

Package ‘phylopred’

October 14, 2026

Type Package

Title Predicting Species Interactions from Phylogenetic Distance

Version 0.1.0

Description Tools to model and predict host-parasite (and other pairwise species) interactions from phylogenetic distance under presence-only data. Provides pairwise host-sharing logistic models with cluster-aware bootstrap and slope comparison between clades, presence-only performance metrics (a threshold efficiency metric and a phylogenetic partial receiver operating characteristic (ROC) adapted from Peterson, Papes and Soberon (2008) [doi:10.1016/j.ecolmodel.2007.11.008](https://doi.org/10.1016/j.ecolmodel.2007.11.008)), positive-unlabeled host prediction, and leave-hosts-out validation. Methods build on Robles-Fernandez and Lira-Noriega (2017) [doi:10.3389/fams.2017.00017](https://doi.org/10.3389/fams.2017.00017).

License MIT + file LICENSE

URL <https://github.com/alrobles/phylopred>,
<https://alrobles.github.io/phylopred/>

BugReports <https://github.com/alrobles/phylopred/issues>

Encoding UTF-8

LazyData true

Imports graphics, stats, utils, xplus

Suggests ape, cofil, knitr, rmarkdown, testthat (>= 3.0.0)

Depends R (>= 4.1.0)

Config/testthat/edition 3

Config/roxygen2/version 8.0.0

NeedsCompilation no

Author Angel Robles [aut, cre] (ORCID:
<https://orcid.org/0000-0002-4674-4270>)

Maintainer Angel Robles <a.l.robles.fernandez@gmail.com>

Repository CRAN

Date/Publication 2026-10-14 12:12:07 UTC

Contents

align_phylopred_inputs	2
beetleTreeInteractions	3
bootstrap_phylopred_model	3
cluster_bootstrap_phylopred	4
compare_phylopred_slopes	5
evaluate_phylopred_model	7
extract_phylopred_coefficients	8
fish_tree	9
fit_phylopred_model	9
fit_phylopred_pu	10
host_specificity_index	12
host_threshold_metric	13
phylopred_confidence_ellipse	14
phylo_partial_roc	15
phy_dist	16
poulin_std	17
predict_phylopred_probability	18
prepare_incidence_matrix	19
prepare_pair_data	19
taxonomic_step_matrix	21
validate_host_prediction	22
Index	24

align_phylopred_inputs	<i>Align incidence matrix and phylogenetic distance matrix</i>
------------------------	--

Description

Validates both inputs and reorders the columns of incidence to match the column order of phydist.

Usage

```
align_phylopred_inputs(incidence, phydist)
```

Arguments

- incidence A binary incidence matrix (rows = parasites, cols = hosts).
- phydist A square phylogenetic distance matrix among hosts.

Value

A named list with:

- incidence** The incidence matrix with columns reordered to match phydist.
- phydist** The (unchanged) phylogenetic distance matrix.

Examples

```
incidence <- prepare_incidence_matrix(beetleTreeInteractions)
aligned <- align_phylopred_inputs(incidence, phy_dist)
```

beetleTreeInteractions

Beetle tree interaction table.

Description

A dataset of bark beetle species and their host tree genera. Original information is available in <https://www.barkbeetles.info/>

Usage

```
beetleTreeInteractions
```

Format

A data frame with 565 rows and 2 variables:

beetle Species of bark beetles

genus Genus of host tree of bark beetles

Details

We use this interaction tables to construct an incidence matrix. This is a table representation of a bipartite graph

Source

<https://www.barkbeetles.info/>

bootstrap_phylopred_model

Bootstrap a phylopred logistic regression model

Description

Runs the phylopred logistic regression n times using random focal host sampling, returning a distribution of coefficient estimates.

Usage

```
bootstrap_phylopred_model(incidence, phydist, n = 1000, seed = NULL)
```

Arguments

<code>incidence</code>	A binary incidence matrix (rows = parasites, cols = hosts).
<code>phydist</code>	A square, numeric phylogenetic distance matrix among hosts.
<code>n</code>	Integer. Number of bootstrap iterations. Default 1000.
<code>seed</code>	Integer or NULL. Random seed for reproducibility. Default NULL.

Value

An S3 object of class "phylopred_bootstrap", a list with:

call The matched call.

n Number of bootstrap iterations.

seed The seed used.

coefficients A matrix of n rows x 12 columns (coefficient info from each iteration).

summary A data.frame with the column-wise means (same 12 columns).

metadata List with `n_parasites` and `n_hosts`.

See Also

[fit_phylopred_model](#), [predict_phylopred_probability](#)

Examples

```
incidence <- prepare_incidence_matrix(beetleTreeInteractions)
aligned   <- align_phylopred_inputs(incidence, phy_dist)
bootstrap_phylopred_model(aligned$incidence, aligned$phydist, n = 5, seed = 42)
```

cluster_bootstrap_phylopred

Cluster bootstrap for pairwise host-sharing models

Description

Nonparametric bootstrap that resamples whole parasites (clusters) with replacement and refits the logistic regression on each replicate. This respects the non-independence of rows that share a parasite in a [prepare_pair_data](#) data set, unlike a naive row-level bootstrap. When a group column is present, resampling is stratified within groups and the model includes a group-by-distance interaction.

Usage

```
cluster_bootstrap_phylopred(
  pairs,
  n_boot = 1000,
  seed = NULL,
  conf_level = 0.95
)
```

Arguments

<code>pairs</code>	A data.frame from prepare_pair_data .
<code>n_boot</code>	Integer. Number of bootstrap replicates. Default 1000.
<code>seed</code>	Integer or NULL. Random seed for reproducibility. Default NULL.
<code>conf_level</code>	Confidence level for percentile intervals. Default 0.95.

Value

An object of class "phylopred_cluster_bootstrap", a list with:

draws Matrix of bootstrap coefficient draws (replicates x terms).

estimates Coefficients of the model fitted to the full data.

confidence_intervals Percentile confidence intervals per term.

n_boot, seed, conf_level Bootstrap settings.

See Also

[prepare_pair_data](#), [compare_phylopred_slopes](#)

Examples

```
# Three parasites keep the example fast
top <- names(sort(table(beetleTreeInteractions[[1]]),
                      decreasing = TRUE))[1:3]
sub <- beetleTreeInteractions[
  as.character(beetleTreeInteractions[[1]]) %in% top, ]
pairs <- prepare_pair_data(sub, phy_dist)
cluster_bootstrap_phylopred(pairs, n_boot = 5, seed = 42)
```

`compare_phylopred_slopes`

Compare host-sharing slopes between two clades

Description

Fits a single logistic regression with a group-by-phylogenetic-distance interaction to a pairwise data set built with [prepare_pair_data](#) and tests whether the slopes of the two groups differ. A more negative slope means the probability of sharing a parasite decays faster with phylogenetic distance, i.e. the clade is more of a phylogenetic specialist. Uncertainty for the slope difference comes from a parasite-level cluster bootstrap.

Usage

```
compare_phylopred_slopes(pairs, n_boot = 1000, seed = NULL, conf_level = 0.95)
```

Arguments

<code>pairs</code>	A data.frame from prepare_pair_data with a group column containing exactly two groups.
<code>n_boot</code>	Integer. Number of cluster bootstrap replicates. Default 1000.
<code>seed</code>	Integer or NULL. Random seed for reproducibility. Default NULL.
<code>conf_level</code>	Confidence level. Default 0.95.

Value

An object of class "phylopred_slope_comparison", a list with:

groups The two group labels (reference first).

slopes Named vector with the slope of each group.

slope_difference Slope of the second group minus the reference.

interaction_p_value Wald p-value of the interaction term.

bootstrap The [cluster_bootstrap_phylopred](#) object.

slope_cis Bootstrap percentile CIs for each group's slope and for the difference.

fitted_model The underlying glm object.

See Also

[prepare_pair_data](#), [cluster_bootstrap_phylopred](#)

Examples

```
# Compare Xyleborus beetles against other genera (subset kept small
# so the bootstrap example runs quickly)
parasites <- unique(as.character(beetleTreeInteractions[[1]]))
keep <- c(grep("^Xyleborus", parasites, value = TRUE)[1:3],
         parasites[!grepl("^Xyleborus", parasites)][1:3])
sub <- beetleTreeInteractions[
  as.character(beetleTreeInteractions[[1]]) %in% keep, ]
pairs <- prepare_pair_data(
  sub, phy_dist,
  group = function(x) ifelse(grepl("^Xyleborus", x), "Xyleborus", "other")
)
compare_phylopred_slopes(pairs, n_boot = 5, seed = 42)
```

evaluate_phylopred_model

Evaluate the predictive capacity of a pairwise host-sharing model

Description

Computes goodness-of-fit and discrimination metrics for a logistic host-sharing model fitted to a [prepare_pair_data](#) data set: McFadden and Tjur pseudo-R-squared, the area under the ROC curve (AUC), and optionally a grouped cross-validated AUC where whole parasites are held out (parasite-grouped k-fold), which measures the ability to predict the host range of unseen parasites.

Usage

```
evaluate_phylopred_model(pairs, cv = FALSE, k = 5, seed = NULL)
```

Arguments

<code>pairs</code>	A data.frame from prepare_pair_data .
<code>cv</code>	Logical. If TRUE, also compute the parasite-grouped cross-validated AUC. Default FALSE (refitting can be slow on large data sets).
<code>k</code>	Integer. Number of cross-validation folds. Default 5.
<code>seed</code>	Integer or NULL. Random seed for fold assignment. Default NULL.

Value

An object of class "phylopred_evaluation", a list with:

mcFadden_r2 McFadden pseudo-R-squared, $1 - \logLik(model)/\logLik(null)$.

tjur_r2 Tjur coefficient of discrimination (mean fitted probability for 1s minus mean for 0s).

auc In-sample area under the ROC curve.

cv_auc Parasite-grouped k-fold cross-validated AUC (NA if cv = FALSE).

fitted_model The underlying glm object.

n_pairs, n_parasites Data set size.

See Also

[prepare_pair_data](#), [cluster_bootstrap_phylopred](#)

Examples

```
# Three parasites keep the example fast
top <- names(sort(table(beetleTreeInteractions[[1]]),
                     decreasing = TRUE))[1:3]
sub <- beetleTreeInteractions[
  as.character(beetleTreeInteractions[[1]]) %in% top, ]
pairs <- prepare_pair_data(sub, phy_dist)
evaluate_phylopred_model(pairs)
evaluate_phylopred_model(pairs, cv = TRUE, k = 3, seed = 42)
```

`extract_phylopred_coefficients`*Extract coefficients from a phylopred model or bootstrap object*

Description

Returns a named list of model coefficients and their associated statistics from a `phylopred_model` or `phylopred_bootstrap` object.

Usage

```
extract_phylopred_coefficients(model)
```

Arguments

`model` A `phylopred_model` or `phylopred_bootstrap` object.

Value

A named list with elements:

intercept Intercept coefficient (mean for bootstrap).

slope Slope coefficient (mean for bootstrap).

intercept_se Standard error of intercept.

slope_se Standard error of slope.

intercept_ci Two-element vector with 2.5% and 97.5% CI for intercept.

slope_ci Two-element vector with 2.5% and 97.5% CI for slope.

Examples

```
inc <- prepare_incidence_matrix(beetleTreeInteractions)
al  <- align_phylopred_inputs(inc, phy_dist)
model <- fit_phylopred_model(al$incidence, al$phydist, seed = 42)
extract_phylopred_coefficients(model)
```

fish_tree	<i>Phylogenetic tree of marine fish hosts of Caligus and Lepeophtheirus</i>
-----------	---

Description

A phylogenetic tree (class phylo) of 625 marine fish species that are recorded hosts of sea lice of the genera *Caligus* and *Lepeophtheirus* in the cofid copepod-fish interaction database. The tree is pruned from the Fish Tree of Life actinopterygian phylogeny (Rabosky and colleagues 2018, actinopt_12k_raxml). Tip labels are binomial species names with spaces.

Usage

```
fish_tree
```

Format

A phylo object with 625 tips.

Source

<https://fishtreeoflife.org/>. The underlying data are released under the CC0 1.0 public domain dedication (Dryad, [doi:10.5061/dryad.fc71cp4](https://doi.org/10.5061/dryad.fc71cp4)). Please cite Rabosky, D. L. and colleagues (2018) An inverse latitudinal gradient in speciation rate for marine fishes. Nature 559, 392-395. [doi:10.1038/s4158601802731](https://doi.org/10.1038/s4158601802731)

See Also

The cofid package for the interaction records.

fit_phylopred_model	<i>Fit a phylopred logistic regression model</i>
---------------------	--

Description

Fits a logistic regression model relating host susceptibility to phylogenetic distance.

Usage

```
fit_phylopred_model(incidence, phydist, seed = NULL)
```

Arguments

incidence	A binary incidence matrix (rows = parasites/pathogens, cols = hosts).
phydist	A square, numeric phylogenetic distance matrix among hosts. Row/col names must match the column names of incidence.
seed	Integer or NULL. Random seed for reproducibility. Default NULL.

Value

An S3 object of class "phylopred_model", a list with:

call The matched call.

coefficients Named numeric vector c(intercept, slope).

standard_errors Named numeric vector c(intercept, slope).

confidence_intervals Matrix with rows c("intercept", "slope") and cols c("2.5%", "97.5%").

fitted_model The underlying glm object.

training_data The data.frame used for model fitting.

convergence Logical; whether the GLM converged.

metadata List with n_parasites, n_hosts, and seed.

See Also

[bootstrap_phylopred_model](#), [predict_phylopred_probability](#)

Examples

```
incidence <- prepare_incidence_matrix(beetleTreeInteractions)
aligned   <- align_phylopred_inputs(incidence, phy_dist)
fit_phylopred_model(aligned$incidence, aligned$phydist, seed = 42)
```

fit_phylopred_pu

Positive-unlabeled host prediction via xplus (PLUS algorithm)

Description

Predicts the hosts of a parasite from the phylogeny using positive-unlabeled (PU) learning, as a parallel alternative to the pairwise logistic model. Known hosts are treated as positives and every other species in the phylogeny as unlabeled (never as a true absence). Each species is described by phylogenetic features and the PLUS algorithm (xplus::xplus) iteratively re-labels the unlabeled species while fitting a penalized logistic model.

Usage

```
fit_phylopred_pu(
  interactions,
  phydist,
  parasite,
  n_eigen = 10,
  seed = NULL,
  ...
)
```

Arguments

<code>interactions</code>	A data.frame whose first two columns are the parasite and host species names (as in prepare_pair_data).
<code>phydist</code>	Square numeric matrix of phylogenetic distances with species names as row and column names.
<code>parasite</code>	Character scalar: the parasite whose hosts are predicted.
<code>n_eigen</code>	Number of PCoA axes used as features. Default 10.
<code>seed</code>	Optional random seed passed to <code>xplus::xplus</code> .
<code>...</code>	Further arguments passed to <code>xplus::xplus</code> (e.g. <code>max_iter</code> , <code>learning_rate</code> , <code>qq</code>).

Details

The feature matrix contains, for every species in `phydist`:

- the minimum and mean phylogenetic distance to the parasite's known hosts (excluding the species itself), and
- the first `n_eigen` principal coordinates (PCoA) of the phylogenetic distance matrix, describing the global position of the species in the phylogeny (the analogue of environmental layers in an SDM).

Value

An object of class "phylopred_pu": a list with

prediction data.frame with species, pred (predicted host suitability) and is_host.

fit The underlying `xplus` object.

parasite The parasite name.

The prediction columns can be passed directly to [host_threshold_metric](#) and [phylo_partial_roc](#).

See Also

[host_threshold_metric](#), [phylo_partial_roc](#), [prepare_pair_data](#)

Examples

```
pu <- fit_phylopred_pu(beetleTreeInteractions, phy_dist,
                      parasite = beetleTreeInteractions[[1]][1],
                      max_iter = 20, seed = 1)
host_threshold_metric(pu$prediction$pred, pu$prediction$is_host)
```

 host_specificity_index

Host specificity index (S_{TD}) for every parasite in an interaction table

Description

Applies `poulin_std` to every parasite in an interaction table, following the same expansion logic as `prepare_pair_data`. Useful for contrasting the Poulin and Mouillot (2003) specificity index against phylopred's model-based host-range metric b^* (`host_threshold_metric`) across many parasites.

Usage

```
host_specificity_index(interactions, phydist, group = NULL)
```

Arguments

<code>interactions</code>	A data.frame with at least two columns: the first the parasite (or symbiont) species, the second the host species. Extra columns are ignored. Duplicated rows are removed.
<code>phydist</code>	A square, numeric distance matrix among hosts (taxonomic steps or phylogenetic distances), with matching row and column names.
<code>group</code>	Optional. Either the name of a column in <code>interactions</code> giving a grouping factor for each parasite (e.g. genus or clade), or a function applied to the parasite name to derive the group.

Value

A data.frame of class "phylopred_specificity" with columns `parasite`, `n_hosts` (known hosts present in `phydist`), `std` (S_{TD}), `var_std` ($Var(S_{TD})$), and `group` (only if `group` is supplied).

References

Poulin, R. and Mouillot, D. (2003) Parasite specialization from a phylogenetic perspective: a new index of host specificity. *Parasitology* 126, 473-480. doi:10.1017/S0031182003002993

See Also

`poulin_std`, `prepare_pair_data`, `host_threshold_metric`

Examples

```
std_by_parasite <- host_specificity_index(beetleTreeInteractions, phy_dist)
head(std_by_parasite)
```

host_threshold_metric *Threshold metric f_γ for predicted host suitability*

Description

Computes the threshold-dependent performance metric

$$f_\gamma(t) = a(t)/(\gamma b(t) + c(t))$$

where, for a cutoff t on the predicted host-suitability scores:

$a(t)$ proportion of known hosts with prediction $\geq t$ (sensitivity / recall);

$b(t)$ proportion of ALL species in the phylogeny with prediction $\geq t$ (predicted host breadth, the phylogenetic analogue of predicted area in SDM threshold metrics);

$c(t) = 1 - a(t)$ proportion of known hosts lost (omission).

The metric is designed for presence-only / positive-unlabeled settings: it never uses absences, and γ tunes the trade-off between retaining known hosts and predicting a compact host set. The maximum of $f_\gamma(t)$ gives an optimal cutoff t^* .

Usage

```
host_threshold_metric(pred, is_host, gamma = 1)
```

Arguments

pred	Numeric vector of predicted suitability scores, one per species in the phylogeny (any real values; typically probabilities).
is_host	Logical or 0/1 vector of the same length: TRUE/1 for species that are known hosts (presences).
gamma	Positive scalar weight on the predicted breadth term. Default 1.

Value

An object of class "phylopred_threshold_metric": a data.frame with columns threshold, a, b, c, and f_gamma, evaluated at every distinct prediction value. The optimal cutoff is stored in attributes "optimal_threshold" and "optimal_f", and gamma in attribute "gamma".

See Also

[phylo_partial_roc](#), [evaluate_phylopred_model](#)

Examples

```

pairs <- prepare_pair_data(beetleTreeInteractions, phy_dist)
fit <- stats::glm(suscept ~ phydist, data = pairs, family = stats::binomial())
# Suitability of every host species for one focal parasite
one <- pairs[pairs$parasite == pairs$parasite[1], ]
one <- one[!duplicated(one$target), ]
pred <- stats::predict(fit, newdata = one, type = "response")
m <- host_threshold_metric(pred, one$suscept, gamma = 1)
plot(m)

```

phylopred_confidence_ellipse

Joint confidence ellipse for intercept and slope

Description

Computes the boundary of the joint confidence region ("ellipse") for a pair of coefficients, either from a fitted glm (using the asymptotic covariance matrix) or from a [cluster_bootstrap_phylopred](#) object (using the empirical covariance of the bootstrap draws, which respects parasite-level clustering).

Usage

```

phylopred_confidence_ellipse(
  object,
  terms = c("(Intercept)", "phydist"),
  level = 0.95,
  n_points = 200
)

```

Arguments

object	A glm fit or a phylopred_cluster_bootstrap object.
terms	Character vector of length 2 with the coefficient names. Default c("(Intercept)", "phydist").
level	Confidence level. Default 0.95.
n_points	Number of points on the ellipse boundary. Default 200.

Value

A data.frame with n_points rows and two columns named after terms, tracing the ellipse boundary. The center (point estimates) is returned as the "center" attribute.

See Also

[cluster_bootstrap_phylopred](#)

Examples

```
# Three parasites keep the example fast
top <- names(sort(table(beetleTreeInteractions[[1]]),
                      decreasing = TRUE))[1:3]
sub <- beetleTreeInteractions[
  as.character(beetleTreeInteractions[[1]]) %in% top, ]
pairs <- prepare_pair_data(sub, phy_dist)

# From a glm: asymptotic ellipse
fit <- glm(suscept ~ phydist, data = pairs, family = binomial())
ell <- phylopred_confidence_ellipse(fit)
plot(ell, type = "l")
points(t(attr(ell, "center")), pch = 19)

# From a cluster bootstrap
boot <- cluster_bootstrap_phylopred(pairs, n_boot = 5, seed = 42)
ell_b <- phylopred_confidence_ellipse(boot)
lines(ell_b, lty = 2)
```

phylo_partial_roc	<i>Phylogenetic partial ROC (Peterson, Papes & Soberon 2008 analogue)</i>
-------------------	---

Description

Partial ROC test for presence-only host-suitability predictions, following Peterson, Papes & Soberon (2008, Ecological Modelling 213, 63-72). The ROC space is redefined for presence-only data: the y axis is the sensitivity $a(t)$ (proportion of known hosts predicted at cutoff t) and the x axis is the proportion of ALL species in the phylogeny predicted as hosts, $b(t)$ (the analogue of proportional predicted area). The curve is restricted to the low-omission region $a(t) \geq 1 - E$ for a user-chosen omission tolerance error_rate E , and performance is summarized as the ratio of the partial AUC of the observed curve to the partial AUC of the null (random-ranking) expectation, the diagonal $a = b$. A ratio of 1 equals random performance; ratios above 1 indicate predictive signal. Significance is assessed by bootstrap resampling of the known hosts.

Usage

```
phylo_partial_roc(
  pred,
  is_host,
  error_rate = 0.05,
  n_boot = 500,
  boot_prop = 0.5,
  seed = NULL
)
```

Arguments

<code>pred</code>	Numeric vector of predicted suitability scores, one per species in the phylogeny.
<code>is_host</code>	Logical or 0/1 vector of the same length: TRUE/1 for known hosts.
<code>error_rate</code>	Allowed omission error E in $[0, 1)$. The partial AUC is computed over the region with sensitivity $\geq 1 - E$. Default 0.05.
<code>n_boot</code>	Number of bootstrap replicates. Default 500.
<code>boot_prop</code>	Proportion of known hosts resampled (with replacement) in each bootstrap replicate, as in Peterson, Papes and Soberon (2008). Default 0.5.
<code>seed</code>	Integer or NULL. Random seed for the bootstrap. Default NULL.

Value

An object of class "phylopred_partial_roc", a list with:

auc_ratio Observed partial AUC ratio (observed / null).

p_value Bootstrap proportion of replicates with ratio ≤ 1 .

boot_ratios Vector of bootstrap AUC ratios.

curve data.frame with threshold, a (sensitivity) and b (predicted breadth).

error_rate, n_boot, boot_prop The configuration used.

See Also

[host_threshold_metric](#)

Examples

```
pairs <- prepare_pair_data(beetleTreeInteractions, phy_dist)
fit <- stats::glm(suscept ~ phydist, data = pairs, family = stats::binomial())
one <- pairs[pairs$parasite == pairs$parasite[1], ]
one <- one[!duplicated(one$target), ]
pred <- stats::predict(fit, newdata = one, type = "response")
phylo_partial_roc(pred, one$suscept, n_boot = 50, seed = 42)
```

phy_dist

Phylogenetic distance matrix among host data on phylopred

Description

Data of phylogenetic distance matrix for bark beetles host species

Usage

phy_dist

Format

A matrix with 235 rows and 235 columns. Contain Phylogenetic distance matrix between bark beetle host genera

Source

<https://www.barkbeetles.info/>

poulin_std	<i>Poulin and Mouillot taxonomic/phylogenetic distinctness index (S_{TD})</i>
------------	--

Description

Computes the average taxonomic (or phylogenetic) distinctness S_{TD} of Poulin and Mouillot (2003) for a single parasite's known host set:

$$S_{TD} = \frac{2 \sum_{i < j} \omega_{ij}}{s(s-1)}$$

where ω_{ij} is the distance between hosts i and j taken from phydist, and s is the number of known hosts. When phydist is a taxonomic step matrix (see [taxonomic_step_matrix](#)), S_{TD} is the original 1-5 taxonomic index; when phydist is a matrix of phylogenetic (patristic) distances, this is the continuous phylogenetic generalization noted as possible by Poulin and Mouillot (2003, after Clarke and Warwick, 2001), expressed directly in branch-length units. The corresponding sampling variance $Var(S_{TD}) = \sum_{i \neq j} (\omega_{ij} - S_{TD})^2 / [s(s-1)]$ is also returned.

Usage

```
poulin_std(phydist, hosts)
```

Arguments

phydist	A square, numeric distance matrix among hosts with matching row and column names (taxonomic steps from taxonomic_step_matrix , or phylogenetic/patristic distances such as <code>phy_dist</code>).
hosts	Character vector of known host names for one parasite. Hosts not present in phydist are dropped with a warning.

Value

An object of class "phylopred_std": a list with elements `n_hosts`, `std` (S_{TD}), and `var_std` ($Var(S_{TD})$, NA when `n_hosts` < 3). S_{TD} is undefined (NA, with a warning) when `n_hosts` < 2, as noted by Poulin and Mouillot (2003) for single-host parasites.

References

- Poulin, R. and Mouillot, D. (2003) Parasite specialization from a phylogenetic perspective: a new index of host specificity. *Parasitology* 126, 473-480. doi:10.1017/S0031182003002993
- Clarke, K.R. and Warwick, R.M. (2001) A further biodiversity index applicable to species lists: variation in taxonomic distinctness. *Marine Ecology Progress Series* 216, 265-278.

See Also

[host_specificity_index](#), [taxonomic_step_matrix](#), [host_threshold_metric](#)

Examples

```
hosts <- unique(beetleTreeInteractions$genus[
  beetleTreeInteractions$beetle == beetleTreeInteractions$beetle[1]])
poulin_std(phy_dist, hosts)
```

predict_phylopred_probability

Predict host susceptibility probabilities from logistic regression coefficients

Description

Applies a logit function using intercept and slope coefficients to a vector of phylogenetic distances. Accepts a numeric vector, a phylopred_model, or a phylopred_bootstrap object.

Usage

```
predict_phylopred_probability(coef, dist)
```

Arguments

coef	A named numeric vector with elements "intercept" and "slope", OR a phylopred_model object, OR a phylopred_bootstrap object (in which case the bootstrap summary means are used). An unnamed vector of length ≥ 2 is interpreted as c(intercept, slope).
dist	A numeric vector of phylogenetic distances.

Value

A named numeric vector of predicted probabilities of the same length as dist.

See Also

[fit_phylopred_model](#)

Examples

```
inc <- prepare_incidence_matrix(beetleTreeInteractions)
al <- align_phylopred_inputs(inc, phy_dist)
model <- fit_phylopred_model(al$incidence, al$phydist, seed = 42)
dist_vec <- seq(0, max(al$phydist), length.out = 50)
predict_phylopred_probability(model, dist_vec)
```

```
prepare_incidence_matrix
```

Prepare an incidence matrix from an interaction table

Description

Builds an incidence (presence/absence) matrix from a two-column data.frame of species interactions.

Usage

```
prepare_incidence_matrix(db)
```

Arguments

db A data.frame with exactly two columns (e.g., parasite and host).

Value

An incidence matrix (rows = first column unique values, cols = second column unique values) filled with 0/1.

Examples

```
prepare_incidence_matrix(beetleTreeInteractions)
```

```
prepare_pair_data
```

Build a pairwise host-sharing data set from an interaction table

Description

Expands an interaction table into all (focal host, target host) pairs per parasite. For every parasite, each of its known hosts is used as a focal host, and every other host in phydist is a target host whose response is 1 if it is also a known host of that parasite and 0 otherwise.

Usage

```
prepare_pair_data(interactions, phydist, group = NULL)
```

Arguments

interactions	A data.frame with at least two columns: the first the parasite (or symbiont) species, the second the host species. Extra columns are ignored. Duplicated rows are removed.
phydist	A square, numeric phylogenetic (or taxonomic) distance matrix among hosts, with matching row and column names. Hosts in interactions absent from phydist are dropped with a warning.
group	Optional. Either the name of a column in interactions giving a grouping factor for each parasite (e.g. genus or clade), or a function applied to the parasite name to derive the group (e.g. <code>function(x) sub(".*", "", x)</code> for the genus).

Details

Unlike the legacy workflow (which samples a single focal host per parasite), this uses every known host as a focal host, so no information is discarded and results are deterministic. Rows belonging to the same parasite are not independent; account for this with [cluster_bootstrap_phylopred](#) when computing uncertainty.

Value

A data.frame of class "phylopred_pairs" with columns:

- parasite** Parasite species (cluster identifier).
- focal** Focal (known) host.
- target** Target host.
- phydist** Distance between focal and target host.
- suscept** 1 if the target host is a known host of the parasite.
- group** Grouping factor (only if group is supplied).

See Also

[fit_phylopred_model](#), [cluster_bootstrap_phylopred](#), [compare_phylopred_slopes](#)

Examples

```
pairs <- prepare_pair_data(beetleTreeInteractions, phy_dist)
head(pairs)
```

taxonomic_step_matrix *Build a taxonomic step-distance matrix (Poulin & Mouillot scale)*

Description

Constructs a symmetric taxonomic distance matrix among species from a taxonomic hierarchy table, following the step scale of Poulin and Mouillot (2003): 1 if two species share a genus, 2 if they share a family but not a genus, 3 if they share an order but not a family, 4 if they share a class but not an order, and 5 if they share none of the above (different class / phylum). Any subset of genus, family, order, class may be supplied; missing ranks are skipped and the maximum step for that pair falls back to $\text{length}(\text{ranks}) + 1$.

Usage

```
taxonomic_step_matrix(tax_table)
```

Arguments

tax_table A data.frame with one row per species and columns species plus at least two of genus, family, order, class. No NA values are allowed in the rank columns (see [host_specificity_index](#) for handling hosts with incomplete taxonomy).

Value

A square numeric matrix with row/column names equal to `tax_table$species`, diagonal 0, suitable as the `phydist` argument of [poulin_std](#) or [host_specificity_index](#).

References

Poulin, R. and Mouillot, D. (2003) Parasite specialization from a phylogenetic perspective: a new index of host specificity. *Parasitology* 126, 473-480. doi:[10.1017/S0031182003002993](https://doi.org/10.1017/S0031182003002993)

See Also

[poulin_std](#), [host_specificity_index](#)

Examples

```
tax <- data.frame(
  species = c("Caligus elongatus", "Caligus curtus",
              "Lepeophtheirus salmonis", "Ergasilus sieboldi"),
  genus   = c("Caligus", "Caligus", "Lepeophtheirus", "Ergasilus"),
  family  = c("Caligidae", "Caligidae", "Caligidae", "Ergasilidae"),
  order   = c("Siphonostomatoida", "Siphonostomatoida",
              "Siphonostomatoida", "Cyclopoida")
)
taxonomic_step_matrix(tax)
```

 validate_host_prediction

Leave-hosts-out validation of host predictions

Description

Presence-only predictive validation: a fraction of the parasite's known hosts is hidden, the model is refitted on the remaining hosts, and the hidden hosts are scored against the unlabeled species. Because true absences do not exist, performance is measured as the AUC of the hidden (held-out) hosts versus the unlabeled species — the probability that a truly interacting species outranks a random unlabeled one. This measures actual predictive ability, not in-sample fit.

Usage

```
validate_host_prediction(
  interactions,
  phydist,
  parasite,
  method = c("glm", "pu"),
  prop_holdout = 0.25,
  n_rep = 10,
  seed = NULL,
  ...
)
```

Arguments

<code>interactions</code>	A data.frame whose first two columns are the parasite and host species names.
<code>phydist</code>	Square numeric matrix of phylogenetic distances with species names as row and column names.
<code>parasite</code>	Character scalar: the parasite whose hosts are predicted.
<code>method</code>	"glm" for the pairwise logistic model (target species scored by their maximum predicted sharing probability across the training focal hosts) or "pu" for the xplus PU-learning model (fit_phylopred_pu).
<code>prop_holdout</code>	Proportion of known hosts hidden in each replicate. Default 0.25.
<code>n_rep</code>	Number of holdout replicates. Default 10.
<code>seed</code>	Optional random seed.
<code>...</code>	Further arguments passed to fit_phylopred_pu when method = "pu".

Value

An object of class "phylopred_holdout": a list with auc (vector of holdout AUCs, one per replicate), mean_auc, method, parasite, prop_holdout, and n_rep.

Index

* datasets

- beetleTreeInteractions, [3](#)
- fish_tree, [9](#)
- phy_dist, [16](#)

align_phylopred_inputs, [2](#)

beetleTreeInteractions, [3](#)

bootstrap_phylopred_model, [3](#), [10](#)

cluster_bootstrap_phylopred, [4](#), [6](#), [7](#), [14](#),
[20](#)

compare_phylopred_slopes, [5](#), [5](#), [20](#)

evaluate_phylopred_model, [7](#), [13](#)

extract_phylopred_coefficients, [8](#)

fish_tree, [9](#)

fit_phylopred_model, [4](#), [9](#), [18](#), [20](#)

fit_phylopred_pu, [10](#), [22](#), [23](#)

host_specificity_index, [12](#), [18](#), [21](#)

host_threshold_metric, [11](#), [12](#), [13](#), [16](#), [18](#),
[23](#)

phy_dist, [16](#)

phylo_partial_roc, [11](#), [13](#), [15](#), [23](#)

phylopred_confidence_ellipse, [14](#)

poulin_std, [12](#), [17](#), [21](#)

predict_phylopred_probability, [4](#), [10](#), [18](#)

prepare_incidence_matrix, [19](#)

prepare_pair_data, [4-7](#), [11](#), [12](#), [19](#)

taxonomic_step_matrix, [17](#), [18](#), [21](#)

validate_host_prediction, [22](#)