_maxC` *Trims the H-S association matrix maximizing the congruence*

Description

For N runs, it randomly chooses `n` unique one-to-one associations and trims the H-S association matrix to include only the `n` associations.

Usage

```
trimHS_maxC(N, HS, n, check.unique = TRUE, strat = "sequential", cl = 1)
```

Arguments

`N`	Number of runs.
`HS`	Host-Symbiont association matrix.
`n`	Number of unique associations.
`check.unique`	if `TRUE` discards duplicated trimmed matrices. This alternative is recommended if `n` is small, because the probability of obtaining the same trimmed matrix in different runs increases as `n` decreases.
`strat`	Flag indicating whether execution is to be `"sequential"` or `"parallel"`. Default is `"sequential"`, resolves R expressions sequentially in the current R process. If `"parallel"` resolves R expressions in parallel in separate R sessions running in the background.
`cl`	Number of cluster to be used for parallel computing. `parallelly::availableCores()` returns the number of clusters available. Default is `cl = 1` resulting in `"sequential"` execution.

Value

A list of the N trimmed matrices.

Examples

```
data(nuc_cp)
N = 10  #for the example, we recommend 1e+4 value
n = 15
TNC <- trimHS_maxC(N, np_matrix, n, check.unique = TRUE)
```

`trimHS_maxI` *Trims the H-S association matrix maximizing the incongruence*

Description

For N runs, it randomly chooses `n` associations and trims the H-S association matrix to include them, allowing both single and multiple associations.

Usage

```
trimHS_maxI(N, HS, n, check.unique = TRUE, strat = "sequential", cl = 1)
```

Arguments

`N`	Number of runs.
`HS`	Host-Symbiont association matrix.
`n`	Number of associations.
`check.unique`	if `TRUE` discards duplicated trimmed matrices. This alternative is recommended if `n` is small, because the probability of obtaining the same trimmed matrix in different runs increases as `n` decreases.
`strat`	Flag indicating whether execution is to be `"sequential"` or `"parallel"`. Default is `"sequential"`, resolves R expressions sequentially in the current R process. If `"parallel"` resolves R expressions in parallel in separate R sessions running in the background.
`cl`	Number of cluster to be used for parallel computing. `parallelly::availableCores()` returns the number of clusters available. Default is `cl = 1` resulting in `"sequential"` execution.

Value

A list of the N trimmed matrices.

Examples

```
data(nuc_cp)
N = 10  #for the example, we recommend 1e+4 value
n = 15
TNC <- trimHS_maxI(N, np_matrix, n, check.unique = TRUE)
```

Index