Package 'castor'

March 11, 2021

Type Package

Title Efficient Phylogenetics on Large Trees

Version 1.6.7

Date 2021-03-10

Author Stilianos Louca

Maintainer Stilianos Louca <louca@zoology.ubc.ca>

Description Efficient phylogenetic analyses on massive phylogenies comprising up to millions of tips. Functions include pruning, rerooting, calculation of most-recent common ancestors, calculating distances from the tree root and calculating pairwise distances. Calculation of phylogenetic signal and mean trait depth (trait conservatism), ancestral state reconstruction and hidden character prediction of discrete characters, simulating and fitting models of trait evolution, fitting and simulating diversification models, dating trees, comparing trees, and reading/writing trees in Newick format. Citation: Louca, Stilianos and Doebeli, Michael (2017) <doi:10.1093/bioinformatics/btx701>.

License GPL (>= 2)

Depends Rcpp (>= 0.12.10)

Imports parallel, naturalsort, stats, nloptr, Matrix, RSpectra

SystemRequirements C++11

LinkingTo Rcpp

RoxygenNote 7.1.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2021-03-11 06:20:17 UTC

R topics documented:

| castor-package | *Efficient computations on large phylogenetic trees.* |

Description

This package provides efficient tree manipulation functions including pruning, rerooting, calculation of most-recent common ancestors, calculating distances from the tree root and calculating pairwise distance matrices. Calculation of phylogenetic signal and mean trait depth (trait conservatism). Efficient ancestral state reconstruction and hidden character prediction of discrete characters on phylogenetic trees, using Maximum Likelihood and Maximum Parsimony methods. Simulating models of trait evolution, and generating random trees.

Details

The most important data unit is a phylogenetic tree of class "phylo", with the tree topology encoded in the member variable tree.edge. See the ape package manual for details on the "phylo" format. The castor package was designed to be efficient for large phylogenetic trees (>10,000 tips), and scales well to trees with millions of tips. Most functions have asymptotically linear time complexity O(N) in the number of edges N. This efficiency is achived via temporary auxiliary data structures, use of dynamic programing, heavy use of C++, and integer-based indexing instead of name-based indexing of arrays. All functions support trees that include monofurcations (nodes with a single child) as well as multifurcations (nodes with more than 2 children). See the associated paper by Louca et al. for a comparison with other packages.

Throughout this manual, "Ntips" refers to the number of tips, "Nnodes" to the number of nodes and "Nedges" to the number of edges in a tree. In the context of discrete trait evolution/reconstruction, "Nstates" refers to the number of possible states of the trait. In the context of multivariate trait evolution, "Ntraits" refers to the number of traits.

Author(s)

Stilianos Louca

Maintainer: Stilianos Louca <louca@zoology.ubc.ca>

References

S. Louca and M. Doebeli (2017). Efficient comparative phylogenetics on large trees. Bioinformatics. DOI:10.1093/bioinformatics/btx701

---

asr_empirical_probabilities

Empirical ancestral state probabilities.

Description

Given a rooted phylogenetic tree and the states of a discrete trait for each tip, calculate the empirical state frequencies/probabilities for each node in the tree, i.e. the frequencies/probabilities of states across all tips descending from that node. This may be used as a very crude estimate of ancestral state probabilities.

Usage

```
asr_empirical_probabilities(tree, tip_states, Nstates=NULL,
                            probabilities=TRUE, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	An integer vector of size Ntips, specifying the state of each tip in the tree as an integer from 1 to Nstates, where Nstates is the possible number of states (see below).
Nstates	Either NULL, or an integer specifying the number of possible states of the trait. If NULL, then it will be computed based on the maximum value encountered in tip_states
probabilities	Logical, specifying whether empirical frequencies should be normalized to represent probabilities. If FALSE, then the raw occurrence counts are returned.
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

For this function, the trait's states must be represented by integers within 1,..,Nstates, where Nstates is the total number of possible states. If the states are originally in some other format (e.g., characters or factors), you should map them to a set of integers 1,..,Nstates. You can easily map any set of discrete states to integers using the function map_to_state_space.

The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). The function has asymptotic time complexity O(Nedges x Nstates).

Tips must be represented in `tip_states` in the same order as in `tree$tip.label`. The vector `tip_states` need not include names; if it does, however, they are checked for consistency (if `check_input==TRUE`).

Value

A list with the following elements:

`ancestral_likelihoods`

A 2D integer (if `probabilities==FALSE`) or numeric (if `probabilities==TRUE`) matrix, listing the frequency or probability of each state for each node. This matrix will have size Nnodes x Nstates, where Nstates was either explicitly provided as an argument or inferred from `tip_states`. The rows in this matrix will be in the order in which nodes are indexed in the tree, i.e. the [n,s]-th entry will be the frequency or probability of the s-th state for the n-th node. Note that the name was chosen for compatibility with other ASR functions.

Author(s)

Stilianos Louca

See Also

[asr_max_parsimony](), [asr_squared_change_parsimony]() [asr_mk_model](), [map_to_state_space]()

Examples

```
## Not run:
# generate a random tree
Ntips = 100
tree  = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# create a random transition matrix
Nstates = 3
Q = get_random_mk_transition_matrix(Nstates, rate_model="ER", max_rate=0.01)
cat(sprintf("Simulated ER transition rate=%g\n",Q[1,2]))

# simulate the trait's evolution
simulation = simulate_mk_model(tree, Q)
tip_states = simulation$tip_states

# calculate empirical probabilities of tip states
asr_empirical_probabilities(tree, tip_states=tip_states, Nstates=Nstates)

## End(Not run)
```

```
asr_independent_contrasts
```

Ancestral state reconstruction via phylogenetic independent contrasts.

Description

Reconstruct ancestral states for a continuous (numeric) trait using phylogenetic independent contrasts (PIC; Felsenstein, 1985).

Usage

```
asr_independent_contrasts(tree,
                          tip_states,
                          weighted    = TRUE,
                          include_CI  = FALSE,
                          check_input = TRUE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

tip_states   A numeric vector of size Ntips, specifying the known state of each tip in the tree.

weighted   Logical, specifying whether to weight tips and nodes by the inverse length of their incoming edge, as in the original method by Felsenstein (1985). If FALSE, edge lengths are treated as if they were 1.

include_CI   Logical, specifying whether to also calculate standard errors and confidence intervals for the reconstructed states under a Brownian motion model, as described by Garland et al (1999).

check_input   Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

The function traverses the tree in postorder (tips–>root) and estimates the state of each node as a convex combination of the estimated states of its chilren. These estimates are the intermediate "X" variables introduced by Felsenstein (1985) in his phylogenetic independent contrasts method. For the root, this yields the same globally parsimonious state as the squared-changes parsimony algorithm implemented in `asr_squared_change_parsimony` (Maddison 1991). For any other node, PIC only yields locally parsimonious reconstructions, i.e. reconstructed states only depend on the subtree descending from the node (see discussion by Maddison 1991).

If `weighted==TRUE`, then this function yields the same ancestral state reconstructions as

`ape::ace(phy=tree,x=tip_states,type="continuous",method="pic",model="BM",CI=FALSE)`

in the ape package (v. 0.5-64). Note that in contrast to the CI95 returned by ape::ace, the confidence intervals calculated here have the same units as the trait and depend both on the tree topology as well as the tip states.

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. This is the same as setting weighted=FALSE. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). Edges with length 0 will be adjusted internally to some tiny length if needed (if weighted==TRUE).

Tips must be represented in tip_states in the same order as in tree$tip.label. The vector tip_states need not include item names; if it does, however, they are checked for consistency (if check_input==TRUE). All tip states must be non-NA; otherwise, consider using one of the functions for hidden-state-prediction (e.g., hsp_independent_contrasts).

The function has asymptotic time complexity O(Nedges).

Value

A list with the following elements:

ancestral_states

> A numeric vector of size Nnodes, listing the reconstructed state of each node. The entries in this vector will be in the order in which nodes are indexed in the tree.

standard_errors

> Numeric vector of size Nnodes, listing the phylogenetically estimated standard error for the state in each node, under a Brownian motion model. The standard errors have the same units as the trait and depend both on the tree topology as well as the tip states. Calculated as described by Garland et al. (1999, page 377). Only included if include_CI==TRUE.

CI95

> Numeric vector of size Nnodes, listing the radius (half width) of the 95% confidence interval of the state in each node. Confidence intervals have same units as the trait and depend both on the tree topology as well as the tip states. For each node, the confidence interval is calculated according to the Student's t-distribution with Npics degrees of freedom, where Npics is the number of internally calculated independent contrasts descending from the node [Garland et al, 1999]. Only included if include_CI==TRUE.

Author(s)

Stilianos Louca

References

J. Felsenstein (1985). Phylogenies and the Comparative Method. The American Naturalist. 125:1-15.

W. P. Maddison (1991). Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Systematic Zoology. 40:304-314.

T. Garland Jr., P. E. Midford, A. R. Ives (1999). An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. American Zoologist. 39:374-388.

See Also

[asr_squared_change_parsimony](), [asr_max_parsimony](), [asr_mk_model]()

Examples

```
# generate random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a continuous trait
tip_states = simulate_ou_model(tree, stationary_mean=0, spread=1, decay_rate=0.001)$tip_states

# reconstruct node states via weighted PIC
asr = asr_independent_contrasts(tree, tip_states, weighted=TRUE, include_CI=TRUE)
node_states = asr$ancestral_states

# get lower bounds of 95% CIs
lower_bounds = node_states - asr$CI95
```

---

asr_max_parsimony　　　　*Maximum-parsimony ancestral state reconstruction.*

Description

Reconstruct ancestral states for a discrete trait using maximum parsimony. Transition costs can vary between transitions, and can optionally be weighted by edge length.

Usage

```
asr_max_parsimony(tree, tip_states, Nstates=NULL,
                  transition_costs="all_equal",
                  edge_exponent=0, weight_by_scenarios=TRUE,
                  check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	An integer vector of size Ntips, specifying the state of each tip in the tree as an integer from 1 to Nstates, where Nstates is the possible number of states (see below).
Nstates	Either NULL, or an integer specifying the number of possible states of the trait. If NULL, then Nstates will be computed based on the maximum value encountered in tip_states

transition_costs

> Either "all_equal","sequential", "proportional", "exponential", or a quadratic non-negatively valued matrix of size Nstates x Nstates, specifying the transition costs between all possible states (which can include 0 as well as Inf). The [r,c]-th entry of the matrix is the cost of transitioning from state r to state c. The option "all_equal" specifies that all transitions are permitted and are equally costly. "sequential" means that only transitions between adjacent states are permitted and are all equally costly. "proportional" means that all transitions are permitted, but the cost increases proportional to the distance between states. "exponential" means that all transitions are permitted, but the cost increases exponentially with the distance between states. The options "sequential" and "proportional" only make sense if states exhibit an order relation (as reflected in their integer representation).

edge_exponent   Non-negative real-valued number. Optional exponent for weighting transition costs by the inverse length of edge lengths. If 0, edge lengths do not influence the ancestral state reconstruction (this is the conventional max-parsimony). If >0, then at each edge the transition costs are multiplied by $1/L^e$, where L is the edge length and e is the edge exponent. This parameter is mostly experimental; modify at your own discretion.

weight_by_scenarios

> Logical, indicating whether to weight each optimal state of a node by the number of optimal maximum-parsimony scenarios in which the node is in that state. If FALSE, then all optimal states of a node are weighted equally (i.e. are assigned equal probabilities).

check_input   Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

For this function, the trait's states must be represented by integers within 1,..,Nstates, where Nstates is the total number of possible states. If the states are originally in some other format (e.g. characters or factors), you should map them to a set of integers 1,..,Nstates. The order of states (if relevant) should be reflected in their integer representation. For example, if your original states are "small", "medium" and "large" and transition_costs=="sequential", it is advised to represent these states as integers 1,2,3. You can easily map any set of discrete states to integers using the function map_to_state_space.

This function utilizes Sankoff's (1975) dynamic programming algorithm for determining the smallest number (or least costly if transition costs are uneven) of state changes along edges needed to reproduce the observed tip states. The function has asymptotic time complexity O(Ntips+Nnodes x Nstates).

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. If edge_exponent is 0, then edge lengths do not influence the result. If edge_exponent!=0, then all edges must have non-zero length. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child).

Tips must be represented in tip_states in the same order as in tree$tip.label. None of the input vectors or matrixes need include row or column names; if they do, however, they are checked for consistency (if check_input==TRUE).

This function is meant for reconstructing ancestral states in all nodes of a tree, when the state of each tip is known. If some of the tips have unknown state, consider either pruning the tree to keep only tips with known states, or using the function `hsp_max_parsimony`.

Value

A list with the following elements:

success
: Boolean, indicating whether ASR was successful. If `FALSE`, the remaining returned elements may be undefined.

ancestral_likelihoods
: A 2D numeric matrix, listing the probability of each node being in each state. This matrix will have size Nnodes x Nstates, where Nstates was either explicitly provided as an argument or inferred from `tip_states`. The rows in this matrix will be in the order in which nodes are indexed in the tree, i.e. the [n,s]-th entry will be the probability of the s-th state for the n-th node. These probabilities are calculated based on `scenario_counts` (see below), assuming that every maximum parsimony scenario is equally likely. Note that the name was chosen for compatibility with other ASR functions.

scenario_counts
: A 2D numeric matrix of size Nnodes x Nstates, listing for each node and each state the number of maximum parsimony scenarios in which the node was in the specific state. If only a single maximum parsimony scenario exists for the whole tree, then the sum of entries in each row will be one.

total_cost
: Real number, specifying the total transition cost across the tree for the most parsimonious scenario. In the classical case where `transition_costs="all_equal"`, the `total_cost` equals the total number of state changes in the tree under the most parsimonious scenario.

Author(s)

Stilianos Louca

References

D. Sankoff (1975). Minimal mutation trees of sequences. SIAM Journal of Applied Mathematics. 28:35-42.

J. Felsenstein (2004). Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

See Also

`hsp_max_parsimony`, `asr_squared_change_parsimony` `asr_mk_model`, `hsp_mk_model`, `map_to_state_space`

Examples

```
## Not run:
# generate random tree
Ntips = 10
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree
```

```
# simulate a discrete trait
Nstates = 5
Q = get_random_mk_transition_matrix(Nstates, rate_model="ER")
tip_states = simulate_mk_model(tree, Q)$tip_states

# reconstruct node states via MPR
results = asr_max_parsimony(tree, tip_states, Nstates)
node_states = max.col(results$ancestral_likelihoods)

# print reconstructed node states
print(node_states)

## End(Not run)
```

---

asr_mk_model   *Ancestral state reconstruction with Mk models and rerooting*

Description

Ancestral state reconstruction of a discrete trait using a fixed-rates continuous-time Markov model
(a.k.a. "Mk model"). This function can estimate the (instantaneous) transition matrix using maxi-
mum likelihood, or take a specified transition matrix. The function can optionally calculate marginal
ancestral state likelihoods for each node in the tree, using the rerooting method by Yang et al.
(1995).

Usage

```
asr_mk_model( tree,
              tip_states,
              Nstates = NULL,
              tip_priors = NULL,
              rate_model = "ER",
              transition_matrix = NULL,
              include_ancestral_likelihoods = TRUE,
              reroot = TRUE,
              root_prior = "empirical",
              Ntrials = 1,
              optim_algorithm = "nlminb",
              optim_max_iterations = 200,
              optim_rel_tol = 1e-8,
              store_exponentials = TRUE,
              check_input =TRUE,
              Nthreads = 1)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with
                no incoming edge.

tip_states   An integer vector of size Ntips, specifying the state of each tip in the tree in terms of an integer from 1 to Nstates, where Ntips is the number of tips and Nstates is the number of possible states (see below). Can also be NULL. If tip_states==NULL, then tip_priors must not be NULL (see below).

Nstates   Either NULL, or an integer specifying the number of possible states of the trait. If Nstates==NULL, then it will be computed based on the maximum value encountered in tip_states or based on the number of columns in tip_priors (whichever is non-NULL).

tip_priors   A 2D numeric matrix of size Ntips x Nstates, where Nstates is the possible number of states for the character modelled. Hence, tip_priors[i,s] is the likelihood of the observed state of tip i, if the tip's true state was in state s. For example, if you know for certain that a tip is in state k, then set tip_priors[i,s]=1 for s=k and tip_priors[i,s]=0 for all other s.

rate_model   Rate model to be used for fitting the transition rate matrix. Can be "ER" (all rates equal), "SYM" (transition rate i–>j is equal to transition rate j–>i), "ARD" (all rates can be different), "SUEDE" (only stepwise transitions i–>i+1 and i–>i-1 allowed, all 'up' transitions are equal, all 'down' transitions are equal) or "SRD" (only stepwise transitions i–>i+1 and i–>i-1 allowed, and each rate can be different). Can also be an index matrix that maps entries of the transition matrix to the corresponding independent rate parameter to be fitted. Diagonal entries should map to 0, since diagonal entries are not treated as independent rate parameters but are calculated from the remaining entries in the transition matrix. All other entries that map to 0 represent a transition rate of zero. The format of this index matrix is similar to the format used by the ace function in the ape package. rate_model is only relevant if transition_matrix==NULL.

transition_matrix
                  Either a numeric quadratic matrix of size Nstates x Nstates containing fixed transition rates, or NULL. The [r,c]-th entry in this matrix should store the transition rate from state r to state c. Each row in this matrix must have sum zero. If NULL, then the transition rates will be estimated using maximum likelihood, based on the rate_model specified.

root_prior   Prior probability distribution of the root's states, used to calculate the model's overall likelihood from the root's marginal ancestral state likelihoods. Can be "flat" (all states equal), "empirical" (empirical probability distribution of states across the tree's tips), "stationary" (stationary probability distribution of the transition matrix), "likelihoods" (use the root's state likelihoods as prior) or "max_likelihood" (put all weight onto the state with maximum likelihood). If "stationary" and transition_matrix==NULL, then a transition matrix is first fitted using a flat root prior, and then used to calculate the stationary distribution. root_prior can also be a non-negative numeric vector of size Nstates and with total sum equal to 1.

include_ancestral_likelihoods
                  Include the marginal ancestral likelihoods for each node (conditional scaled state likelihoods) in the return values. Note that this may increase the computation time and memory needed, so you may set this to FALSE if you don't need marginal ancestral states.

reroot   Reroot tree at each node when computing marginal ancestral likelihoods, according to Yang et al. (1995). This is the default and recommended behavior, but leads to increased computation time. If FALSE, ancestral likelihoods at each node are computed solely based on the subtree descending from that node, without rerooting.

Ntrials   Number of trials (starting points) for fitting the transition matrix. Only relevant if transition_matrix=NULL. A higher number may reduce the risk of landing in a local non-global optimum of the likelihood function, but will increase computation time during fitting.

optim_algorithm
                  Either "optim" or "nlminb", specifying which optimization algorithm to use for maximum-likelihood estimation of the transition matrix. Only relevant if transition_matrix==NULL.

optim_max_iterations
                  Maximum number of iterations (per fitting trial) allowed for optimizing the likelihood function.

optim_rel_tol   Relative tolerance (stop criterion) for optimizing the likelihood function.

store_exponentials
                  Logical, specifying whether to pre-calculate and store exponentials of the transition matrix during calculation of ancestral likelihoods. This may reduce computation time because each exponential is only calculated once, but requires more memory since all exponentials are stored.

                  Only relevant if include_ancestral_likelihoods==TRUE, otherwise exponentials are never stored.

check_input   Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Nthreads   Number of parallel threads to use for running multiple fitting trials simultaneously. This only makes sense if your computer has multiple cores/CPUs and if Ntrials>1, and is only relevant if transition_matrix==NULL. This option is ignored on Windows, because Windows does not support forking.

Details

For this function, the trait's states must be represented by integers within 1,..,Nstates, where Nstates is the total number of possible states. If the states are originally in some other format (e.g. characters or factors), you should map them to a set of integers 1,..,Nstates. The order of states (if relevant) should be reflected in their integer representation. For example, if your original states are "small", "medium" and "large" and rate_model=="SUEDE", it is advised to represent these states as integers 1,2,3. You can easily map any set of discrete states to integers using the function [map_to_state_space](#).

This function allows the specification of the precise tip states (if these are known) using the vector tip_states. Alternatively, if some tip states are only known in terms of a probability distribution, you can pass these probability distributions using the matrix tip_priors. Note that exactly one of the two arguments, tip_states or tip_priors, must be non-NULL.

Tips must be represented in `tip_states` or `tip_priors` in the same order as in `tree$tip.label`. None of the input vectors or matrixes need include row or column names; if they do, however, they are checked for consistency (if `check_input==TRUE`).

The tree is either assumed to be complete (i.e. include all possible species), or to represent a random subset of species chosen independently of their states. The rerooting method by Yang et al (1995) is used to calculate the marginal ancestral state likelihoods for each node by treating the node as a root and calculating its conditional scaled likelihoods. Note that the re-rooting algorithm is strictly speaking only valid for reversible Mk models, that is, satisfying the criterion

$$\pi_i Q_{ij} = \pi_j Q_{ji}, \quad \forall i, j,$$

where Q is the transition rate matrix and π is the stationary distribution of the model. The rate models "ER", 'SYM', "SUEDE" and "SRD" are reversible. For example, for "SUEDE" or "SRD" choose $\pi_{i+1} = \pi_i Q_{i,i+1}/Q_{i+1,i}$. In contrast, "ARD" models are generally not reversible.

If `tree$edge.length` is missing, each edge in the tree is assumed to have length 1. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). This function is similar to `rerootingMethod` in the phytools package (v0.5-64) and similar to `ape::ace` (v4.1) with options `method="ML",type="discrete"` and `marginal=FALSE`, but tends to be much faster than `rerootingMethod` and ace for large trees.

Value

A list with the following elements:

success	Logical, indicating whether ASR was successful. If `FALSE`, all other return values may be `NULL`.
Nstates	Integer, specifying the number of modeled trait states.

transition_matrix

> A numeric quadratic matrix of size Nstates x Nstates, containing the transition rates of the Markov model. The [r,c]-th entry is the transition rate from state r to state c. Will be the same as the input `transition_matrix`, if the latter was not `NULL`.

loglikelihood	Log-likelihood of the observed tip states under the fitted (or provided) Mk model. If `transition_matrix` was `NULL` in the input, then this will be the log-likelihood maximized during fitting.

ancestral_likelihoods

> Optional, only returned if `include_ancestral_likelihoods` was `TRUE`. A 2D numeric matrix, listing the likelihood of each state at each node (marginal ancestral likelihoods). This matrix will have size Nnodes x Nstates, where Nstates was either explicitly provided as an argument, or inferred from `tip_states` or `tip_priors` (whichever was non-NULL). The rows in this matrix will be in the order in which nodes are indexed in the tree, i.e. the [n,s]-th entry will be the likelihood of the s-th state at the n-th node. For example, `likelihoods[1,3]` will store the likelihood of observing the tree's tip states (if `reroot=TRUE`) or the descending subtree's tip states (if `reroot=FALSE`), if the first node was in state 3. Note that likelihoods are rescaled (normalized) to sum to 1 for convenience and numerical stability. The marginal likelihoods at a node should not, however, be interpreted as a probability distribution among states.

Author(s)

Stilianos Louca

References

Z. Yang, S. Kumar and M. Nei (1995). A new method for inference of ancestral nucleotide and amino acid sequences. Genetics. 141:1641-1650.

See Also

[hsp_mk_model](), [asr_max_parsimony](), [asr_squared_change_parsimony](), [hsp_max_parsimony](), [map_to_state_space]()

Examples

```
## Not run:
# generate random tree
Ntips = 1000
tree  = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# create random transition matrix
Nstates = 5
Q = get_random_mk_transition_matrix(Nstates, rate_model="ER", max_rate=0.01)
cat(sprintf("Simulated ER transition rate=%g\n",Q[1,2]))

# simulate the trait's evolution
simulation = simulate_mk_model(tree, Q)
tip_states = simulation$tip_states
cat(sprintf("Simulated states for last 20 nodes:\n"))
print(tail(simulation$node_states,20))

# reconstruct node states from simulated tip states
# at each node, pick state with highest marginal likelihood
results = asr_mk_model(tree, tip_states, Nstates, rate_model="ER", Ntrials=2)
node_states = max.col(results$ancestral_likelihoods)

# print Mk model fitting summary
cat(sprintf("Mk model: log-likelihood=%g\n",results$loglikelihood))
cat(sprintf("Fitted ER transition rate=%g\n",results$transition_matrix[1,2]))

# print reconstructed node states for last 20 nodes
print(tail(node_states,20))

## End(Not run)
```

---

asr_squared_change_parsimony

Squared-change parsimony ancestral state reconstruction.

Description

Reconstruct ancestral states for a continuous (numeric) trait using squared-change maximum parsimony (Maddison, 1991). Transition costs can optionally be weighted by the inverse edge lengths ("weighted squared-change parsimony" by Maddison).

Usage

```
asr_squared_change_parsimony(tree, tip_states, weighted=TRUE, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	A numeric vector of size Ntips, specifying the known state of each tip in the tree.
weighted	Logical, specifying whether to weight transition costs by the inverted edge lengths. This corresponds to the "weighted squared-change parsimony" reconstruction by Maddison (1991) for a Brownian motion model of trait evolution.
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

The function traverses the tree in postorder (tips–>root) to calculate the quadratic parameters described by Maddison (1991) and obtain the globally parsimonious squared-change parsimony state for the root. The function then reroots at each node, updates all affected quadratic parameters in the tree and calculates the node's globally parsimonious squared-change parsimony state. The function has asymptotic time complexity O(Nedges).

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. This is the same as setting weighted=FALSE. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). Edges with length 0 will be adjusted internally to some tiny length if needed (if weighted==TRUE).

Tips must be represented in tip_states in the same order as in tree$tip.label. The vector tip_states need not include item names; if it does, however, they are checked for consistency (if check_input==TRUE).

If weighted==FALSE, then this function yields the same ancestral state reconstructions as

```
ape::ace(tip_states,tree,type="continuous",method="ML",model="BM",CI=FALSE)
```

in the ape package (v. 0.5-64), assuming the tree as unit edge lengths. If weighted==TRUE, then this function yields the same ancestral state reconstructions as the maximum likelihood estimates under a Brownian motion model, as implemented by the Rphylopars package (v. 0.2.10):

```
Rphylopars::anc.recon(tip_states,tree,vars=FALSE,CI=FALSE).
```

Value

A list with the following elements:

ancestral_states

> A numeric vector of size Nnodes, listing the reconstructed state of each node. The entries in this vector will be in the order in which nodes are indexed in the tree.

total_sum_of_squared_changes

> The total sum of squared changes, minimized by the (optionally weighted) squared-change parsimony algorithm. This is equation 7 in (Maddison, 1991).

Author(s)

Stilianos Louca

References

W. P. Maddison (1991). Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Systematic Zoology. 40:304-314.

See Also

[asr_independent_contrasts](asr_max_parsimony), asr_mk_model

Examples

```
# generate random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a continuous trait
tip_states = simulate_ou_model(tree, stationary_mean=0, spread=1, decay_rate=0.001)$tip_states

# reconstruct node states based on simulated tip states
node_states = asr_squared_change_parsimony(tree, tip_states, weighted=TRUE)$ancestral_states
```

---

asr_subtree_averaging   *Ancestral state reconstruction via subtree averaging.*

Description

Reconstruct ancestral states in a phylogenetic tree for a continuous (numeric) trait by averaging trait values over descending subtrees. That is, for each node the reconstructed state is set to the arithmetic average state of all tips descending from that node.

Usage

```
asr_subtree_averaging(tree, tip_states, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	A numeric vector of size Ntips, specifying the known state of each tip in the tree.
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

The function returns the estimated ancestral states (=averages) as well as the corresponding standard deviations. Note that reconstructed states are local estimates, i.e. they only take into account the tips descending from the reconstructed node.

The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). Edge lengths and distances between tips and nodes are not taken into account. All tip states are assumed to be known, and NA or NaN are not allowed in tip_states.

Tips must be represented in tip_states in the same order as in tree$tip.label. The vector tip_states need not include item names; if it does, however, they are checked for consistency (if check_input==TRUE).

Value

A list with the following elements:

success	Logical, indicating whether ASR was sucessful. If all input data are valid then this will always be TRUE, but it is provided for consistency with other ASR functions.
ancestral_states	A numeric vector of size Nnodes, listing the reconstructed state (=average over descending tips) for each node. The entries in this vector will be in the order in which nodes are indexed in the tree.
ancestral_stds	A numeric vector of size Nnodes, listing the standard deviations corresponding to ancestral_stds.
ancestral_counts	A numeric vector of size Nnodes, listing the number of (descending) tips used to reconstruct the state of each node.

Author(s)

Stilianos Louca

See Also

[asr_independent_contrasts](), [asr_squared_change_parsimony]()

Examples

```
# generate random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a continuous trait
tip_states = simulate_ou_model(tree, stationary_mean=0, spread=1, decay_rate=0.001)$tip_states

# reconstruct node states by averaging simulated tip states
node_states = asr_subtree_averaging(tree, tip_states)$ancestral_states
```

---

collapse_monofurcations

Remove monofurcations from a tree.

Description

Eliminate monofurcations (nodes with only a single child) from a phylogenetic tree, by connecting their incoming and outgoing edge.

Usage

```
collapse_monofurcations(tree, force_keep_root=TRUE, as_edge_counts=FALSE)
```

Arguments

tree   A rooted tree of class "phylo".

force_keep_root

                    Logical, indicating whether the root node should always be kept (i.e., even if it only has a single child).

as_edge_counts   Logical, indicating whether all edges should be assumed to have length 1. If TRUE, the outcome is the same as if the tree had no edges.

Details

All tips in the input tree retain their original indices, however the returned tree may include fewer nodes and edges. Edge and node indices may change.

If tree$edge.length is missing, then all edges in the input tree are assumed to have length 1.

Value

A list with the following elements:

tree   A new tree of class "phylo", containing only bifurcations (and multifurcations, if these existed in the input tree). The number of nodes in this tree, Nnodes_new, may be lower than of the input tree.

new2old_node   Integer vector of length Nnodes_new, mapping node indices in the new tree to node indices in the old tree.

Nnodes_removed   Integer. Number of nodes (monofurcations) removed from the tree.

Author(s)

Stilianos Louca

See Also

multifurcations_to_bifurcations

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1), max_tips=Ntips)$tree

# prune the tree to generate random monofurcations
random_tips = sample.int(n=Ntips, size=0.5 * Ntips, replace=FALSE)
tree = get_subtree_with_tips(tree, only_tips=random_tips, collapse_monofurcations=FALSE)$subtree

# collapse monofurcations
new_tree = collapse_monofurcations(tree)$tree

# print summary of old and new tree
cat(sprintf("Old tree has %d nodes\n",tree$Nnode))
cat(sprintf("New tree has %d nodes\n",new_tree$Nnode))
```

---

collapse_tree_at_resolution

Collapse nodes of a tree at a phylogenetic resolution.

Description

Given a rooted tree and a phylogenetic resolution threshold, collapse all nodes whose distance to all descending tips does not exceed the threshold (or whose sum of descending edge lengths does not exceed the threshold), into new tips. This function can be used to obtain a "coarser" version of the tree, or to cluster closely related tips into a single tip.

Usage

```
collapse_tree_at_resolution(tree,
                            resolution            = 0,
                            by_edge_count         = FALSE,
                            shorten               = TRUE,
                            rename_collapsed_nodes = FALSE,
                            criterion             = 'max_tip_depth')
```

Arguments

`tree`	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
`resolution`	Numeric, specifying the phylogenetic resolution at which to collapse the tree. This is the maximum distance a descending tip can have from a node, such that the node is collapsed into a new tip. If set to 0 (default), then only nodes whose descending tips are identical to the node will be collapsed.
`by_edge_count`	Logical. Instead of considering edge lengths, consider edge counts as phylogenetic distance between nodes and tips. This is the same as if all edges had length equal to 1.
`shorten`	Logical, indicating whether collapsed nodes should be turned into tips at the same location (thus potentially shortening the tree). If FALSE, then the incoming edge of each collapsed node is extended by some length L, where L is the distance of the node to its farthest descending tip (thus maintaining the height of the tree).
`rename_collapsed_nodes`	
	Logical, indicating whether collapsed nodes should be renamed using a representative tip name (the farthest descending tip). See details below.
`criterion`	Character, specifying the criterion to use for collapsing (i.e. how to interpret `resolution`). 'max_tip_depth': Collapse nodes based on their maximum distance to any descending tip. 'sum_tip_paths': Collapse nodes based on the sum of descending edges (each edge counted once). 'max_tip_pair_dist': Collapse nodes based on the maximum distance between any pair of descending tips.

Details

The tree is traversed from root to tips and nodes are collapsed into tips as soon as the criterion equals or falls below the resolution threshold.

The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). Tip labels and uncollapsed node labels of the collapsed tree are inheritted from the original tree. If rename_collapsed_nodes==FALSE, then labels of collapsed nodes will be the node labels from the original tree (in this case the original tree should include node labels). If rename_collapsed_nodes==TRUE, each collapsed node is given the label of its farthest descending tip. If shorten==TRUE, then edge lengths are the same as in the original tree. If shorten==FALSE, then edges leading into collapsed nodes may be longer than before.

Value

A list with the following elements:

`tree`	A new rooted tree of class "phylo", containing the collapsed tree.
`root_shift`	Numeric, indicating the phylogenetic distance between the old and the new root. Will always be non-negative.
`collapsed_nodes`	
	Integer vector, listing indices of collapsed nodes in the original tree (subset of 1,..,Nnodes).

farthest_tips	Integer vector of the same length as `collapsed_nodes`, listing indices of the farthest tips for each collapsed node. Hence, `farthest_tips[n]` will be the index of a tip in the original tree that descended from node `collapsed_nodes[n]` and had the greatest distance from that node among all descending tips.
new2old_clade	Integer vector of length equal to the number of tips+nodes in the collapsed tree, with values in 1,..,Ntips+Nnodes, mapping tip/node indices of the collapsed tree to tip/node indices in the original tree.
new2old_edge	Integer vector of length equal to the number of edges in the collapsed tree, with values in 1,..,Nedges, mapping edge indices of the collapsed tree to edge indices in the original tree.

Author(s)

Stilianos Louca

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=1000)$tree

# print number of nodes
cat(sprintf("Simulated tree has %d nodes\n",tree$Nnode))

# collapse any nodes with tip-distances < 20
collapsed = collapse_tree_at_resolution(tree, resolution=20)$tree

# print number of nodes
cat(sprintf("Collapsed tree has %d nodes\n",collapsed$Nnode))
```

---

congruent_divergence_times
                    Extract dating anchors for a target tree, using a dated reference tree

---

Description

Given a reference tree and a target tree, this function maps target nodes to concordant reference nodes when possible, and extracts divergence times of the mapped reference nodes from the reference tree. This function can be used to define secondary dating constraints for a larger target tree, based on a time-calibrated smaller reference tree (Eastman et al. 2013). This only makes sense if the reference tree is time-calibrated. A provided mapping specifies which and how tips in the target tree correspond to tips in the reference tree.

Usage

```
congruent_divergence_times(reference_tree, target_tree, mapping)
```

Arguments

reference_tree   A rooted tree object of class "phylo". Usually this tree will be time-calibrated
                  (i.e. edge lengths represent time intervals).

target_tree   A rooted tree object of class "phylo".

mapping   A table mapping a subset of target tips to a subset of reference tips, as described
                  by Eastman et al (2013). Multiple target tips may map to the same reference
                  tip, but not vice versa (i.e. every target tip can appear at most once in the map-
                  ping). In general, a tip mapped to in the reference tree is assumed to represent a
                  monophyletic group of tips in the target tree, although this assumption may be
                  violated in practice (Eastman et al. 2013).

                  The mapping must be in one of the following formats:

                  Option 1: A 2D integer array of size NM x 2 (with NM being the number of
                  mapped target tips), listing target tip indices mapped to reference tip indices
                  (mapping[m,1] (target tip) –> mapping[m,2] (reference tip)).

                  Option 2: A 2D character array of size NM x 2, listing target tip labels mapped
                  to reference tip labels.

                  Option 3: A data frame of size NM x 1, whose row names are target tip la-
                  bels and whose entries are either integers (reference tip indices) or characters
                  (reference tip labels). This is the format used by geiger::congruify.phylo
                  (v.206).

                  Option 4: A vector of size NM, whose names are target tip labels and whose
                  entries are either integers (reference tip indices) or characters (reference tip la-
                  bels).

Details

Both the reference and target tree may include monofurcations and/or multifurcations. In principle,
neither of the two trees needs to be ultrametric, although in most applications reference_tree will
be ultrametric.

In special cases each reference tip may be found in the target tree, i.e. the reference tree is a subtree
of the target tree. This may occur e.g. if a smaller subtree of the target tree has been extracted and
dated, and subsequently the larger target tree is to be dated using secondary constraints inferred
from the dated subtree.

The function returns a table that maps a subset of target nodes to an equally sized subset of con-
cordant reference nodes. Ages (divergence times) of the mapped reference nodes are extracted and
associated with the concordant target nodes.

For bifurcating trees the average time complexity of this function is O(TNtips x log(RNtips) x NM),
where TNtips and RNtips are the number of tips in the target and reference tree, respectively. This
function is similar to geiger::congruify.phylo (v.206). For large trees, this function tends to be
much faster than geiger::congruify.phylo.

Value

A named list with the following elements:

Rnodes   Integer vector of length NC (where NC is the number of concordant node pairs
                  found) and with values in 1,..,RNnodes, listing indices of reference nodes that

could be matched with (i.e. were concordant to) a target node. Entries in Rnodes will correspond to entries in Tnodes and ages.

Tnodes   Integer vector of length NC and with values in 1,..,TNnodes, listing indices of target nodes that could be matched with (i.e. were concordant to) a reference node. Entries in Tnodes will correspond to entries in Rnodes and ages.

ages   Numeric vector of length NC, listing divergence times (ages) of the reference nodes listed in Rnodes. These ages can be used as fixed anchors for time-calibrating the target tree using a separate program (such as PATHd8).

Author(s)

Stilianos Louca

References

J. M. Eastman, L. J. Harmon, D. C. Tank (2013). Congruification: support for time scaling large phylogenetic trees. Methods in Ecology and Evolution. 4:688-691.

See Also

[extend_tree_to_height](), [date_tree_red](), [get_tips_for_mrcas](), [tree_distance]()

Examples

```
# generate random tree (target tree)
Ntips = 10000
tree = castor::generate_random_tree(parameters=list(birth_rate_intercept=1), max_tips=Ntips)$tree

# extract random subtree (reference tree)
Nsubtips    = 10
subtips     = sample.int(n=Ntips,size=Nsubtips,replace=FALSE)
subtreeing  = castor::get_subtree_with_tips(tree, only_tips=subtips)
subtree     = subtreeing$subtree

# map subset of target tips to reference tips
mapping = matrix(c(subtreeing$new2old_tip,(1:Nsubtips)),ncol=2,byrow=FALSE)

# extract divergence times by congruification
congruification = congruent_divergence_times(subtree, tree, mapping)

cat("Concordant target nodes:\n")
print(congruification$target_nodes)

cat("Ages of concordant nodes:\n")
print(congruification$ages)
```

---

congruent_hbds_model   *Generate a congruent homogenous-birth-death-sampling model.*

Description

Given a homogenous birth-death-sampling (HBDS) model (or abstract congruence class), obtain the congruent model (or member of the congruence class) with a specific speciation rate λ, or extinction rate μ, or sampling rate ψ, or effective reproduction ratio R_e or removal rate $\mu + \psi$ (aka. "become uninfectious"" rate). All input and output time-profiles are specified as piecewise polynomial curves (splines), defined on a discrete grid of ages. This function allows exploration of a model's congruence class, by obtaining various members of the congruence class depending on the specified λ, μ, ψ, R_e or removal rate. For more details on HBDS models and congruence classes see the documentation of `simulate_deterministic_hbds`.

Usage

```
congruent_hbds_model(age_grid,
                     PSR,
                     PDR,
                     lambda_psi,
                     lambda           = NULL,
                     mu               = NULL,
                     psi              = NULL,
                     Reff             = NULL,
                     removal_rate     = NULL,
                     lambda0          = NULL,
                     CSA_ages         = NULL,
                     CSA_pulled_probs = NULL,
                     CSA_PSRs         = NULL,
                     splines_degree   = 1,
                     ODE_relative_dt  = 0.001,
                     ODE_relative_dy  = 1e-4)
```

Arguments

age_grid   Numeric vector, listing discrete ages (time before present) on which the various model variables (e.g., λ, μ etc) are specified. Listed ages must be strictly increasing, and must include the present-day (i.e. age 0).

PSR   Numeric vector, of the same size as age_grid, specifying the pulled speciation rate (PSR) (in units 1/time) at the ages listed in age_grid. The PSR is assumed to vary polynomially between grid points (see argument splines_degree). Can also be a single number, in which case PSR is assumed to be time-independent.

PDR   Numeric vector, of the same size as age_grid, specifying the pulled diversification rate (PDR) (in units 1/time) at the ages listed in age_grid. The PDR is assumed to vary polynomially between grid points (see argument splines_degree). The PDR of a HBDS model is defined as $PDR = \lambda - \mu - \psi + (1/\lambda)d\lambda/dt$

	(where t denotes age). Can also be a single number, in which case PDR is assumed to be time-independent.
lambda_psi	Numeric vector, of the same size as age_grid, specifying the product of speciation rate and sampling rate ($\lambda\psi$, in units 1/time^2) at the ages listed in age_grid. $\lambda\psi$ is assumed to vary polynomially between grid points (see argument splines_degree). Can also be a single number, in which case $\lambda\psi$ is assumed to be time-independent.
lambda	Numeric vector, of the same size as age_grid, specifying the speciation rate (λ, in units 1/time) at the ages listed in age_grid. The speciation rate is assumed to vary polynomially between grid points (see argument splines_degree). Can also be a single number, in which case λ is assumed to be time-independent. By providing λ, one can select a specific model from the congruence class. Note that exactly one of lambda, mu, psi, Reff or removal_rate must be provided.
mu	Numeric vector, of the same size as age_grid, specifying the extinction rate (μ, in units 1/time) at the ages listed in age_grid. The extinction rate is assumed to vary polynomially between grid points (see argument splines_degree). Can also be a single number, in which case μ is assumed to be time-independent. In an epidemiological context, μ typically corresponds to the recovery rate plus the death rate of infected individuals. By providing μ (together with lambda0, see below), one can select a specific model from the congruence class. Note that exactly one of lambda, mu, psi, Reff or removal_rate must be provided.
psi	Numeric vector, of the same size as age_grid, specifying the (Poissonian) sampling rate (ψ, in units 1/time) at the ages listed in age_grid. The sampling rate is assumed to vary polynomially between grid points (see argument splines_degree). Can also be a single number, in which case ψ is assumed to be time-independent. By providing ψ, one can select a specific model from the congruence class. Note that exactly one of lambda, mu, psi, Reff or removal_rate must be provided.
Reff	Numeric vector, of the same size as age_grid, specifying the effective reproduction ratio (R_e, unitless) at the ages listed in age_grid. The R_e is assumed to vary polynomially between grid points (see argument splines_degree). Can also be a single number, in which case R_e is assumed to be time-independent. By providing R_e (together with lambda0, see below), one can select a specific model from the congruence class. Note that exactly one of lambda, mu, psi, Reff or removal_rate must be provided.
removal_rate	Numeric vector, of the same size as age_grid, specifying the removal rate ($\mu + \psi$, in units 1/time) at the ages listed in age_grid. IN an epidemiological context this is also known as "become uninfectious" rate. The removal rate is assumed to vary polynomially between grid points (see argument splines_degree). Can also be a single number, in which case the removal rate is assumed to be time-independent. By providing $\mu + \psi$ (together with lambda0, see below), one can select a specific model from the congruence class. Note that exactly one of lambda, mu, psi, Reff or removal_rate must be provided.
lambda0	Numeric, specifying the speciation rate at the present-day (i.e., at age 0). Must be provided if and only if one of mu, Reff or removal_rate is provided.
CSA_ages	Optional numeric vector, listing the ages of concentrated sampling attempts, in ascending order. Concentrated sampling is assumed to occur in addition to any continuous (Poissonian) sampling specified by psi.

CSA_pulled_probs

> Optional numeric vector of the same size as CSA_ages, listing pulled sampling probabilities at each concentrated sampling attempt (CSA). Note that in contrast to the sampling rates psi, the CSA_pulled_probs are interpreted as probabilities and must thus be between 0 and 1. CSA_pulled_probs must be provided if and only if CSA_ages is provided.

CSA_PSRs   Optional numeric vector of the same size as CSA_ages, specifying the pulled sampling rate (PSR) during each concentrated sampling attempt. While in principle the PSR is already provided by the argument PSR, the PSR may be non-continuous at CSAs, which makes a representation as piecewise polynomial function difficult; hence, you must explicitly provide the correct PSR at each CSA. CSA_PSRs must be provided if and only if CSA_ages is provided.

splines_degree   Integer, either 0,1,2 or 3, specifying the polynomial degree of the provided time-dependent variables between grid points in age_grid. For example, if splines_degree==1, then the provided PDR, PSR and so on are interpreted as piecewise-linear curves; if splines_degree==2 they are interpreted as quadratic splines; if splines_degree==3 they are interpreted as cubic splines. The splines_degree influences the analytical properties of the curve, e.g. splines_degree==1 guarantees a continuous curve, splines_degree==2 guarantees a continuous curve and continuous derivative, and so on.

ODE_relative_dt

> Positive unitless number, specifying the default relative time step for internally used ordinary differential equation solvers. Typical values are 0.01-0.001.

ODE_relative_dy

> Positive unitless number, specifying the relative difference between subsequent simulated and interpolated values, in internally used ODE solvers. Typical values are 1e-2 to 1e-5. A smaller ODE_relative_dy increases interpolation accuracy, but also increases memory requirements and adds runtime.

Details

The PDR, PSR and the product $\lambda\psi$ are variables that are invariant across the entire congruence class of an HBDS model, i.e. any two congruent models have the same PSR, PDR and product $\lambda\psi$. Reciprocally, any HBDS congruence class is fully determined by its PDR, PSR and $\lambda\psi$. This function thus allows "collapsing" a congruence class down to a single member (a specific HBDS model) by specifying one or more additional variables over time (such as λ, or ψ, or μ and λ_0). Alternatively, this function may be used to obtain alternative models that are congruent to some reference model, for example to explore the robustness of specific epidemiological quantities of interest. The function returns a specific HBDS model in terms of the time profiles of various variables (such as λ, μ and ψ).

In the current implementation it is assumed that any sampled lineages are immediately removed from the pool, that is, this function cannot accommodate models with a non-zero retention probability upon sampling. This is a common assumption in molecular epidemiology. Note that in this function age always refers to time before present, i.e., present day age is 0, and age increases towards the root.

Value

A named list with the following elements:

success	Logical, indicating whether the calculation was successful. If FALSE, then the returned list includes an additional 'error' element (character) providing a description of the error; all other return variables may be undefined.
valid	Logical, indicating whether the returned model appears to be biologically valid (for example, does not have negative λ, μ or ψ). In principle, a congruence class may include biologically invalid models, which might be returned depending on the input to congruent_hbds_model. Note that only biologically valid models can be subsequently simulated using [simulate_deterministic_hbds](#).
ages	Numeric vector of size NG, specifying the discrete ages (time before present) on which all returned time-curves are specified. Will always be equal to age_grid.
lambda	Numeric vector of size NG, listing the speciation rates λ of the returned model at the ages given in ages[].
mu	Numeric vector of size NG, listing the extinction rates μ of the returned model at the ages given in ages[].
psi	Numeric vector of size NG, listing the (Poissonian) sampling rates ψ of the returned model at the ages given in ages[].
lambda_psi	Numeric vector of size NG, listing the product $\lambda\psi$ at the ages given in ages[].
Reff	Numeric vector of size NG, listing the effective reproduction ratio R_e of the returned model at the ages given in ages[].
removal_rate	Numeric vector of size NG, listing the removal rate ($\mu + \psi$, aka. "become uninfectious" rate) of the returned model at the ages given in ages[].
Pmissing	Numeric vector of size NG, listing the probability that a past lineage extant during ages[] will be missing from a tree generated by the model.
CSA_probs	Numeric vector of the same size as CSA_ages, listing the sampling probabilities at each of the CSAs.
CSA_Pmissings	Numeric vector of the same size as CSA_ages, listing the probability that a past lineage extant during each of CSA_ages[] will be missing from a tree generated by the model.

Author(s)

Stilianos Louca

References

T. Stadler, D. Kuehnert, S. Bonhoeffer, A. J. Drummond (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). PNAS. 110:228-233.

A. MacPherson, S. Louca, A. McLaughlin, J. B. Joy, M. W. Pennell (in review as of 2020). A general birth-death-sampling model for epidemiology and macroevolution. DOI:10.1101/2020.10.10.334383

See Also

[generate_tree_hbds](#), [fit_hbds_model_parametric](#), [simulate_deterministic_hbds](#)

Examples

```
# define an HBDS model with exponentially decreasing speciation/extinction rates
# and constant Poissonian sampling rate psi
oldest_age= 10
age_grid  = seq(from=0,to=oldest_age,by=0.1) # choose a sufficiently fine age grid
lambda    = 1*exp(0.01*age_grid) # define lambda on the age grid
mu        = 0.2*lambda # assume similarly shaped but smaller mu

# simulate deterministic HBD model
# scale LTT such that it is 100 at age 1
sim = simulate_deterministic_hbds(age_grid  = age_grid,
                                  lambda    = lambda,
                                  mu        = mu,
                                  psi       = 0.1,
                                  age0      = 1,
                                  LTT0      = 100)

# calculate a congruent HBDS model with an alternative sampling rate
# use the previously simulated variables to define the congruence class
new_psi = 0.1*exp(-0.01*sim$ages) # consider a psi decreasing with age
congruent = congruent_hbds_model(age_grid   = sim$ages,
                                 PSR        = sim$PSR,
                                 PDR        = sim$PDR,
                                 lambda_psi = sim$lambda_psi,
                                 psi        = new_psi)

# compare the deterministic LTT of the two models
# to confirm that the models are indeed congruent
if(!congruent$valid){
    cat("WARNING: Congruent model is not biologically valid\n")
}else{
    # simulate the congruent model to get the LTT
    csim = simulate_deterministic_hbds(age_grid = congruent$ages,
                                      lambda   = congruent$lambda,
                                      mu       = congruent$mu,
                                      psi      = congruent$psi,
                                      age0     = 1,
                                      LTT0     = 100)

    # plot deterministic LTT of the original and congruent model
    plot( x = sim$ages, y = sim$LTT, type='l',
          main='dLTT', xlab='age', ylab='lineages',
          xlim=c(oldest_age,0), col='red')
    lines(x= csim$ages, y=csim$LTT,
          type='p', pch=21, col='blue')
}
```

---

count_lineages_through_time

Count number of lineages through time (LTT).

Description

Given a rooted timetree (i.e., a tree whose edge lengths represent time intervals), calculate the number of lineages represented in the tree at various time points, otherwise known as "lineages through time"" (LTT) curve. The root is interpreted as time 0, and the distance of any node or tip from the root is interpreted as time elapsed since the root. Optionally, the slopes and relative slopes of the LTT curve are also returned.

Usage

```
count_lineages_through_time(  tree,
                              Ntimes        = NULL,
                              min_time      = NULL,
                              max_time      = NULL,
                              times         = NULL,
                              include_slopes= FALSE,
                              degree        = 1,
                              regular_grid  = TRUE)
```

Arguments

tree	A rooted tree of class "phylo", where edge lengths represent time intervals (or similar).
Ntimes	Integer, number of equidistant time points at which to count lineages. Can also be NULL, in which case times must be provided.
min_time	Minimum time (distance from root) to consider. If NULL, this will be set to the minimum possible (i.e. 0). Only relevant if times==NULL.
max_time	Maximum time (distance from root) to consider. If NULL, this will be set to the maximum possible. Only relevant if times==NULL.
times	Integer vector, listing time points (in ascending order) at which to count lineages. Can also be NULL, in which case Ntimes must be provided.
include_slopes	Logical, specifying whether the slope and the relative slope of the returned clades-per-time-point curve should also be returned.
degree	Integer, specifying the "degree" of the LTT curve: LTT(t) will be the number of lineages in the tree at time t that have at least n descending tips in the tree. Typically order=1, which corresponds to the classical LTT curve.
regular_grid	Logical, specifying whether the automatically generated time grid should be regular (equal distances between grid points). This option only matters if times==NULL. If regular_grid==FALSE and times==NULL, the time grid will be irregular, with grid point density being roughly proportional to the square root of the number of lineages at any particular time (i.e., the grid becomes finer towards the tips).

Details

Given a sequence of time points between a tree's root and tips, this function essentially counts how many edges "cross" each time point (if degree==1). The slopes and relative slopes are calculated from this curve using finite differences.

Note that the classical LTT curve (degree=1) is non-decreasing over time, whereas higher-degree LTT's may be decreasing as well as increasing over time.

If `tree$edge.length` is missing, then every edge in the tree is assumed to be of length 1. The tree may include multifurcations as well as monofurcations (i.e. nodes with only one child). The tree need not be ultrametric, although in general this function only makes sense for dated trees (e.g., where edge lengths are time intervals or similar).

Either `Ntimes` or `times` must be non-NULL, but not both. If `times!=NULL`, then `min_time` and `max_time` must be NULL.

Value

A list with the following elements:

`Ntimes`	Integer, indicating the number of returned time points. Equal to the provided `Ntimes` if applicable.
`times`	Numeric vector of size Ntimes, listing the time points at which the LTT was calculated. If `times` was provided as an argument to the function, then this will be the same as provided.
`lineages`	Integer vector of size Ntimes, listing the number of lineages represented in the tree at each time point that have at least degree descending tips, i.e. the LTT curve.
`slopes`	Numeric vector of size Ntimes, listing the slopes (finite-difference approximation of 1st derivative) of the LTT curve.
`relative_slopes`	
	Numeric vector of size Ntimes, listing the relative slopes of the LTT curve, i.e. `slopes` divided by a sliding-window average of `lineages`.

Author(s)

Stilianos Louca

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1), max_tips=1000)$tree

# calculate classical LTT curve
results = count_lineages_through_time(tree, Ntimes=100)

# plot classical LTT curve
plot(results$times, results$lineages, type="l", xlab="time", ylab="# clades")
```

---

count_tips_per_node *Count descending tips.*

Description

Given a rooted phylogenetic tree, count the number of tips descending (directy or indirectly) from each node.

Usage

```
count_tips_per_node(tree)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

Details

The asymptotic time complexity of this function is O(Nedges), where Nedges is the number of edges.

Value

An integer vector of size Nnodes, with the i-th entry being the number of tips descending (directly or indirectly) from the i-th node.

Author(s)

Stilianos Louca

See Also

get_subtree_at_node

Examples

```
# generate a tree using a simple speciation model
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=1000)$tree

# count number of tips descending from each node
tips_per_node = count_tips_per_node(tree);

# plot histogram of tips-per-node
barplot(table(tips_per_node[tips_per_node<10]), xlab="# tips", ylab="# nodes")
```

---

count_transitions_between_clades

Count the number of state transitions between tips or nodes.

Description

Given a rooted phylogenetic tree, one or more pairs of tips and/or nodes, and the state of some discrete trait at each tip and node, calculate the number of state transitions along the shortest path between each pair of tips/nodes.

Usage

```
count_transitions_between_clades(tree, A, B, states, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
A	An integer vector or character vector of size Npairs, specifying the first of the two members of each pair of tips/nodes. If an integer vector, it must list indices of tips (from 1 to Ntips) and/or nodes (from Ntips+1 to Ntips+Nnodes). If a character vector, it must list tip and/or node names.
B	An integer vector or character vector of size Npairs, specifying the second of the two members of each pair of tips/nodes. If an integer vector, it must list indices of tips (from 1 to Ntips) and/or nodes (from Ntips+1 to Ntips+Nnodes). If a character vector, it must list tip and/or node names.
states	Integer vector of length Ntips+Nnodes, listing the discrete state of each tip and node in the tree. The order of entries must match the order of tips and nodes in the tree; this requirement is only verified if states has names and check_input==TRUE.
check_input	Logical, whether to perform basic validations of the input data. If you know for certain that your input is valid, you can set this to FALSE to reduce computation time.

Details

The discrete state must be represented by integers (both negatives and positives are allowed); characters and other data types are not allowed. If tip/node states are originally encoded as characters rather than integers, you can use map_to_state_space to convert these to integers (for example "male" & "female" may be represented as 1 & 2). Also note that a state must be provided for each tip and ancestral node, not just for the tips. If you only know the states of tips, you can use an ancestral state reconstruction tool to estimate ancestral states first.

The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child). If A and/or B is a character vector, then tree$tip.label must exist. If node names are included in A and/or B, then tree$node.label must also exist.

Value

An integer vector of size Npairs, with the i-th element being the number of state transitions between tips/nodes A[i] and B[i] (along their shortest connecting path).

Author(s)

Stilianos Louca

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# pick 3 random pairs of tips or nodes
Npairs = 3
A = sample.int(n=(Ntips+tree$Nnode), size=Npairs, replace=FALSE)
B = sample.int(n=(Ntips+tree$Nnode), size=Npairs, replace=FALSE)

# assign a random state to each tip & node in the tree
# consider a binary trait
states = sample.int(n=2, size=Ntips+tree$Nnode, replace=TRUE)

# calculate number of transitions for each tip pair
Ntransitions = count_transitions_between_clades(tree, A, B, states=states)
```

---

date_tree_red   *Date a tree based on relative evolutionary divergences.*

Description

Given a rooted phylogenetic tree and a single node ('anchor') of known age (distance from the present), rescale all edge lengths so that the tree becomes ultrametric and edge lengths correspond to time intervals. The function is based on relative evolutionary divergences (RED), which measure the relative position of each node between the root and its descending tips (Parks et al. 2018). If no anchor node is provided, the root is simply assumed to have age 1. This function provides a heuristic quick-and-dirty way to date a phylogenetic tree.

Usage

```
date_tree_red(tree, anchor_node = NULL, anchor_age = 1)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
anchor_node	Integer, ranging between 1 and Nnodes. Index of the node to be used as dating anchor. If NULL, the tree's root is used as anchor.
anchor_age	Positive numeric. Age of the anchor node.

Details

The RED of a node measures its relative placement between the root and the node's descending tips (Parks et al. 2018). The root's RED is set to 0. Traversing from root to tips (preorder traversal), for each node the RED is set to $P + (a/(a+b)) \cdot (1-P)$, where P is the RED of the node's parent, a is the edge length connecting the node to its parent, and b is the average distance from the node to its descending tips. The RED of all tips is set to 1.

For each edge, the RED difference between child & parent is used to set the new length of that edge, multiplied by some common scaling factor to translate RED units into time units. The scaling factor is chosen such that the new distance of the anchor node from its descending tips equals `anchor_age`. All tips will have age 0. The topology of the dated tree, as well as tip/node/edge indices, remain unchanged.

This function provides a heuristic approach to making a tree ultrametric, and has not been derived from a specific evolutionary model. In particular, its statistical properties are unknown to the author.

The time complexity of this function is O(Nedges). The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). If `tree$edge.length` is `NULL`, then all edges in the input tree are assumed to have length 1.

Value

A list with the following elements:

success	Logical, indicating whether the dating was successful. If `FALSE`, all other return values (except for `error`) may be undefined.
tree	A new rooted tree of class "phylo", representing the dated tree.
REDs	Numeric vector of size Nnodes, listing the RED of each node in the input tree.
error	Character, listing any error message if `success==FALSE`.

Author(s)

Stilianos Louca

References

D. H. Parks, M. Chuvochina et al. (2018). A proposal for a standardized bacterial taxonomy based on genome phylogeny. bioRxiv 256800. DOI:10.1101/256800

See Also

[congruent_divergence_times](#)

Examples

```
# generate a random non-ultrametric tree
params = list(birth_rate_intercept=1, death_rate_intercept=0.8)
tree = generate_random_tree(params, max_time=1000, coalescent=FALSE)$tree

# make ultrametric, by setting the root to 2 million years
dated_tree = date_tree_red(tree, anchor_age=2e6)
```

---

evaluate_spline *Evaluate a scalar spline at arbitrary locations.*

Description

Given a natural spline function $Y : \mathbb{R} \to \mathbb{R}$, defined as a series of Y values on a discrete X grid, evaluate its values (or derivative) at arbitrary X points. Supported splines degrees are 0 (Y is piecewise constant), 1 (piecewise linear), 2 (piecewise quadratic) and 3 (piecewise cubic).

Usage

```
evaluate_spline(Xgrid,
                Ygrid,
                splines_degree,
                Xtarget,
                extrapolate = "const",
                derivative  = 0)
```

Arguments

Xgrid	Numeric vector, listing x-values in ascending order.
Ygrid	Numeric vector of the same length as Xgrid, listing the values of Y on Xgrid.
splines_degree	Integer, either 0, 1, 2 or 3, specifying the polynomial degree of the spline curve Y between grid points. For example, 0 means Y is piecewise constant, 1 means Y is piecewise linear and so on.
Xtarget	Numeric vector, listing arbitrary X values on which to evaluate Y.
extrapolate	Character, specifying how to extrapolate Y beyond Xgrid if needed. Available options are "const" (i.e. use the value of Y on the nearest Xgrid point) or "splines" (i.e. use the polynomial coefficients from the nearest grid point).
derivative	Integer, specifying which derivative to return. To return the spline's value, set derivative=0. Currently only the options 0,1,2 are supported.

Details

Spline functions are returned by some of castor's fitting routines, so evaluate_spline is meant to aid with the evaluation and plotting of such functions. A spline function of degree $D \geq 1$ has continuous derivatives up to degree $D - 1$. The function evaluate_spline is much more efficient if Xtarget is monotonically increasing or decreasing.

Value

A numeric vector of the same length as Xtarget, listing the values (or derivatives, if derivative>0) of Y on Xtarget.

Author(s)

Stilianos Louca

Examples

```
# specify Y on a coarse X grid
Xgrid = seq(from=0,to=10,length.out=10)
Ygrid = sin(Xgrid)

# define a fine grid of target X values
Xtarget = seq(from=0,to=10,length.out=1000)

# evaluate Y on Xtarget, either as piecewise linear or piecewise cubic function
Ytarget_lin = evaluate_spline(Xgrid,Ygrid,splines_degree=1,Xtarget=Xtarget)
Ytarget_cub = evaluate_spline(Xgrid,Ygrid,splines_degree=3,Xtarget=Xtarget)

# plot both the piecewise linear and piecewise cubic curves
plot(x=Xtarget, y=Ytarget_cub, type='l', col='red', xlab='X', ylab='Y')
lines(x=Xtarget, y=Ytarget_lin, type='l', col='blue', xlab='X', ylab='Y')
```

---

exponentiate_matrix   *Exponentiate a matrix.*

Description

Calculate the exponential $\exp(T \cdot A)$ of some quadratic real-valued matrix A for one or more scalar scaling factors T.

Usage

```
exponentiate_matrix(A, scalings=1, max_absolute_error=1e-3,
                    min_polynomials=1, max_polynomials=1000)
```

Arguments

A	A real-valued quadratic matrix of size N x N.
scalings	Vector of real-valued scalar scaling factors T, for each of which the exponential $\exp(T \cdot A)$ should be calculated.
max_absolute_error	
	Maximum allowed absolute error for the returned approximations. A smaller allowed error implies a greater computational cost as more matrix polynomials need to be included (see below). The returned approximations may have a greater error if the parameter `max_polynomials` is set too low.
min_polynomials	
	Minimum number of polynomials to include in the approximations (see equation below), even if `max_absolute_error` may be satisfied with fewer polynomials. If you don't know how to choose this, in most cases the default is fine. Note that regardless of `min_polynomials` and `max_absolute_error`, the number of polynomials used will not exceed `max_polynomials`.

max_polynomials

>   Maximum allowed number of polynomials to include in the approximations (see equation below). Meant to provide a safety limit for the amount of memory and the computation time required. For example, a value of 1000 means that up to 1000 matrices (powers of A) of size N x N may be computed and stored temporarily in memory. Note that if max_polynomials is too low, the requested accuracy (via max_absolute_error) may not be achieved. That said, for large trees more memory may be required to store the actual result rather than the intermediate polynomials, i.e. for purposes of saving RAM it doesn't make much sense to choose max_polynomials much smaller than the length of scalings.

Details

Discrete character evolution Markov models often involve repeated exponentiations of the same transition matrix along each edge of the tree (i.e. with the scaling T being the edge length). Matrix exponentiation can become a serious computational bottleneck for larger trees or large matrices (i.e. spanning multiple discrete states). This function pre-calculates polynomials $A^p/p!$ of the matrix, and then uses linear combinations of the same polynomials for each requested T:

$$\exp(T \cdot A) = \sum_{p=0}^{P} T^p \frac{A^p}{p!} + ...$$

This function thus becomes very efficient when the number of scaling factors is large (e.g. >10,000). The number of polynomials included is determined based on the specified max_absolute_error, and based on the largest (by magnitude) scaling factor requested. The function utilizes the balancing algorithm proposed by James et al (2014, Algorithm 3) and the scaling & squaring method (Moler and Van Loan, 2003) to improve the conditioning of the matrix prior to exponentiation.

Value

A 3D numeric matrix of size N x N x S, where N is the number of rows & column of the input matrix A and S is the length of scalings. The [r,c,s]-th element of this matrix is the entry in the r-th row and c-th column of $\exp(scalings[s] \cdot A)$.

Author(s)

Stilianos Louca

References

R. James, J. Langou and B. R. Lowery (2014). On matrix balancing and eigenvector computation. arXiv:1401.5766

C. Moler and C. Van Loan (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review. 45:3-49.

Examples

```
# create a random 5 x 5 matrix
A = get_random_mk_transition_matrix(Nstates=5, rate_model="ER")
```

```
# calculate exponentials exp(0.1*A) and exp(10*A)
exponentials = exponentiate_matrix(A, scalings=c(0.1,10))

# print 1st exponential: exp(0.1*A)
print(exponentials[,,1])

# print 2nd exponential: exp(10*A)
print(exponentials[,,2])
```

---

extend_tree_to_height   *Extend a rooted tree up to a specific height.*

Description

Given a rooted phylogenetic tree and a specific distance from the root ("new height"), elongate terminal edges (i.e. leading into tips) as needed so that all tips have a distance from the root equal to the new height. If a tip already extends beyond the specified new height, its incoming edge remains unchanged.

Usage

```
extend_tree_to_height(tree, new_height=NULL)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
new_height	Numeric, specifying the phylogenetic distance from the root to which tips are to be extended. If NULL or negative, then it is set to the maximum distance of any tip from the root.

Details

The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). All tip, edge and node indices remain unchanged. This function provides a quick-and-dirty way to make a tree ultrametric, or to correct small numerical inaccuracies in supposed-to-be ultrametric trees.

Value

A list with the following elements:

tree	A new rooted tree of class "phylo", representing the extended tree.
max_extension	Numeric. The largest elongation added to a terminal edge.

Author(s)

Stilianos Louca

See Also

trim_tree_at_height

Examples

```
# generate a random non-ultrametric tree
tree = generate_random_tree(list(birth_rate_intercept=1,death_rate_intercept=0.5),
                            max_time=1000,
                            coalescent=FALSE)$tree

# print min & max distance from root
span = get_tree_span(tree)
cat(sprintf("Min & max tip height = %g & %g\n",span$min_distance,span$max_distance))

# make tree ultrametric by extending terminal edges
extended = extend_tree_to_height(tree)$tree

# print new min & max distance from root
span = get_tree_span(extended)
cat(sprintf("Min & max tip height = %g & %g\n",span$min_distance,span$max_distance))
```

---

extract_fasttree_constraints

Extract tree constraints in FastTree alignment format.

Description

Given a rooted phylogenetic tree, extract binary constraints in FastTree alignment format. Every internal bifurcating node with more than 2 descending tips will constitute an separate constraint.

Usage

```
extract_fasttree_constraints(tree)
```

Arguments

tree   A rooted tree of class "phylo".

Details

This function can be used to define constraints based on a backbone subtree, to be used to generate a larger tree using FastTree (as of v2.1.11). Only bifurcating nodes with at least 3 descending tips are used as constraints.

The constraints are returned as a 2D matrix; the actual fasta file with the constraint alignments can be written easily from this matrix. For more details on FastTree constraints see the original FastTree documentation.

Value

A list with the following elements:

Nconstraints   Integer, specifying the number of constraints extracted.

constraints   2D character matrix of size Ntips x Nconstraints, with values '0', '1' or '-', specifying which side ("left" or "right") of a constraint (node) each tip is found on.

constraint2node

Integer vector of size Nconstraints, with values in 1,..,Nnodes, specifying the original node index used to define each constraint.

Author(s)

Stilianos Louca

Examples

```
# generate a simple rooted tree, with tip names tip.1, tip.2, ...
Ntips = 10
tree  = generate_random_tree(list(birth_rate_intercept=1),
                             max_tips=Ntips,
                             tip_basename="tip.")$tree

# extract constraints
constraints = castor::extract_fasttree_constraints(tree)$constraints

# print constraints to screen in fasta format
cat(paste(sapply(1:Ntips,
    FUN=function(tip) sprintf(">%s\n%s\n",tree$tip.label[tip],
    paste(as.character(constraints[tip,]),collapse=""))),collapse=""))
```

---

find_farthest_tips   *Find farthest tip to each tip & node of a tree.*

Description

Given a rooted phylogenetic tree and a subset of potential target tips, for each tip and node in the tree find the farthest target tip. The set of target tips can also be taken as the whole set of tips in the tree.

Usage

```
find_farthest_tips( tree,
                    only_descending_tips = FALSE,
                    target_tips          = NULL,
                    as_edge_counts       = FALSE,
                    check_input          = TRUE)
```

Arguments

tree
:   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

only_descending_tips
:   A logical indicating whether the farthest tip to a node or tip should be chosen from its descending tips only. If FALSE, then the whole set of possible target tips is considered.

target_tips
:   Optional integer vector or character vector listing the subset of target tips to restrict the search to. If an integer vector, this should list tip indices (values in 1,..,Ntips). If a character vector, it should list tip names (in this case tree$tip.label must exist). If target_tips is NULL, then all tips of the tree are considered as target tips.

as_edge_counts
:   Logical, specifying whether to count phylogenetic distance in terms of edge counts instead of cumulative edge lengths. This is the same as setting all edge lengths to 1.

check_input
:   Logical, whether to perform basic validations of the input data. If you know for certain that your input is valid, you can set this to FALSE to reduce computation time.

Details

If only_descending_tips is TRUE, then only descending target tips are considered when searching for the farthest target tip of a node/tip. In that case, if a node/tip has no descending target tip, its farthest target tip is set to NA. If tree$edge.length is missing or NULL, then each edge is assumed to have length 1. The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child).

The asymptotic time complexity of this function is O(Nedges), where Nedges is the number of edges in the tree.

Value

A list with the following elements:

farthest_tip_per_tip
:   An integer vector of size Ntips, listing the farthest target tip for each tip in the tree. Hence, farthest_tip_per_tip[i] is the index of the farthest tip (from the set of target tips), with respect to tip i (where i=1,..,Ntips). Some values may appear multiple times in this vector, if multiple tips share the same farthest target tip.

farthest_tip_per_node
:   An integer vector of size Nnodes, listing the index of the farthest target tip for each node in the tree. Hence, farthest_tip_per_node[i] is the index of the farthest tip (from the set of target tips), with respect to node i (where i=1,..,Nnodes). Some values may appear multiple times in this vector, if multiple nodes share the same farthest target tip.

farthest_distance_per_tip

> Integer vector of size Ntips. Phylogenetic ("patristic") distance of each tip in the tree to its farthest target tip. If only_descending_tips was set to TRUE, then farthest_distance_per_tip[i] will be set to infinity for any tip i that is not a target tip.

farthest_distance_per_node

> Integer vector of size Nnodes. Phylogenetic ("patristic") distance of each node in the tree to its farthest target tip. If only_descending_tips was set to TRUE, then farthest_distance_per_node[i] will be set to infinity for any node i that has no descending target tips.

Author(s)

Stilianos Louca

References

M. G. I. Langille, J. Zaneveld, J. G. Caporaso et al (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology. 31:814-821.

See Also

find_nearest_tips

Examples

```
# generate a random tree
Ntips = 1000
parameters = list(birth_rate_intercept=1,death_rate_intercept=0.9)
tree = generate_random_tree(parameters,Ntips,coalescent=FALSE)$tree

# pick a random set of "target" tips
target_tips = sample.int(n=Ntips, size=5, replace=FALSE)

# find farthest target tip to each tip & node in the tree
results = find_farthest_tips(tree, target_tips=target_tips)

# plot histogram of distances to target tips (across all tips of the tree)
distances = results$farthest_distance_per_tip
hist(distances, breaks=10, xlab="farthest distance", ylab="number of tips", prob=FALSE);
```

---

find_nearest_tips   *Find nearest tip to each tip & node of a tree.*

Description

Given a rooted phylogenetic tree and a subset of potential target tips, for each tip and node in the tree find the nearest target tip. The set of target tips can also be taken as the whole set of tips in the tree.

Usage

```
find_nearest_tips(tree,
                  only_descending_tips = FALSE,
                  target_tips          = NULL,
                  as_edge_counts       = FALSE,
                  check_input          = TRUE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with
                no incoming edge.

only_descending_tips

                A logical indicating whether the nearest tip to a node or tip should be chosen
                from its descending tips only. If FALSE, then the whole set of possible target
                tips is considered.

target_tips   Optional integer vector or character vector listing the subset of target tips to
                restrict the search to. If an integer vector, this should list tip indices (val-
                ues in 1,..,Ntips). If a character vector, it should list tip names (in this case
                tree$tip.label must exist). If target_tips is NULL, then all tips of the tree
                are considered as target tips.

as_edge_counts Logical, specifying whether to count phylogenetic distance in terms of edge
                counts instead of cumulative edge lengths. This is the same as setting all edge
                lengths to 1.

check_input   Logical, whether to perform basic validations of the input data. If you know for
                certain that your input is valid, you can set this to FALSE to reduce computation
                time.

Details

Langille et al. (2013) introduced the Nearest Sequenced Taxon Index (NSTI) as a measure for how
well a set of microbial operational taxonomic units (OTUs) is represented by a set of sequenced
genomes of related organisms. Specifically, the NSTI of a microbial community is the average phy-
logenetic distance of any OTU in the community, to the closest relative with an available sequenced
genome ("target tips"). In analogy to the NSTI, the function find_nearest_tips provides a means
to find the nearest tip (from a subset of target tips) to each tip and node in a phylogenetic tree,
together with the corresponding phylogenetic ("patristic") distance.

If only_descending_tips is TRUE, then only descending target tips are considered when searching
for the nearest target tip of a node/tip. In that case, if a node/tip has no descending target tip, its
nearest target tip is set to NA. If tree$edge.length is missing or NULL, then each edge is assumed
to have length 1. The tree may include multi-furcations as well as mono-furcations (i.e. nodes with
only one child).

The asymptotic time complexity of this function is O(Nedges), where Nedges is the number of
edges in the tree.

Value

A list with the following elements:

nearest_tip_per_tip

>   An integer vector of size Ntips, listing the nearest target tip for each tip in the
>   tree. Hence, `nearest_tip_per_tip[i]` is the index of the nearest tip (from the
>   set of target tips), with respect to tip i (where i=1,..,Ntips). Some values may
>   appear multiple times in this vector, if multiple tips share the same nearest target
>   tip.

nearest_tip_per_node

>   An integer vector of size Nnodes, listing the index of the nearest target tip
>   for each node in the tree. Hence, `nearest_tip_per_node[i]` is the index
>   of the nearest tip (from the set of target tips), with respect to node i (where
>   i=1,..,Nnodes). Some values may appear multiple times in this vector, if multi-
>   ple nodes share the same nearest target tip.

nearest_distance_per_tip

>   Integer vector of size Ntips. Phylogenetic ("patristic") distance of each tip in the
>   tree to its nearest target tip. If `only_descending_tips` was set to TRUE, then
>   `nearest_distance_per_tip[i]` will be set to infinity for any tip i that is not a
>   target tip.

nearest_distance_per_node

>   Integer vector of size Nnodes. Phylogenetic ("patristic") distance of each node
>   in the tree to its nearest target tip. If `only_descending_tips` was set to TRUE,
>   then `nearest_distance_per_node[i]` will be set to infinity for any node i that
>   has no descending target tips.

Author(s)

Stilianos Louca

References

M. G. I. Langille, J. Zaneveld, J. G. Caporaso et al (2013). Predictive functional profiling of micro-
bial communities using 16S rRNA marker gene sequences. Nature Biotechnology. 31:814-821.

See Also

find_farthest_tips

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# pick a random set of "target" tips
target_tips = sample.int(n=Ntips, size=as.integer(Ntips/10), replace=FALSE)

# find nearest target tip to each tip & node in the tree
results = find_nearest_tips(tree, target_tips=target_tips)

# plot histogram of distances to target tips (across all tips of the tree)
distances = results$nearest_distance_per_tip
```

```
hist(distances, breaks=10, xlab="nearest distance", ylab="number of tips", prob=FALSE);
```

---

find_root   *Find the root of a tree.*

Description

Find the root of a phylogenetic tree. The root is defined as the unique node with no parent.

Usage

```
find_root(tree)
```

Arguments

tree   A tree of class "phylo". If the tree is not rooted, the function will return NA.

Details

By convention, the root of a "phylo" tree is typically the first node (i.e. with index Ntips+1), however this is not always guaranteed. This function finds the root of a tree by searching for the node with no parent. If no such node exists, NA is returned. If multiple such nodes exist, NA is returned. If any node has more than 1 parent, NA is returned. Hence, this function can be used to test if a tree is rooted purely based on the edge structure, assuming that the tree is connected (i.e. not a forest).

The asymptotic time complexity of this function is O(Nedges), where Nedges is the number of edges in the tree.

Value

Index of the tree's root, as listed in tree$edge. An integer ranging from Ntips+1 to Ntips+Nnodes, where Ntips and Nnodes is the number of tips and nodes in the tree, respectively. By convention, the root will typically be Ntips+1 but this is not guaranteed.

Author(s)

Stilianos Louca

See Also

[find_root_of_monophyletic_tips](), [root_at_node](), [root_at_midpoint]()

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# reroot the tree at the 20-th node
new_root_node = 20
tree = root_at_node(tree, new_root_node, update_indices=FALSE)

# find new root index and compare with expectation
cat(sprintf("New root is %d, expected at %d\n",find_root(tree),new_root_node+Ntips))
```

---

find_root_of_monophyletic_tips

Find the node or tip that, as root, would make a set of target tips monophyletic.

Description

Given a tree (rooted or unrooted) and a specific set of target tips, this function finds the tip or node that, if turned into root, would make a set of target tips a monophyletic group that either descends from a single child of the new root (if as_MRCA==FALSE) or whose MRCA is the new root (if as_MRCA==TRUE).

Usage

```
find_root_of_monophyletic_tips(tree, monophyletic_tips, as_MRCA=TRUE, is_rooted=FALSE)
```

Arguments

tree   A tree object of class "phylo". Can be unrooted or rooted.

monophyletic_tips

Character or integer vector, specifying the names or indices, respectively, of the target tips that should be turned monophyletic. If an integer vector, its elements must be between 1 and Ntips. If a character vector, its elements must be elements in tree$tip.label.

as_MRCA   Logical, specifying whether the new root should become the MRCA of the target tips. If FALSE, the new root is chosen such that the MRCA of the target tips is the child of the new root.

is_rooted   Logical, specifying whether the input tree can be assumed to be rooted. If you are sure that the input tree is rooted, set this to TRUE for computational efficiency, otherwise to be on the safe side set this to FALSE.

Details

The input tree may include an arbitrary number of incoming and outgoing edges per node (but only one edge per tip), and the direction of these edges can be arbitrary. Of course, the undirected graph defined by all edges must still be a valid tree (i.e. a connected acyclic graph). Note that this function does not change the tree, it just determines which tip or node should be made root for the target tips to be a monophyletic group.

The asymptotic time complexity of this function is O(Nedges).

Value

A single integer between 1 and (Ntips+Nnodes), specifying the index of the tip or node that, if made root, would make the target tips monophyletic. If this was not possible, NA is returned.

Author(s)

Stilianos Louca

See Also

find_root

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# pick a random node and find all descending tips
MRCA = sample.int(tree$Nnode,size=1)
monophyletic_tips = get_subtree_at_node(tree, MRCA)$new2old_tip

# change root of tree (change edge directions)
tree = root_at_node(tree, new_root_node=10, update_indices=FALSE)

# determine root that would make target tips monophyletic
new_root = find_root_of_monophyletic_tips(tree, monophyletic_tips, as_MRCA=TRUE, is_rooted=FALSE)

# compare expectation with result
cat(sprintf("MRCA = %d, new root node=%d\n",MRCA,new_root-Ntips))
```

---

fit_and_compare_bm_models

Fit and compare Brownian Motion models for multivariate trait evolution between two data sets.

Description

Given two rooted phylogenetic trees and states of one or more continuous (numeric) traits on the trees' tips, fit a multivariate Brownian motion model of correlated evolution to each data set and compare the fitted models. This function estimates the diffusivity matrix for each data set (i.e., each tree/tip-states set) via maximum-likelihood and assesses whether the log-difference between the two fitted diffusivity matrixes is statistically significant, under the null hypothesis that the two data sets exhibit the same diffusivity. Optionally, multiple trees can be used as input for each data set, under the assumption that the trait evolved on each tree according to the same BM model. For more details on how BM is fitted to each data set see the function `fit_bm_model`.

Usage

```
fit_and_compare_bm_models( trees1,
                           tip_states1,
                           trees2,
                           tip_states2,
                           Nbootstraps   = 0,
                           Nsignificance = 0,
                           check_input   = TRUE,
                           verbose       = FALSE,
                           verbose_prefix = "")
```

Arguments

trees1	Either a single rooted tree or a list of rooted trees, of class "phylo", corresponding to the first data set on which a BM model is to be fitted. Edge lengths are assumed to represent time intervals or a similarly interpretable phylogenetic distance.
tip_states1	Numeric state of each trait at each tip in each tree in the first data set. If trees1 is a single tree, then tip_states1 must either be a numeric vector of size Ntips or a 2D numeric matrix of size Ntips x Ntraits, listing the trait states for each tip in the tree. If trees1 is a list of Ntrees trees, then tip_states1 must be a list of length Ntrees, each element of which lists the trait states for the corresponding tree (as a vector or 2D matrix, similarly to the single-tree case).
trees2	Either a single rooted tree or a list of rooted trees, of class "phylo", corresponding to the second data set on which a BM model is to be fitted. Edge lengths are assumed to represent time intervals or a similarly interpretable phylogenetic distance.
tip_states2	Numeric state of each trait at each tip in each tree in the second data set, similarly to tip_states1.
Nbootstraps	Integer, specifying the number of parametric bootstraps to perform for calculating the confidence intervals of BM diffusivities fitted to each data set. If <=0, no bootstrapping is performed.
Nsignificance	Integer, specifying the number of simulations to perform for assessing the statistical significance of the log-transformed difference between the diffusivities fitted to the two data sets, i.e. of $|\log(D_1) - \log(D_2)|$. Set to 0 to not calculate the statistical significance. See below for additional details.

check_input   Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

verbose   Logical, specifying whether to print progress report messages to the screen.

verbose_prefix Character, specifying a prefix to include in front of progress report messages on each line. Only relevant if verbose==TRUE.

Details

For details on the Brownian Motion model see [fit_bm_model](#) and [simulate_bm_model](#). This function separately fits a single-variate or multi-variate BM model with constant diffusivity (diffusivity matrix, in the multivariate case) to each data set; internally, this function applies fit_bm_model to each data set.

If Nsignificance>0, the statistical significance of the log-transformed difference of the two fitted diffusivity matrixes, $|\log(D_1) - \log(D_2)|$, is assessed, under the null hypothesis that both data sets were generated by the same common BM model. The diffusivity of this common BM model is estimated by fitting to both datasets at once, i.e. after merging the two datasets into a single dataset of trees and tip states (see return variable fit_common below). For each of the Nsignificance random simulations of the common BM model on the two tree sets, the diffusivities are again separately fitted on the two simulated sets and the resulting log-difference is compared to the one of the original data sets. The returned significance is the probability that the diffusivities would have a log-difference larger than the observed one, if the two data sets had been generated under the common BM model.

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. The tree may include multifurcations (i.e. nodes with more than 2 children) as well as monofurcations (i.e. nodes with only one child). Note that multifurcations are internally expanded to bifurcations, prior to model fitting.

Value

A list with the following elements:

success   Logical, indicating whether the fitting was successful for both data sets. If FALSE, then an additional return variable, error, will contain a description of the error; in that case all other return variables may be undefined.

fit1   A named list containing the fitting results for the first data set, in the same format as returned by [fit_bm_model](#). In particular, the diffusivity fitted to the first data set will be stored in fit1$diffusivity.

fit2   A named list containing the fitting results for the second data set, in the same format as returned by [fit_bm_model](#). In particular, the diffusivity fitted to the second data set will be stored in fit2$diffusivity.

log_difference The absolute difference between the log-transformed diffusivities, i.e. $|\log(D_1) - \log(D_2)|$. In the multivariate case, this will be a matrix of size Ntraits x Ntraits.

significance   Numeric, statistical significance of the observed log-difference under the null hypothesis that the two data sets were generated by a common BM model. Only returned if Nsignificance>0.

fit_common   A named list containing the fitting results for the two data sets combined, in the
                  same format as returned by [fit_bm_model](). The common diffusivity, fit_common$diffusivity
                  is used for the random simulations when assessing the statistical significance of
                  the log-difference of the separately fitted diffusivities. Only returned if Nsignificance>0.

Author(s)

Stilianos Louca

References

J. Felsenstein (1985). Phylogenies and the Comparative Method. The American Naturalist. 125:1-
15.

See Also

[simulate_bm_model](), [fit_bm_model](), [get_independent_contrasts]()

Examples

```
# simulate distinct BM models on two random trees
D1 = 1
D2 = 2
tree1 = generate_random_tree(list(birth_rate_factor=1),max_tips=100)$tree
tree2 = generate_random_tree(list(birth_rate_factor=1),max_tips=100)$tree
tip_states1 = simulate_bm_model(tree1, diffusivity = D1)$tip_states
tip_states2 = simulate_bm_model(tree2, diffusivity = D2)$tip_states

# fit and compare BM models between the two data sets
fit = fit_and_compare_bm_models(trees1         = tree1,
                                tip_states1    = tip_states1,
                                trees2         = tree2,
                                tip_states2    = tip_states2,
                                Nbootstraps    = 100,
                                Nsignificance  = 100)

# print summary of results
cat(sprintf("Fitted D1 = %g, D2 = %g, significance of log-diff. = %g\n",
            fit$fit1$diffusivity, fit$fit2$diffusivity, fit$significance))
```

---

fit_and_compare_sbm_const

                       Fit and compare Spherical Brownian Motion models for diffusive ge-
                       ographic dispersal between two data sets.

Description

Given two rooted phylogenetic trees and geographic coordinates of the trees' tips, fit a Spherical Brownian Motion (SBM) model of diffusive geographic dispersal with constant diffusivity to each tree and compare the fitted models. This function estimates the diffusivity (D) for each data set (i.e., each set of trees + tip-coordinates) via maximum-likelihood and assesses whether the log-difference between the two fitted diffusivities is statistically significant, under the null hypothesis that the two data sets exhibit the same diffusivity. Optionally, multiple trees can be used as input for each data set, under the assumption that dispersal occurred according to the same diffusivity in each tree of that dataset. For more details on how SBM is fitted to each data set see the function `fit_sbm_const`.

Usage

```
fit_and_compare_sbm_const( trees1,
                           tip_latitudes1,
                           tip_longitudes1,
                           trees2,
                           tip_latitudes2,
                           tip_longitudes2,
                           radius,
                           planar_approximation  = FALSE,
                           only_basal_tip_pairs  = FALSE,
                           only_distant_tip_pairs = FALSE,
                           min_MRCA_time         = 0,
                           max_MRCA_age          = Inf,
                           max_phylodistance     = Inf,
                           min_diffusivity       = NULL,
                           max_diffusivity       = NULL,
                           Nbootstraps           = 0,
                           Nsignificance         = 0,
                           SBM_PD_functor        = NULL,
                           verbose               = FALSE,
                           verbose_prefix        = "")
```

Arguments

trees1   Either a single rooted tree or a list of rooted trees, of class "phylo", corresponding to the first data set on which an SBM model is to be fitted. Edge lengths are assumed to represent time intervals or a similarly interpretable phylogenetic distance.

tip_latitudes1 Numeric vector listing the latitude (in decimal degrees) of each tip in each tree in the first data set. If `trees1` is a single tree, then `tip_latitudes1` must be a numeric vector of size Ntips, listing the latitudes for each tip in the tree. If `trees1` is a list of Ntrees trees, then `tip_latitudes1` must be a list of length Ntrees, each element of which lists the latitudes for the corresponding tree (as a vector, similarly to the single-tree case).

tip_longitudes1

                Similar to `tip_latitudes1`, but listing longitudes (in decimal degrees) of each

                          tip in each tree in the first data set.

trees2   Either a single rooted tree or a list of rooted trees, of class "phylo", correspond-
                          ing to the second data set on which an SBM model is to be fitted. Edge lengths
                          are assumed to represent time intervals or a similarly interpretable phylogenetic
                          distance.

tip_latitudes2   Numeric vector listing the latitude (in decimal degrees) of each tip in each tree
                          in the second data set, similarly to tip_latitudes1.

tip_longitudes2
                          Numeric vector listing the longitude (in decimal degrees) of each tip in each tree
                          in the second data set, similarly to tip_longitudes1.

radius   Strictly positive numeric, specifying the radius of the sphere. For Earth, the
                          mean radius is 6371 km.

planar_approximation
                          Logical, specifying whether to estimate the diffusivity based on a planar approx-
                          imation of the SBM model, i.e. by assuming that geographic distances between
                          tips are as if tips are distributed on a 2D cartesian plane. This approximation is
                          only accurate if geographical distances between tips are small compared to the
                          sphere's radius.

only_basal_tip_pairs
                          Logical, specifying whether to only compare immediate sister tips, i.e., tips con-
                          nected through a single parental node.

only_distant_tip_pairs
                          Logical, specifying whether to only compare tips at distinct geographic loca-
                          tions.

min_MRCA_time   Numeric, specifying the minimum allowed time (distance from root) of the most
                          recent common ancestor (MRCA) of sister tips considered in the fitting. In other
                          words, an independent contrast is only considered if the two sister tips' MRCA
                          has at least this distance from the root. Set min_MRCA_time<=0 to disable this
                          filter.

max_MRCA_age   Numeric, specifying the maximum allowed age (distance from youngest tip) of
                          the MRCA of sister tips considered in the fitting. In other words, an independent
                          contrast is only considered if the two sister tips' MRCA has at most this age
                          (time to present). Set max_MRCA_age=Inf to disable this filter.

max_phylodistance
                          Numeric, maximum allowed geodistance for an independent contrast to be in-
                          cluded in the SBM fitting. Set max_phylodistance=Inf to disable this filter.

min_diffusivity
                          Non-negative numeric, specifying the minimum possible diffusivity. If NULL,
                          this is automatically chosen.

max_diffusivity
                          Non-negative numeric, specifying the maximum possible diffusivity. If NULL,
                          this is automatically chosen.

Nbootstraps   Integer, specifying the number of parametric bootstraps to perform for calculat-
                          ing the confidence intervals of SBM diffusivities fitted to each data set. If <=0,
                          no bootstrapping is performed.

Nsignificance	Integer, specifying the number of simulations to perform for assessing the statistical significance of the linear difference and log-transformed difference between the diffusivities fitted to the two data sets, i.e. of $|D_1 - D_2|$ and of $|\log(D_1) - \log(D_2)|$. Set to 0 to not calculate statistical significances. See below for additional details.
SBM_PD_functor	SBM probability density functor object. Used internally and for debugging purposes. Unless you know what you're doing, you should keep this NULL.
verbose	Logical, specifying whether to print progress report messages to the screen.
verbose_prefix	Character, specifying a prefix to include in front of progress report messages on each line. Only relevant if verbose==TRUE.

Details

For details on the Spherical Brownian Motion model see `fit_sbm_const` and `simulate_sbm`. This function separately fits an SBM model with constant diffusivity to each of two data sets; internally, this function applies fit_sbm_const to each data set.

If Nsignificance>0, the statistical significance of the linear difference ($|D_1 - D_2|$) and log-transformed difference ($|\log(D_1) - \log(D_2)|$) of the two fitted diffusivities is assessed under the null hypothesis that both data sets were generated by the same common SBM model. The diffusivity of this common SBM model is estimated by fitting to both datasets at once, i.e. after merging the two datasets into a single dataset of trees and tip coordinates (see return variable fit_common below). For each of the Nsignificance random simulations of the common SBM model on the two tree sets, the diffusivities are again separately fitted on the two simulated sets and the resulting difference and log-difference is compared to those of the original data sets. The returned lin_significance (or log_significance) is the probability that the diffusivities would have a difference (or log-difference) larger than the observed one, if the two data sets had been generated under the common SBM model.

If edge.length is missing from one of the input trees, each edge in the tree is assumed to have length 1. Trees may include multifurcations as well as monofurcations, however multifurcations are internally expanded into bifurcations by adding dummy nodes.

Value

A list with the following elements:

success	Logical, indicating whether the fitting was successful for both data sets. If FALSE, then an additional return variable, error, will contain a description of the error; in that case all other return variables may be undefined.
fit1	A named list containing the fitting results for the first data set, in the same format as returned by `fit_sbm_const`. In particular, the diffusivity fitted to the first data set will be stored in fit1$diffusivity.
fit2	A named list containing the fitting results for the second data set, in the same format as returned by `fit_sbm_const`. In particular, the diffusivity fitted to the second data set will be stored in fit2$diffusivity.
lin_difference	The absolute difference between the two diffusivities, i.e. $|D_1 - D_2|$.
log_difference	The absolute difference between the two log-transformed diffusivities, i.e. $|\log(D_1) - \log(D_2)|$.

lin_significance

> Numeric, statistical significance of the observed lin-difference under the null hypothesis that the two data sets were generated by a common SBM model. Only returned if Nsignificance>0.

log_significance

> Numeric, statistical significance of the observed log-difference under the null hypothesis that the two data sets were generated by a common SBM model. Only returned if Nsignificance>0.

fit_common   A named list containing the fitting results for the two data sets combined, in the same format as returned by `fit_sbm_const`. The common diffusivity, fit_common$diffusivity is used for the random simulations when assessing the statistical significance of the lin-difference and log-difference of the separately fitted diffusivities. Only returned if Nsignificance>0.

Author(s)

Stilianos Louca

References

S. Louca (in review as of 2020). Phylogeographic estimation and simulation of global diffusive dispersal. Systematic Biology.

See Also

[simulate_sbm](), [fit_sbm_const](), [fit_sbm_linear](), [fit_sbm_parametric]()

Examples

```
## Not run:
# simulate distinct SBM models on two random trees
radius = 6371   # Earth's radius
D1     = 1      # diffusivity on 1st tree
D2     = 3      # diffusivity on 2nd tree
tree1 = generate_random_tree(list(birth_rate_factor=1),max_tips=100)$tree
tree2 = generate_random_tree(list(birth_rate_factor=1),max_tips=100)$tree
sim1  = simulate_sbm(tree=tree1, radius=radius, diffusivity=D1)
sim2  = simulate_sbm(tree=tree2, radius=radius, diffusivity=D2)
tip_latitudes1  = sim1$tip_latitudes
tip_longitudes1 = sim1$tip_longitudes
tip_latitudes2  = sim2$tip_latitudes
tip_longitudes2 = sim2$tip_longitudes

# fit and compare SBM models between the two hypothetical data sets
fit = fit_and_compare_sbm_const(trees1          = tree1,
                                tip_latitudes1  = tip_latitudes1,
                                tip_longitudes1 = tip_longitudes1,
                                trees2          = tree2,
                                tip_latitudes2  = tip_latitudes2,
                                tip_longitudes2 = tip_longitudes2,
                                radius          = radius,
```

```
                                      Nbootstraps    = 0,
                                      Nsignificance  = 100)

# print summary of results
cat(sprintf("Fitted D1 = %g, D2 = %g, significance of log-diff. = %g\n",
            fit$fit1$diffusivity, fit$fit2$diffusivity, fit$log_significance))

## End(Not run)
```

---

fit_bm_model   *Fit a Brownian Motion model for multivariate trait evolution.*

Description

Given a rooted phylogenetic tree and states of one or more continuous (numeric) traits on the tree's tips, fit a multivariate Brownian motion model of correlated co-evolution of these traits. This estimates a single diffusivity matrix, which describes the variance-covariance structure of each trait's random walk. The model assumes a fixed diffusivity matrix on the entire tree. Optionally, multiple trees can be used as input, under the assumption that the trait evolved on each tree according to the same BM model.

Usage

```
fit_bm_model(  trees,
               tip_states,
               isotropic   = FALSE,
               Nbootstraps = 0,
               check_input = TRUE)
```

Arguments

trees   Either a single rooted tree or a list of rooted trees, of class "phylo". The root of
                 each tree is assumed to be the unique node with no incoming edge. Edge lengths
                 are assumed to represent time intervals or a similarly interpretable phylogenetic
                 distance.

tip_states   Numeric state of each trait at each tip in each tree. If trees was a single tree,
                 then tip_states must either be a numeric vector of size Ntips or a 2D numeric
                 matrix of size Ntips x Ntraits, listing the trait states for each tip in the tree. If
                 trees is a list of Ntrees trees, then tip_states must be a list of length Ntrees,
                 each element of which lists the trait states for the corresponding tree (as a vector
                 or 2D matrix, similarly to the single-tree case).

isotropic   Logical, specifying whether diffusion should be assumed to be isotropic (i.e., in-
                 dependent of the direction). Hence, if isotropic=TRUE, then the diffusivity ma-
                 trix is forced to be diagonal, with all entries being equal. If isotropic=FALSE,
                 an arbitrary diffusivity matrix is fitted (i.e., the diffusivity matrix is only con-
                 strained to be symmetric and non-negative definite).

Nbootstraps	Integer, specifying the number of parametric bootstraps to perform for calculating the confidence intervals. If <=0, no bootstrapping is performed.
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to `FALSE` to reduce computation.

Details

The BM model is defined by the stochastic differential equation

$$dX = \sigma \cdot dW$$

where W is a multidimensional Wiener process with Ndegrees independent components and σ is a matrix of size Ntraits x Ndegrees. Alternatively, the same model can be defined as a Fokker-Planck equation for the probability density ρ:

$$\frac{\partial \rho}{\partial t} = \sum_{i,j} D_{ij} \frac{\partial^2 \rho}{\partial x_i \partial x_j}.$$

The matrix D is referred to as the diffusivity matrix (or diffusion tensor), and $2D = \sigma \cdot \sigma^T$. Note that σ can be obtained from D by means of a Cholesky decomposition.

The function uses phylogenetic independent contrasts (Felsenstein, 1985) to retrieve independent increments of the multivariate random walk. The diffusivity matrix D is then fitted using maximum-likelihood on the intrinsic geometry of positive-definite matrices. If multiple trees are provided as input, then independent contrasts are extracted from all trees and combined into a single set of independent contrasts (i.e., as if they had been extracted from a single tree).

If `tree$edge.length` is missing, each edge in the tree is assumed to have length 1. The tree may include multifurcations (i.e. nodes with more than 2 children) as well as monofurcations (i.e. nodes with only one child). Note that multifurcations are internally expanded to bifurcations, prior to model fitting.

Value

A list with the following elements:

success	Logical, indicating whether the fitting was successful. If `FALSE`, then an additional return variable, `error`, will contain a description of the error; in that case all other return variables may be undefined.
diffusivity	Either a single non-negative number (if `tip_states` was a vector) or a 2D quadratic non-negative-definite matrix (if `tip_states` was a 2D matrix). The fitted diffusivity matrix of the multivariate Brownian motion model.
loglikelihood	The log-likelihood of the fitted model, given the provided tip states data.
Ncontrasts	Integer, number of independent contrasts used to estimate the diffusivity. This corresponds to the number of independent data points used.
standard_errors	
	Either a single numeric or a 2D numeric matrix of size Ntraits x Ntraits, listing the estimated standard errors of the estimated diffusivity, based on parametric bootstrapping. Only returned if Nbootstraps>0.

CI50lower	Either a single numeric or a 2D numeric matrix of size Ntraits x Ntraits, listing the lower bounds of the 50% confidence interval for the estimated diffusivity (25-75% percentile), based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI50upper	Either a single numeric or a 2D numeric matrix of size Ntraits x Ntraits, listing the upper bound of the 50% confidence interval for the estimated diffusivity, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI95lower	Either a single numeric or a 2D numeric matrix of size Ntraits x Ntraits, listing the lower bound of the 95% confidence interval for the estimated diffusivity (2.5-97.5% percentile), based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI95upper	Either a single numeric or a 2D numeric matrix of size Ntraits x Ntraits, listing the upper bound of the 95% confidence interval for the estimated diffusivity, based on parametric bootstrapping. Only returned if Nbootstraps>0.
consistency	Numeric between 0 and 1, estimated consistency of the data with the fitted model. If L denotes the loglikelihood of new data generated by the fitted model (under the same model) and M denotes the expectation of L, then consistency is the probability that $\|L - M\|$ will be greater or equal to $\|X - M\|$, where X is the loglikelihood of the original data under the fitted model. Only returned if Nbootstraps>0. A low consistency (e.g., <0.05) indicates that the fitted model is a poor description of the data. See Lindholm et al. (2019) for background.

Author(s)

Stilianos Louca

References

J. Felsenstein (1985). Phylogenies and the Comparative Method. The American Naturalist. 125:1-15.

A. Lindholm, D. Zachariah, P. Stoica, T. B. Schoen (2019). Data consistency approach to model validation. IEEE Access. 7:59788-59796.

See Also

simulate_bm_model, get_independent_contrasts

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1), 10000)$tree


# Example 1: Scalar case
# - - - - - - - - - - - - - -
# simulate scalar continuous trait on the tree
D = 1
tip_states = simulate_bm_model(tree, diffusivity=D)$tip_states
```

```
# estimate original diffusivity from the generated data
fit = fit_bm_model(tree, tip_states)
cat(sprintf("True D=%g, fitted D=%g\n",D,fit$diffusivity))


# Example 2: Multivariate case
# - - - - - - - - - - - - - - - -
# simulate vector-valued continuous trait on the tree
D = get_random_diffusivity_matrix(Ntraits=5)
tip_states = simulate_bm_model(tree, diffusivity=D)$tip_states

# estimate original diffusivity matrix from the generated data
fit = fit_bm_model(tree, tip_states)

# compare true and fitted diffusivity matrices
cat("True D:\n"); print(D)
cat("Fitted D:\n"); print(fit$diffusivity)
```

---

fit_hbds_model_on_grid

Fit a homogenous birth-death-sampling model on a discrete time grid.

Description

Given a timetree (potentially sampled through time and not necessarily ultrametric), fit a homogenous birth-death-sampling (HBDS) model in which speciation, extinction and lineage sampling occurs at some continuous (Poissonian) rates λ, μ and ψ, which are defined on a fixed grid of discrete time points and assumed to vary polynomially between grid points. Sampled lineages are kept in the pool of extant lineages at some "retention probability" κ, which may also depend on time. In addition, this model can include concentrated sampling attempts (CSAs) at a finite set of discrete time points $t_1, .., t_m$. "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction/sampling rates. Every HBDS model is thus defined based on the values that λ, μ, ψ and κ take over time, as well as the sampling probabilities $\rho_1, .., \rho_m$ and retention probabilities $\kappa_1, .., \kappa_m$ during the concentrated sampling attempts. This function estimates the values of λ, μ, ψ and κ on each grid point, as well as the $\rho_1, .., \rho_m$ and $\kappa_1, .., \kappa_m$, by maximizing the corresponding likelihood of the timetree. Special cases of this model (when rates are piecewise constant through time) are sometimes known as "birth-death-skyline plots" in the literature (Stadler 2013). In epidemiology, these models are often used to describe the phylogenies of viral strains sampled over the course of the epidemic.

Usage

```
fit_hbds_model_on_grid( tree,
                        root_age          = NULL,
                        oldest_age        = NULL,
                        age_grid          = NULL,
                        CSA_ages          = NULL,
                        min_lambda        = 0,
```

```
max_lambda              = +Inf,
min_mu                  = 0,
max_mu                  = +Inf,
min_psi                 = 0,
max_psi                 = +Inf,
min_kappa               = 0,
max_kappa               = 1,
min_CSA_probs           = 0,
max_CSA_probs           = 1,
min_CSA_kappas          = 0,
max_CSA_kappas          = 1,
guess_lambda            = NULL,
guess_mu                = NULL,
guess_psi               = NULL,
guess_kappa             = NULL,
guess_CSA_probs         = NULL,
guess_CSA_kappas        = NULL,
fixed_lambda            = NULL,
fixed_mu                = NULL,
fixed_psi               = NULL,
fixed_kappa             = NULL,
fixed_CSA_probs         = NULL,
fixed_CSA_kappas        = NULL,
fixed_age_grid          = NULL,
const_lambda            = FALSE,
const_mu                = FALSE,
const_psi               = FALSE,
const_kappa             = FALSE,
const_CSA_probs         = FALSE,
const_CSA_kappas        = FALSE,
splines_degree          = 1,
condition               = "auto",
ODE_relative_dt         = 0.001,
ODE_relative_dy         = 1e-3,
CSA_age_epsilon         = NULL,
Ntrials                 = 1,
max_start_attempts      = 1,
Nthreads                = 1,
max_model_runtime       = NULL,
Nbootstraps             = 0,
Ntrials_per_bootstrap   = NULL,
fit_control             = list(),
focal_param_values      = NULL,
verbose                 = FALSE,
diagnostics             = FALSE,
verbose_prefix          = "")
```

Arguments

tree	A timetree of class "phylo", representing the time-calibrated reconstructed phylogeny of a set of extant and/or extinct species. Tips of the tree are interpreted as terminally sampled lineages, while monofurcating nodes are interpreted as non-terminally sampled lineages, i.e., lineages sampled at some past time point and with subsequently sampled descendants.
root_age	Positive numeric, specifying the age of the tree's root. Can be used to define a time offset, e.g. if the last tip was not actually sampled at the present. If NULL, this will be calculated from the tree and it will be assumed that the last tip was sampled at the present.
oldest_age	Strictly positive numeric, specifying the oldest time before present ("age") to consider when calculating the likelihood. If this is equal to or greater than the root age, then oldest_age is interpreted as the stem age. If oldest_age is less than the root age, the tree is split into multiple subtrees at that age by treating every edge crossing that age as the stem of a subtree, and each subtree is considered an independent realization of the HBDS model stemming at that age. This can be useful for avoiding points in the tree close to the root, where estimation uncertainty is generally higher. If oldest_age==NULL, it is automatically set to the root age.
age_grid	Numeric vector, listing ages in ascending order, on which λ, μ, ψ and κ are fitted and allowed to vary independently. This grid must cover at least the age range from the present (age 0) to oldest_age. If NULL or of length <=1 (regardless of value), then λ, μ, ψ and κ are assumed to be time-independent.
CSA_ages	Optional numeric vector, listing ages (in ascending order) at which concentrated sampling attempts (CSAs) occurred. If NULL, it is assumed that no concentrated sampling attempts took place and that all tips were sampled according to the continuous sampling rate psi.
min_lambda	Numeric vector of length Ngrid (=max(1,length(age_grid))), or a single numeric, specifying lower bounds for the fitted speciation rate λ at each point in the age grid. If a single numeric, the same lower bound applies at all ages.
max_lambda	Numeric vector of length Ngrid, or a single numeric, specifying upper bounds for the fitted speciation rate λ at each point in the age grid. If a single numeric, the same upper bound applies at all ages. Use +Inf to omit upper bounds.
min_mu	Numeric vector of length Ngrid, or a single numeric, specifying lower bounds for the fitted extinction rate μ at each point in the age grid. If a single numeric, the same lower bound applies at all ages.
max_mu	Numeric vector of length Ngrid, or a single numeric, specifying upper bounds for the fitted extinction rate μ at each point in the age grid. If a single numeric, the same upper bound applies at all ages. Use +Inf to omit upper bounds.
min_psi	Numeric vector of length Ngrid, or a single numeric, specifying lower bounds for the fitted Poissonian sampling rate ψ at each point in the age grid. If a single numeric, the same lower bound applies at all ages.
max_psi	Numeric vector of length Ngrid, or a single numeric, specifying upper bounds for the fitted Poissonian sampling rate ψ at each point in the age grid. If a single numeric, the same upper bound applies at all ages. Use +Inf to omit upper bounds.

min_kappa	Numeric vector of length Ngrid, or a single numeric, specifying lower bounds for the fitted retention probability κ at each point in the age grid. If a single numeric, the same lower bound applies at all ages.
max_kappa	Numeric vector of length Ngrid, or a single numeric, specifying upper bounds for the fitted retention probability κ at each point in the age grid. If a single numeric, the same upper bound applies at all ages. Use +Inf to omit upper bounds.
min_CSA_probs	Numeric vector of length NCSA (=length(CSA_ages)), or a single numeric, specifying lower bounds for the fitted sampling probabilities $\rho_1,..,\rho_m$ at each concentrated sampling attempt. If a single numeric, the same lower bound applies at all CSAs. Note that, since ρ_1, ρ_2, ... are probabilities, min_CSA_probs should not be negative.
max_CSA_probs	Numeric vector of length NCSA, or a single numeric, specifying upper bounds for the fitted sampling probabilities ρ_1, ρ_2, ... at each concentrated sampling attempt. If a single numeric, the same upper bound applies at all CSAs. Note that, since ρ_1, ρ_2, ... are probabilities, max_CSA_probs should not be greater than 1.
min_CSA_kappas	Numeric vector of length NCSA, or a single numeric, specifying lower bounds for the fitted retention probabilities κ_1, κ_2, ... at each concentrated sampling attempt. If a single numeric, the same lower bound applies at all CSAs. Note that, since κ_1, κ_2, ... are probabilities, min_CSA_kappas should not be negative.
max_CSA_kappas	Numeric vector of length NCSA, or a single numeric, specifying upper bounds for the fitted sampling probabilities κ_1, κ_2, ... at each concentrated sampling attempt. If a single numeric, the same upper bound applies at all CSAs. Note that, since κ_1, κ_2, .. are probabilities, max_CSA_kappas should not be greater than 1.
guess_lambda	Initial guess for λ at each age-grid point. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same λ at all ages) or a numeric vector of size Ngrid specifying a separate guess for λ at each age-grid point. To omit an initial guess for some but not all age-grid points, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.
guess_mu	Initial guess for μ at each age-grid point. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same μ at all ages) or a numeric vector of size Ngrid specifying a separate guess for μ at each age-grid point. To omit an initial guess for some but not all age-grid points, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.
guess_psi	Initial guess for ψ at each age-grid point. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same ψ at all ages) or a numeric vector of size Ngrid specifying a separate guess for ψ at each age-grid point. To omit an initial guess for some but not all age-grid points, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.
guess_kappa	Initial guess for κ at each age-grid point. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same κ at all ages) or a numeric vector of size Ngrid specifying a separate guess for κ at each age-grid

point. To omit an initial guess for some but not all age-grid points, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.

guess_CSA_probs

Initial guess for the ρ_1, ρ_2, ... at each concentrated sampling attempt. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same value at every CSA) or a numeric vector of size NCSA specifying a separate guess at each CSA. To omit an initial guess for some but not all CSAs, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.

guess_CSA_kappas

Initial guess for the κ_1, κ_2, ... at each concentrated sampling attempt. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same value at every CSA) or a numeric vector of size NCSA specifying a separate guess at each CSA. To omit an initial guess for some but not all CSAs, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.

fixed_lambda   Optional fixed (i.e. non-fitted) λ values on one or more age-grid points. Either NULL (λ is not fixed anywhere), or a single numeric (λ fixed to the same value at all grid points) or a numeric vector of size Ngrid (if fixed_age_grid=NULL; λ fixed on one or more age-grid points, use NA for non-fixed values) or a numeric vector of the same size as fixed_age_grid (if fixed_age_grid!=NULL, in which case all entries in fixed_lambda must be finite numbers).

fixed_mu   Optional fixed (i.e. non-fitted) μ values on one or more age-grid points. Either NULL (μ is not fixed anywhere), or a single numeric (μ fixed to the same value at all grid points) or a numeric vector of size Ngrid (if fixed_age_grid=NULL; μ fixed on one or more age-grid points, use NA for non-fixed values) or a numeric vector of the same size as fixed_age_grid (if fixed_age_grid!=NULL, in which case all entries in fixed_mu must be finite numbers).

fixed_psi   Optional fixed (i.e. non-fitted) ψ values on one or more age-grid points. Either NULL (ψ is not fixed anywhere), or a single numeric (ψ fixed to the same value at all grid points) or a numeric vector of size Ngrid (if fixed_age_grid=NULL; ψ fixed on one or more age-grid points, use NA for non-fixed values) or a numeric vector of the same size as fixed_age_grid (if fixed_age_grid!=NULL, in which case all entries in fixed_psi must be finite numbers).

fixed_kappa   Optional fixed (i.e. non-fitted) κ values on one or more age-grid points. Either NULL (κ is not fixed anywhere), or a single numeric (κ fixed to the same value at all grid points) or a numeric vector of size Ngrid (if fixed_age_grid=NULL; κ fixed on one or more age-grid points, use NA for non-fixed values) or a numeric vector of the same size as fixed_age_grid (if fixed_age_grid!=NULL, in which case all entries in fixed_kappa must be finite numbers).

fixed_CSA_probs

Optional fixed (i.e. non-fitted) ρ_1, ρ_2, ... values on one or more age-grid points. Either NULL (none of the ρ_1, ρ_2,... are fixed), or a single numeric (ρ_1, ρ_2,... are fixed to the same value at all CSAs) or a numeric vector of size NCSA (one or more of the ρ_1, ρ_2, ... are fixed, use NA for non-fixed values).

fixed_CSA_kappas

Optional fixed (i.e. non-fitted) κ_1, κ_2, ... values on one or more age-grid points.

Either NULL (none of the κ_1, κ_2,... are fixed), or a single numeric (κ_1, κ_2,... are fixed to the same value at all CSAs) or a numeric vector of size NCSA (one or more of the κ_1, κ_2, ... are fixed, use NA for non-fixed values).

fixed_age_grid Optional numeric vector, specifying an age grid on which fixed_lambda, fixed_mu, fixed_psi and fixed_kappa (whichever is provided) are defined instead of on the age_grid. If fixed_age_grid is provided, then each of fixed_lambda, fixed_mu, fixed_psi and fixed_kappa must be defined (i.e. have a finite non-negative value) on every point in fixed_age_grid. Entries in fixed_age_grid must be in ascending order and must cover at least the ages 0 to oldest_age.

This option may be useful if you want to fit some parameters on a coarse grid, but want to specify (fix) some other parameters on a much finer grid. Also note that if fixed_age_grid is used, all parameters lambda, mu, psi and kappa are internally re-interpolated onto fixed_age_grid when evaluating the likelihood; hence, in general fixed_age_grid should be much finer than age_grid. In most situations you would probably want to keep the default fixed_age_grid=NULL.

const_lambda Logical, specifying whether λ should be assumed constant across the grid, i.e. time-independent. Setting const_lambda=TRUE reduces the number of free (i.e., independently fitted) parameters. If λ is fixed on some grid points (i.e. via fixed_lambda), then only the non-fixed lambdas are assumed to be identical to one another.

const_mu Logical, specifying whether μ should be assumed constant across the grid, i.e. time-independent. Setting const_mu=TRUE reduces the number of free (i.e., independently fitted) parameters. If μ is fixed on some grid points (i.e. via fixed_mu), then only the non-fixed mus are assumed to be identical to one another.

const_psi Logical, specifying whether ψ should be assumed constant across the grid, i.e. time-independent. Setting const_psi=TRUE reduces the number of free (i.e., independently fitted) parameters. If ψ is fixed on some grid points (i.e. via fixed_psi), then only the non-fixed psis are assumed to be identical to one another.

const_kappa Logical, specifying whether κ should be assumed constant across the grid, i.e. time-independent. Setting const_kappa=TRUE reduces the number of free (i.e., independently fitted) parameters. If κ is fixed on some grid points (i.e. via fixed_kappa), then only the non-fixed kappas are assumed to be identical to one another.

const_CSA_probs

Logical, specifying whether the ρ_1, ρ_2, ... should be the same across all CSAs. Setting const_CSA_probs=TRUE reduces the number of free (i.e., independently fitted) parameters. If some of the ρ_1, ρ_2, ... are fixed (i.e. via fixed_CSA_probs), then only the non-fixed CSA_probs are assumed to be identical to one another.

const_CSA_kappas

Logical, specifying whether the κ_1, κ_2, ... should be the same across all CSAs. Setting const_CSA_kappas=TRUE reduces the number of free (i.e., independently fitted) parameters. If some of the κ_1, κ_2, ... are fixed (i.e. via fixed_CSA_kappas), then only the non-fixed CSA_kappas are assumed to be identical to one another.

splines_degree Integer between 0 and 3 (inclusive), specifying the polynomial degree of λ, μ, ψ and κ between age-grid points. If 0, then λ, μ, ψ and κ are considered piecewise

constant, if 1 they are considered piecewise linear, if 2 or 3 they are considered to be splines of degree 2 or 3, respectively. The `splines_degree` influences the analytical properties of the curve, e.g. `splines_degree==1` guarantees a continuous curve, `splines_degree==2` guarantees a continuous curve and continuous derivative, and so on. A degree of 0 is generally not recommended. The case `splines_degree=0` is also known as "skyline" model.

condition   Character, either "crown", "stem", "none" or "auto", specifying on what to condition the likelihood. If "crown", the likelihood is conditioned on the survival of the two daughter lineages branching off at the root. If "stem", the likelihood is conditioned on the survival of the stem lineage. Note that "crown" really only makes sense when `oldest_age` is equal to the root age, while "stem" is recommended if `oldest_age` differs from the root age. "none" is generally not recommended. If "auto", the condition is chosen according to the above recommendations.

ODE_relative_dt
                      Positive unitless number, specifying the default relative time step for the ordinary differential equation solvers. Typical values are 0.01-0.001.

ODE_relative_dy
                      Positive unitless number, specifying the relative difference between subsequent simulated and interpolated values, in internally used ODE solvers. Typical values are 1e-2 to 1e-5. A smaller `ODE_relative_dy` increases interpolation accuracy, but also increases memory requirements and adds runtime (scaling with the tree's age span, not with Ntips).

CSA_age_epsilon
                      Non-negative numeric, in units of time, specfying the age radius around a concentrated sampling attempt, within which to assume that sampling events were due to that concentrated sampling attempt. If NULL, this is chosen automatically based on the anticipated scale of numerical rounding errors. Only relevant if concentrated sampling attempts are included.

Ntrials   Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing `Ntrials` reduces the risk of reaching a non-global local maximum in the fitting objective.

max_start_attempts
                      Integer, specifying the number of times to attempt finding a valid start point (per trial) before giving up on that trial. Randomly choosen extreme start parameters may occasionally result in Inf/undefined likelihoods, so this option allows the algorithm to keep looking for valid starting points.

Nthreads   Integer, specifying the number of parallel threads to use for performing multiple fitting trials simultaneously. This should generally not exceed the number of available CPUs on your machine. Parallel computing is not available on the Windows platform.

max_model_runtime
                      Optional numeric, specifying the maximum number of seconds to allow for each evaluation of the likelihood function. Use this to abort fitting trials leading to parameter regions where the likelihood takes a long time to evaluate (these are often unlikely parameter regions).

Nbootstraps   Integer, specifying the number of parametric bootstraps to perform for estimating standard errors and confidence intervals of estimated parameters. Set to 0 for no bootstrapping.

Ntrials_per_bootstrap

    Integer, specifying the number of fitting trials to perform for each bootstrap sampling. If `NULL`, this is set equal to `max(1,Ntrials)`. Decreasing `Ntrials_per_bootstrap` will reduce computation time, at the expense of potentially inflating the estimated confidence intervals; in some cases (e.g., for very large trees) this may be useful if fitting takes a long time and confidence intervals are very narrow anyway. Only relevant if `Nbootstraps>0`.

fit_control   Named list containing options for the `nlminb` optimization routine, such as `iter.max`, `eval.max` or `rel.tol`. For a complete list of options and default values see the documentation of `nlminb` in the `stats` package.

focal_param_values

    Optional list, listing combinations of parameter values of particular interest and for which the log-likelihoods should be returned. Every element of this list should itself be a named list, containing the elements `lambda`, `mu`, `psi` and `kappa` (each being a numeric vector of size NG) as well as the elements `CSA_probs` and `CSA_kappas` (each being a numeric vector of size NCSA). This may be used e.g. for diagnostic purposes, e.g. to examine the shape of the likelihood function.

verbose   Logical, specifying whether to print progress reports and warnings to the screen. Note that errors always cause a return of the function (see return values `success` and `error`).

diagnostics   Logical, specifying whether to print detailed information (such as model likelihoods) at every iteration of the fitting routine. For debugging purposes mainly.

verbose_prefix   Character, specifying the line prefix for printing progress reports to the screen.

Details

Warning: In the absence of concentrated sampling attempts (NCSA=0), and without well-justified a priori constraints on either λ, μ, ψ and/or κ, it is generally impossible to reliably estimate λ, μ, ψ and κ from timetrees alone. This routine (and any other software that claims to estimate λ, μ, ψ and κ solely from timetrees) should thus be treated with great suspicion. Many epidemiological models make the (often reasonable assumption) that $\kappa = 0$; note that even in this case, one generally can't co-estimate λ, μ and ψ from the timetree alone.

It is advised to provide as much information to the function `fit_hbds_model_on_grid` as possible, including reasonable lower and upper bounds (`min_lambda`, `max_lambda`, `min_mu`, `max_mu`, `min_psi`, `max_psi`, `min_kappa`, `max_kappa`) and reasonable parameter guesses. It is also important that the `age_grid` is sufficiently fine to capture the expected major variations of λ, μ, ψ and κ over time, but keep in mind the serious risk of overfitting when `age_grid` is too fine and/or the tree is too small. The `age_grid` does not need to be uniform, i.e., you may want to use a finer grid in regions where there's more data (tips) available. If strong lower and upper bounds are not available and fitting takes a long time to run, consider using the option `max_model_runtime` to limit how much time the fitting allows for each evaluation of the likelihood.

Note that here "age" refers to time before present, i.e., age increases from tips to root and age 0 is present-day. CSAs are enumerated in the order of increasing age, i.e., from the present to the past. Similarly, the age grid specifies time points from the present towards the past.

Value

A list with the following elements:

success	Logical, indicating whether model fitting succeeded. If FALSE, the returned list will include an additional "error" element (character) providing a description of the error; in that case all other return variables may be undefined.
objective_value	The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.
objective_name	The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".
loglikelihood	The log-likelihood of the fitted model for the given timetree.
guess_loglikelihood	The log-likelihood of the guessed model for the given timetree.
param_fitted	Named list, specifying the fixed and fitted model parameters. This list will contain the elements lambda, mu, psi and kappa (each being a numeric vector of size NG, listing λ, μ, ψ and κ at each age-grid point) as well as the elements CSA_probs and CSA_kappas (each being a numeric vector of size NCSA).
param_guess	Named list, specifying the guessed model parameters. This list will contain the elements lambda, mu, psi and kappa (each being a numeric vector of size NG) as well as the elements CSA_probs and CSA_kappas (each being a numeric vector of size NCSA). Between grid points λ should be interpreted as a piecewise polynomial function (natural spline) of degree splines_degree; to evaluate this function at arbitrary ages use the castor routine [evaluate_spline](#). The same also applies to μ, ψ and κ.
age_grid	Numeric vector of size NG, the age-grid on which λ, μ, ψ and κ are defined. This will be the same as the provided age_grid, unless the latter was NULL or of length <=1.
CSA_ages	Numeric vector of size NCSA, ting listhe ages at which concentrated sampling attempts occurred. This is the same as provided to the function.
NFP	Integer, number of free (i.e., independently) fitted parameters. If none of the λ, μ and ρ were fixed, and const_lambda=FALSE and const_mu=FALSE, then NFP will be equal to 2*Ngrid+1.
Ndata	Integer, the number of data points (sampling and branching events) used for fitting.
AIC	The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.
BIC	The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of branching times), and L is the maximized likelihood.
condition	Character, specifying what conditioning was root for the likelihood (e.g. "crown" or "stem").

converged	Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase `iter.max` and/or `eval.max`).
Niterations	Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.
Nevaluations	Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.
standard_errors	Named list specifying the standard errors of the parameters, based on parametric bootstrapping. This list will contain the elements `lambda`, `mu`, `psi` and `kappa` (each being a numeric vector of size NG) as well as the elements `CSA_probs` and `CSA_kappas` (each being a numeric vector of size NCSA). Only included if `Nbootstraps>0`. Note that the standard errors of non-fitted (i.e., fixed) parameters will be zero.
CI50lower	Named list specifying the lower end of the 50% confidence interval (i.e. the 25% quantile) for each parameter, based on parametric bootstrapping. This list will contain the elements `lambda`, `mu`, `psi` and `kappa` (each being a numeric vector of size NG) as well as the elements `CSA_probs` and `CSA_kappas` (each being a numeric vector of size NCSA). Only included if `Nbootstraps>0`.
CI50upper	Similar to `CI50lower`, but listing the upper end of the 50% confidence interval (i.e. the 75% quantile) for each parameter. For example, the confidence interval for λ at age `age_grid[1]` will be between `CI50lower$lambda[1]` and `CI50upper$lambda[1]`. Only included if `Nbootstraps>0`.
CI95lower	Similar to `CI50lower`, but listing the lower end of the 95% confidence interval (i.e. the 2.5% quantile) for each parameter. Only included if `Nbootstraps>0`.
CI95upper	Similar to `CI50upper`, but listing the upper end of the 95% confidence interval (i.e. the 97.5% quantile) for each parameter. Only included if `Nbootstraps>0`.
consistency	Numeric between 0 and 1, estimated consistency of the data with the fitted model. If L denotes the loglikelihood of new data generated by the fitted model (under the same model) and M denotes the expectation of L, then `consistency` is the probability that $|L - M|$ will be greater or equal to $|X - M|$, where X is the loglikelihood of the original data under the fitted model. Only returned if `Nbootstraps>0`. A low consistency (e.g., <0.05) indicates that the fitted model is a poor description of the data. See Lindholm et al. (2019) for background.

Author(s)

Stilianos Louca

References

T. Stadler, D. Kuehnert, S. Bonhoeffer, A. J. Drummond (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). PNAS. 110:228-233.

A. Lindholm, D. Zachariah, P. Stoica, T. B. Schoen (2019). Data consistency approach to model validation. IEEE Access. 7:59788-59796.

See Also

simulate_deterministic_hbds, fit_hbds_model_parametric

Examples

```
## Not run:
# define lambda & mu & psi as functions of time
# Assuming an exponentially varying lambda & mu, and a constant psi
time2lambda = function(times){ 2*exp(0.1*times) }
time2mu     = function(times){ 0.1*exp(0.09*times) }
time2psi    = function(times){ rep(0.2, times=length(times)) }

# define concentrated sampling attempts
CSA_times   = c(3,4)
CSA_probs   = c(0.1, 0.2)

# generate random tree based on lambda, mu & psi
# assume that all sampled lineages are removed from the pool (i.e. kappa=0)
time_grid = seq(from=0, to=100, by=0.01)
simul = generate_tree_hbds( max_time    = 5,
                            time_grid   = time_grid,
                            lambda      = time2lambda(time_grid),
                            mu          = time2mu(time_grid),
                            psi         = time2psi(time_grid),
                            kappa       = 0,
                            CSA_times   = CSA_times,
                            CSA_probs   = CSA_probs,
                            CSA_kappas  = 0)
tree     = simul$tree
root_age = simul$root_age
cat(sprintf("Tree has %d tips\n",length(tree$tip.label)))

# Define an age grid on which lambda_function & mu_function shall be fitted
fit_age_grid = seq(from=0,to=root_age,length.out=3)

# Fit an HBDS model on a grid
# Assume that psi is known and that sampled lineages are removed from the pool
# Hence, we only fit lambda & mu & CSA_probs
cat(sprintf("Fitting model to tree..\n"))
fit = fit_hbds_model_on_grid(tree,
                             root_age          = root_age,
                             age_grid          = fit_age_grid,
                             CSA_ages          = rev(simul$final_time - CSA_times),
                          fixed_psi          = time2psi(simul$final_time-fit_age_grid),
                             fixed_kappa       = 0,
                             fixed_CSA_kappas  = 0,
                             Ntrials           = 4,
                             Nthreads          = 4,
                             Nbootstraps       = 0,
                             verbose           = TRUE,
                             verbose_prefix    = "  ")
if(!fit$success){
```

```
        cat(sprintf("ERROR: Fitting failed: %s\n",fit$error))
    }else{
        # compare fitted lambda to true lambda
        plot(x=fit$age_grid,
             y=fit$param_fitted$lambda,
             type='l',
             col='#000000',
             xlim=c(root_age,0),
             xlab='age',
             ylab='lambda')
        lines(x=simul$final_time-time_grid,
               y=time2lambda(time_grid),
               type='l',
               col='#0000AA')
    }


    # compare true and fitted model in terms of their LTTs
    LTT      = castor::count_lineages_through_time(tree, Ntimes=100, include_slopes=TRUE)
    LTT$ages = root_age - LTT$times

    cat(sprintf("Simulating deterministic HBDS (true model)..\n"))
    age0 = 0.5 # reference age at which to equate LTTs
    LTT0 = approx(x=LTT$ages, y=LTT$lineages, xout=age0)$y # tree LTT at age0
    fsim = simulate_deterministic_hbds( age_grid       = fit$age_grid,
                                        lambda         = fit$param_fitted$lambda,
                                        mu             = fit$param_fitted$mu,
                                        psi            = fit$param_fitted$psi,
                                        kappa          = fit$param_fitted$kappa,
                                        CSA_ages       = fit$CSA_ages,
                                        CSA_probs      = fit$param_fitted$CSA_probs,
                                        CSA_kappas     = fit$param_fitted$CSA_kappas,
                                        requested_ages = seq(0,root_age,length.out=200),
                                        age0           = age0,
                                        LTT0           = LTT0,
                                        splines_degree = 1)
    if(!fsim$success){
        cat(sprintf("ERROR: Could not simulate fitted model: %s\n",fsim$error))
        stop()
    }
    plot(x=LTT$ages, y=LTT$lineages, type='l', col='#0000AA', lwd=2, xlim=c(root_age,0))
    lines(x=fsim$ages, y=fsim$LTT, type='l', col='#000000', lwd=2)

    ## End(Not run)
```

---

fit_hbds_model_parametric

Fit a parametric homogenous birth-death-sampling model to a time-tree.

Description

Given a timetree (potentially sampled through time and not necessarily ultrametric), fit a homogenous birth-death-sampling (HBDS) model in which speciation, extinction and lineage sampling occurs at some continuous (Poissonian) rates λ, μ and ψ, which are given as parameterized functions of time before present. Sampled lineages are kept in the pool of extant lineages at some "retention probability" κ, which may also depend on time. In addition, this model can include concentrated sampling attempts (CSAs) at a finite set of discrete time points $t_1, .., t_m$. "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction/sampling rates. Every HBDS model is thus defined based on the values that λ, μ, ψ and κ take over time, as well as the sampling probabilities $\rho_1, .., \rho_m$ and retention probabilities $\kappa_1, .., \kappa_m$ during the concentrated sampling attempts; each of these parameters, in turn, is assumed to be determined by a finite set of parameters. This function estimates these parameters by maximizing the corresponding likelihood of the timetree. Special cases of this model are sometimes known as "birth-death-skyline plots" in the literature (Stadler 2013). In epidemiology, these models are often used to describe the phylogenies of viral strains sampled over the course of the epidemic.

Usage

```
fit_hbds_model_parametric(tree,
                          param_values,
                          param_guess          = NULL,
                          param_min            = -Inf,
                          param_max            = +Inf,
                          param_scale          = NULL,
                          root_age             = NULL,
                          oldest_age           = NULL,
                          lambda               = 0,
                          mu                   = 0,
                          psi                  = 0,
                          kappa                = 0,
                          age_grid             = NULL,
                          CSA_ages             = NULL,
                          CSA_probs            = NULL,
                          CSA_kappas           = 0,
                          condition            = "auto",
                          ODE_relative_dt      = 0.001,
                          ODE_relative_dy      = 1e-3,
                          CSA_age_epsilon      = NULL,
                          Ntrials              = 1,
                          max_start_attempts   = 1,
                          Nthreads             = 1,
                          max_model_runtime    = NULL,
                          Nbootstraps          = 0,
                          Ntrials_per_bootstrap = NULL,
                          fit_control          = list(),
                          focal_param_values   = NULL,
                          verbose              = FALSE,
                          diagnostics          = FALSE,
```

```
verbose_prefix       = "")
```

Arguments

tree	A timetree of class "phylo", representing the time-calibrated reconstructed phylogeny of a set of extant and/or extinct species. Tips of the tree are interpreted as terminally sampled lineages, while monofurcating nodes are interpreted as non-terminally sampled lineages, i.e., lineages sampled at some past time point and with subsequently sampled descendants.
param_values	Numeric vector, specifying fixed values for a some or all model parameters. For fitted (i.e., non-fixed) parameters, use NaN or NA. For example, the vector c(1.5,NA,40) specifies that the 1st and 3rd model parameters are fixed at the values 1.5 and 40, respectively, while the 2nd parameter is to be fitted. The length of this vector defines the total number of model parameters. If entries in this vector are named, the names are taken as parameter names. Names should be included if the functions lambda, mu, psi, kappa, CSA_psi and CSA_kappa query parameter values by name (as opposed to numeric index).
param_guess	Numeric vector of size NP, specifying a first guess for the value of each model parameter. For fixed parameters, guess values are ignored. Can be NULL only if all model parameters are fixed.
param_min	Optional numeric vector of size NP, specifying lower bounds for model parameters. If of size 1, the same lower bound is applied to all parameters. Use -Inf to omit a lower bound for a parameter. If NULL, no lower bounds are applied. For fixed parameters, lower bounds are ignored.
param_max	Optional numeric vector of size NP, specifying upper bounds for model parameters. If of size 1, the same upper bound is applied to all parameters. Use +Inf to omit an upper bound for a parameter. If NULL, no upper bounds are applied. For fixed parameters, upper bounds are ignored.
param_scale	Optional numeric vector of size NP, specifying typical scales for model parameters. If of size 1, the same scale is assumed for all parameters. If NULL, scales are determined automatically. For fixed parameters, scales are ignored. It is strongly advised to provide reasonable scales, as this facilitates the numeric optimization algorithm.
root_age	Positive numeric, specifying the age of the tree's root. Can be used to define a time offset, e.g. if the last tip was not actually sampled at the present. If NULL, this will be calculated from the tree and it will be assumed that the last tip was sampled at the present.
oldest_age	Strictly positive numeric, specifying the oldest time before present ("age") to consider when calculating the likelihood. If this is equal to or greater than the root age, then oldest_age is interpreted as the stem age. If oldest_age is less than the root age, the tree is split into multiple subtrees at that age by treating every edge crossing that age as the stem of a subtree, and each subtree is considered an independent realization of the HBDS model stemming at that age. This can be useful for avoiding points in the tree close to the root, where estimation uncertainty is generally higher. If oldest_age==NULL, it is automatically set to the root age.

lambda	Function specifying the speciation rate at any given age (time before present) and for any given parameter values. This function must take exactly two arguments, the 1st one being a numeric vector (one or more ages) and the 2nd one being a numeric vector of size NP (parameter values), and return a numeric vector of the same size as the 1st argument with strictly positive entries. Can also be a single numeric (i.e., lambda is fixed).
mu	Function specifying the extinction rate at any given age and for any given parameter values. This function must take exactly two arguments, the 1st one being a numeric vector (one or more ages) and the 2nd one being a numeric vector of size NP (parameter values), and return a numeric vector of the same size as the 1st argument with non-negative entries. Can also be a single numeric (i.e., mu is fixed).
psi	Function specifying the continuous (Poissonian) lineage sampling rate at any given age and for any given parameter values. This function must take exactly two arguments, the 1st one being a numeric vector (one or more ages) and the 2nd one being a numeric vector of size NP (parameter values), and return a numeric vector of the same size as the 1st argument with non-negative entries. Can also be a single numeric (i.e., psi is fixed).
kappa	Function specifying the retention probability for continuously sampled lineages, at any given age and for any given parameter values. This function must take exactly two arguments, the 1st one being a numeric vector (one or more ages) and the 2nd one being a numeric vector of size NP (parameter values), and return a numeric vector of the same size as the 1st argument with non-negative entries. The retention probability is the probability of a sampled lineage remaining in the pool of extant lineages. Can also be a single numeric (i.e., kappa is fixed).
age_grid	Numeric vector, specifying ages at which the lambda, mu, psi and kappa functionals should be evaluated. This age grid must be fine enough to capture the possible variation in λ, μ, ψ and κ over time, within the permissible parameter range. Listed ages must be strictly increasing, and must cover at least the full considered age interval (from 0 to oldest_age). Can also be NULL or a vector of size 1, in which case λ, μ, ψ and κ are assumed to be time-independent.
CSA_ages	Optional numeric vector, listing ages (in ascending order) at which concentrated sampling attempts occurred. If NULL, it is assumed that no concentrated sampling attempts took place and that all tips were sampled according to the continuous sampling rate psi.
CSA_probs	Function specifying the sampling probabilities during the various concentrated sampling attempts, depending on parameter values. Hence, for any choice of parameters, CSA_probs must return a numeric vector of the same size as CSA_ages. Can also be a single numeric (i.e., concentrated sampling probability is fixed).
CSA_kappas	Function specifying the retention probabilities during the various concentrated sampling attempts, depending on parameter values. Hence, for any choice of parameters, CSA_kappas must return a numeric vector of the same size as CSA_ages. Can also be a single numeric (i.e., retention probability during concentrated samplings is fixed).
condition	Character, either "crown", "stem", "none" or "auto", specifying on what to condition the likelihood. If "crown", the likelihood is conditioned on the survival

of the two daughter lineages branching off at the root. If "stem", the likelihood is conditioned on the survival of the stem lineage. Note that "crown" really only makes sense when `oldest_age` is equal to the root age, while "stem" is recommended if `oldest_age` differs from the root age. "none" is usually not recommended. If "auto", the condition is chosen according to the above recommendations.

ODE_relative_dt

Positive unitless number, specifying the default relative time step for the ordinary differential equation solvers. Typical values are 0.01-0.001.

ODE_relative_dy

Positive unitless number, specifying the relative difference between subsequent simulated and interpolated values, in internally used ODE solvers. Typical values are 1e-2 to 1e-5. A smaller `ODE_relative_dy` increases interpolation accuracy, but also increases memory requirements and adds runtime (scaling with the tree's age span, not with Ntips).

CSA_age_epsilon

Non-negative numeric, in units of time, specfying the age radius around a concentrated sampling attempt, within which to assume that sampling events were due to that concentrated sampling attempt. If NULL, this is chosen automatically based on the anticipated scale of numerical rounding errors. Only relevant if concentrated sampling attempts are included.

Ntrials   Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing `Ntrials` reduces the risk of reaching a non-global local maximum in the fitting objective.

max_start_attempts

Integer, specifying the number of times to attempt finding a valid start point (per trial) before giving up on that trial. Randomly chosen extreme start parameters may occasionally result in Inf/undefined likelihoods, so this option allows the algorithm to keep looking for valid starting points.

Nthreads   Integer, specifying the number of parallel threads to use for performing multiple fitting trials simultaneously. This should generally not exceed the number of available CPUs on your machine. Parallel computing is not available on the Windows platform.

max_model_runtime

Optional numeric, specifying the maximum number of seconds to allow for each evaluation of the likelihood function. Use this to abort fitting trials leading to parameter regions where the likelihood takes a long time to evaluate (these are often unlikely parameter regions).

Nbootstraps   Integer, specifying the number of parametric bootstraps to perform for estimating standard errors and confidence intervals of estimated model parameters. Set to 0 for no bootstrapping.

Ntrials_per_bootstrap

Integer, specifying the number of fitting trials to perform for each bootstrap sampling. If NULL, this is set equal to max(1,Ntrials). Decreasing Ntrials_per_bootstrap will reduce computation time, at the expense of potentially inflating the estimated confidence intervals; in some cases (e.g., for very large trees) this may

	be useful if fitting takes a long time and confidence intervals are very narrow anyway. Only relevant if Nbootstraps>0.
fit_control	Named list containing options for the `nlminb` optimization routine, such as `iter.max`, `eval.max` or `rel.tol`. For a complete list of options and default values see the documentation of `nlminb` in the `stats` package.

focal_param_values

Optional numeric matrix having NP columns and an arbitrary number of rows, listing combinations of parameter values of particular interest and for which the log-likelihoods should be returned. This may be used for diagnostic purposes, e.g., to examine the shape of the likelihood function.

verbose	Logical, specifying whether to print progress reports and warnings to the screen. Note that errors always cause a return of the function (see return values `success` and `error`).
diagnostics	Logical, specifying whether to print detailed information (such as model likelihoods) at every iteration of the fitting routine. For debugging purposes mainly.
verbose_prefix	Character, specifying the line prefix for printing progress reports to the screen.

Details

This function is designed to estimate a finite set of scalar parameters $(p_1, .., p_n \in \mathsf{R})$ that determine the speciation rate λ, the extinction rate μ, the sampling rate ψ, the retention rate κ, the concentrated sampling probabilities $\rho_1, .., \rho_m$ and the concentrated retention probabilities $\kappa_1, .., \kappa_m$, by maximizing the likelihood of observing a given timetree under the HBDS model. Note that the ages (times before present) of the concentrated sampling attempts are assumed to be known and are not fitted.

It is generally advised to provide as much information to the function `fit_hbds_model_parametric` as possible, including reasonable lower and upper bounds (`param_min` and `param_max`), a reasonable parameter guess (`param_guess`) and reasonable parameter scales `param_scale`. If some model parameters can vary over multiple orders of magnitude, it is advised to transform them so that they vary across fewer orders of magnitude (e.g., via log-transformation). It is also important that the `age_grid` is sufficiently fine to capture the variation of λ, μ, ψ and κ over time, since the likelihood is calculated under the assumption that these functions vary linearly between grid points.

Note that in this function age always refers to time before present, i.e., present day age is 0 and age increases from tips to root. The functions `lambda`, `mu`, `psi` and `kappa` should be functions of age, not forward time. Similarly, concentrated sampling attempts (CSAs) are enumerated in order of increasing age, i.e., starting with the youngest CSA and moving towards older CSAs.

Value

A list with the following elements:

success	Logical, indicating whether model fitting succeeded. If `FALSE`, the returned list will include an additional "`error`" element (character) providing a description of the error; in that case all other return variables may be undefined.

objective_value

The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.

objective_name The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".

loglikelihood The log-likelihood of the fitted model for the given timetree.

param_fitted Numeric vector of size NP (number of model parameters), listing all fitted or fixed model parameters in their standard order (see details above). If param_names was provided, elements in fitted_params will be named.

param_guess Numeric vector of size NP, listing guessed or fixed values for all model parameters in their standard order. If param_names was provided, elements in param_guess will be named.

guess_loglikelihood

The loglikelihood of the data for the initial parameter guess (param_guess).

focal_loglikelihoods

A numeric vector of the same size as nrow(focal_param_values), listing loglikelihoods for each of the focal parameter conbinations listed in focal_loglikelihoods.

NFP Integer, number of fitted (i.e., non-fixed) model parameters.

Ndata Number of data points used for fitting, i.e., the number of sampling and branching events that occurred between ages 0 and oldest_age.

AIC The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.

BIC The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (Ndata), and L is the maximized likelihood.

condition Character, specifying what conditioning was root for the likelihood (e.g. "crown" or "stem").

converged Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase iter.max and/or eval.max).

Niterations Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.

Nevaluations Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.

trial_start_objectives

Numeric vector of size Ntrials, listing the initial objective values (e.g., loglikelihoods) for each fitting trial, i.e. at the start parameter values.

trial_objective_values

Numeric vector of size Ntrials, listing the final maximized objective values (e.g., loglikelihoods) for each fitting trial.

trial_Nstart_attempts

Integer vector of size Ntrials, listing the number of start attempts for each fitting trial, until a starting point with valid likelihood was found.

trial_Niterations

Integer vector of size Ntrials, listing the number of iterations needed for each fitting trial.

trial_Nevaluations

> Integer vector of size `Ntrials`, listing the number of likelihood evaluations needed for each fitting trial.

standard_errors

> Numeric vector of size NP, estimated standard error of the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

medians

> Numeric vector of size NP, median the estimated parameters across parametric bootstraps. Only returned if `Nbootstraps>0`.

CI50lower

> Numeric vector of size NP, lower bound of the 50% confidence interval (25-75% percentile) for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI50upper

> Numeric vector of size NP, upper bound of the 50% confidence interval for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI95lower

> Numeric vector of size NP, lower bound of the 95% confidence interval (2.5-97.5% percentile) for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI95upper

> Numeric vector of size NP, upper bound of the 95% confidence interval for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

consistency

> Numeric between 0 and 1, estimated consistency of the data with the fitted model (Lindholm et al. 2019). See the documentation of `fit_hbds_model_on_grid` for an explanation.

Author(s)

Stilianos Louca

References

T. Stadler, D. Kuehnert, S. Bonhoeffer, A. J. Drummond (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). PNAS. 110:228-233.

A. Lindholm, D. Zachariah, P. Stoica, T. B. Schoen (2019). Data consistency approach to model validation. IEEE Access. 7:59788-59796.

See Also

`generate_tree_hbds`, `simulate_deterministic_hbds`

Examples

```
## Not run:
# Generate a random tree with exponentially varying lambda & mu and constant psi
# assume that all sampled lineages are removed from the pool (i.e. kappa=0)
time_grid = seq(from=0, to=100, by=0.01)
root_age  = 5
tree = generate_tree_hbds(max_time  = root_age,
                          time_grid = time_grid,
                          lambda    = 2*exp(0.1*time_grid),
                          mu        = 0.1*exp(0.09*time_grid),
```

```
                                psi           = 0.1,
                                kappa         = 0)$tree
    cat(sprintf("Tree has %d tips\n",length(tree$tip.label)))


    # Define a parametric HBDS model, with exponentially varying lambda & mu
    # Assume that the sampling rate is constant but unknown
    # The model thus has 5 parameters: lambda0, mu0, alpha, beta, psi
    lambda_function = function(ages,params){
        return(params['lambda0']*exp(-params['alpha']*ages));
    }
    mu_function = function(ages,params){
        return(params['mu0']*exp(-params['beta']*ages));
    }
    psi_function = function(ages,params){
        return(rep(params['psi'],length(ages)))
    }

    # Define an age grid on which lambda_function & mu_function shall be evaluated
    # Should be sufficiently fine to capture the variation in lambda & mu
    age_grid = seq(from=0,to=root_age,by=0.01)

    # Perform fitting
    cat(sprintf("Fitting model to tree..\n"))
    fit = fit_hbds_model_parametric(tree,
                          root_age      = root_age,
                          param_values  = c(lambda0=NA, mu0=NA, alpha=NA, beta=NA, psi=NA),
                          param_guess   = c(1,1,0,0,0.5),
                          param_min     = c(0,0,-1,-1,0),
                          param_max     = c(10,10,1,1,10),
                          param_scale   = 1, # all params are in the order of 1
                          lambda        = lambda_function,
                          mu            = mu_function,
                          psi           = psi_function,
                          kappa         = 0,
                          age_grid      = age_grid,
                          Ntrials       = 4, # perform 4 fitting trials
                          Nthreads      = 2) # use 2 CPUs
    if(!fit$success){
        cat(sprintf("ERROR: Fitting failed: %s\n",fit$error))
    }else{
        cat(sprintf("Fitting succeeded:\nLoglikelihood=%g\n",fit$loglikelihood))
        # print fitted parameters
        print(fit$param_fitted)
    }

    ## End(Not run)
```

---

fit_hbd_model_on_grid   *Fit a homogenous birth-death model on a discrete time grid.*

Description

Given an ultrametric timetree, fit a homogenous birth-death (HBD) model in which speciation and extinction rates (λ and mu) are defined on a fixed grid of discrete time points and assumed to vary polynomially between grid points. "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction rates (in the literature this is sometimes referred to simply as "birth-death model"). Every HBD model is defined based on the values that λ and μ take over time as well as the sampling fraction ρ (fraction of extant species sampled). This function estimates the values of λ and μ at each grid point by maximizing the likelihood (Morlon et al. 2011) of the timetree under the resulting HBD model.

Usage

```
fit_hbd_model_on_grid(tree,
                      oldest_age        = NULL,
                      age0              = 0,
                      age_grid          = NULL,
                      min_lambda        = 0,
                      max_lambda        = +Inf,
                      min_mu            = 0,
                      max_mu            = +Inf,
                      min_rho0          = 1e-10,
                      max_rho0          = 1,
                      guess_lambda      = NULL,
                      guess_mu          = NULL,
                      guess_rho0        = 1,
                      fixed_lambda      = NULL,
                      fixed_mu          = NULL,
                      fixed_rho0        = NULL,
                      const_lambda      = FALSE,
                      const_mu          = FALSE,
                      splines_degree    = 1,
                      condition         = "auto",
                      relative_dt       = 1e-3,
                      Ntrials           = 1,
                      Nthreads          = 1,
                      max_model_runtime = NULL,
                      fit_control       = list())
```

Arguments

tree   An ultrametric timetree of class "phylo", representing the time-calibrated recon-
                structed phylogeny of a set of extant species.

oldest_age   Strictly positive numeric, specifying the oldest time before present ("age") to
                consider when calculating the likelihood. If this is equal to or greater than the
                root age, then oldest_age is taken as the stem age, and the classical formula by
                Morlon et al. (2011) is used. If oldest_age is less than the root age, the tree is
                split into multiple subtrees at that age by treating every edge crossing that age as
                the stem of a subtree, and each subtree is considered an independent realization

	of the HBD model stemming at that age. This can be useful for avoiding points in the tree close to the root, where estimation uncertainty is generally higher. If `oldest_age==NULL`, it is automatically set to the root age.
age0	Non-negative numeric, specifying the youngest age (time before present) to consider for fitting, and with respect to which rho is defined. If `age0>0`, then rho0 refers to the sampling fraction at age `age0`, i.e. the fraction of lineages extant at `age0` that are included in the tree. See below for more details.
age_grid	Numeric vector, listing ages in ascending order, on which λ and μ are allowed to vary independently. This grid must cover at least the age range from `age0` to `oldest_age`. If NULL or of length <=1 (regardless of value), then λ and μ are assumed to be time-independent.
min_lambda	Numeric vector of length Ngrid (=max(1,length(age_grid))), or a single numeric, specifying lower bounds for the fitted λ at each point in the age grid. If a single numeric, the same lower bound applies at all ages.
max_lambda	Numeric vector of length Ngrid, or a single numeric, specifying upper bounds for the fitted λ at each point in the age grid. If a single numeric, the same upper bound applies at all ages. Use +Inf to omit upper bounds.
min_mu	Numeric vector of length Ngrid, or a single numeric, specifying lower bounds for the fitted μ at each point in the age grid. If a single numeric, the same lower bound applies at all ages.
max_mu	Numeric vector of length Ngrid, or a single numeric, specifying upper bounds for the fitted μ at each point in the age grid. If a single numeric, the same upper bound applies at all ages. Use +Inf to omit upper bounds.
min_rho0	Numeric, specifying a lower bound for the fitted sampling fraction ρ (fraction of extant species included in the tree).
max_rho0	Numeric, specifying an upper bound for the fitted sampling fraction ρ.
guess_lambda	Initial guess for λ at each age-grid point. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same λ at all ages) or a numeric vector of size Ngrid specifying a separate guess for λ at each age-grid point. To omit an initial guess for some but not all age-grid points, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.
guess_mu	Initial guess for μ at each age-grid point. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same μ at all ages) or a numeric vector of size Ngrid specifying a separate guess for μ at each age-grid point. To omit an initial guess for some but not all age-grid points, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.
guess_rho0	Numeric, specifying an initial guess for the sampling fraction ρ at age0. Setting this to NULL or NA is the same as setting it to 1.
fixed_lambda	Optional fixed (i.e. non-fitted) λ values on one or more age-grid points. Either NULL (λ is not fixed anywhere), or a single numeric (λ fixed to the same value at all grid points) or a numeric vector of size Ngrid (λ fixed on one or more age-grid points, use NA for non-fixed values).
fixed_mu	Optional fixed (i.e. non-fitted) μ values on one or more age-grid points. Either NULL (μ is not fixed anywhere), or a single numeric (μ fixed to the same value

	at all grid points) or a numeric vector of size Ngrid (μ fixed on one or more age-grid points, use NA for non-fixed values).
fixed_rho0	Numeric between 0 and 1, optionallly specifying a fixed value for the sampling fraction ρ. If NULL or NA, the sampling fraction ρ is estimated, however note that this may not always be meaningful (Stadler 2009, Stadler 2013).
const_lambda	Logical, specifying whether λ should be assumed constant across the grid, i.e. time-independent. Setting const_lambda=TRUE reduces the number of free (i.e., independently fitted) parameters. If λ is fixed on some grid points (i.e. via fixed_lambda), then only the non-fixed lambdas are assumed to be identical to one another.
const_mu	Logical, specifying whether μ should be assumed constant across the grid, i.e. time-independent. Setting const_mu=TRUE reduces the number of free (i.e., independently fitted) parameters. If μ is fixed on some grid points (i.e. via fixed_mu), then only the non-fixed mus are assumed to be identical to one another.
splines_degree	Integer between 0 and 3 (inclusive), specifying the polynomial degree of λ and μ between age-grid points. If 0, then λ and μ are considered piecewise constant, if 1 then λ and μ are considered piecewise linear, if 2 or 3 then λ and μ are considered to be splines of degree 2 or 3, respectively. The splines_degree influences the analytical properties of the curve, e.g. splines_degree==1 guarantees a continuous curve, splines_degree==2 guarantees a continuous curve and continuous derivative, and so on. A degree of 0 is generally not recommended. The case splines_degree=0 is also known as "skyline" model.
condition	Character, either "crown", "stem", "none" or "auto", specifying on what to condition the likelihood. If "crown", the likelihood is conditioned on the survival of the two daughter lineages branching off at the root. If "stem", the likelihood is conditioned on the survival of the stem lineage. Note that "crown" really only makes sense when oldest_age is equal to the root age, while "stem" is recommended if oldest_age differs from the root age. "none" is generally not recommended. If "auto", the condition is chosen according to the recommendations mentioned earlier.
relative_dt	Strictly positive numeric (unitless), specifying the maximum relative time step allowed for integration over time, when calculating the likelihood. Smaller values increase integration accuracy but increase computation time. Typical values are 0.0001-0.001. The default is usually sufficient.
Ntrials	Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing Ntrials reduces the risk of reaching a non-global local maximum in the fitting objective.
Nthreads	Integer, specifying the number of parallel threads to use for performing multiple fitting trials simultaneously. This should generally not exceed the number of available CPUs on your machine. Parallel computing is not available on the Windows platform.
max_model_runtime	
	Optional numeric, specifying the maximum number of seconds to allow for each evaluation of the likelihood function. Use this to abort fitting trials leading to parameter regions where the likelihood takes a long time to evaluate (these are often unlikely parameter regions).

fit_control   Named list containing options for the `nlminb` optimization routine, such as `iter.max`, `eval.max` or `rel.tol`. For a complete list of options and default values see the documentation of `nlminb` in the `stats` package.

Details

Warning: Unless well-justified constraints are imposed on either λ and/or μ and ρ, it is generally impossible to reliably estimate λ and μ from extant timetrees alone (Louca and Pennell, 2019). This routine (and any other software that claims to estimate λ and μ solely from extant timetrees) should thus be used with great suspicion. If your only source of information is an extant timetree, and you have no a priori information on how λ or μ might have looked like, you should consider using the more appropriate routines `fit_hbd_pdr_on_grid` and `fit_hbd_psr_on_grid` instead.

If `age0>0`, the input tree is essentially trimmed at `age0` (omitting anything younger than `age0`), and the various variables are fitted to this new (shorter) tree, with time shifted appropriately. For example, the fitted `rho0` is thus the sampling fraction at `age0`, i.e. the fraction of lineages extant at `age0` that are represented in the timetree.

It is generally advised to provide as much information to the function `fit_hbd_model_on_grid` as possible, including reasonable lower and upper bounds (`min_lambda`, `max_lambda`, `min_mu`, `max_mu`, `min_rho0` and `max_rho0`) and a reasonable parameter guess (`guess_lambda`, `guess_mu` and `guess_rho0`). It is also important that the `age_grid` is sufficiently fine to capture the expected major variations of λ and μ over time, but keep in mind the serious risk of overfitting when `age_grid` is too fine and/or the tree is too small.

Value

A list with the following elements:

success   Logical, indicating whether model fitting succeeded. If `FALSE`, the returned list will include an additional "error" element (character) providing a description of the error; in that case all other return variables may be undefined.

objective_value

                The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.

objective_name   The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".

loglikelihood   The log-likelihood of the fitted model for the given timetree.

fitted_lambda   Numeric vector of size Ngrid, listing fitted or fixed speciation rates λ at each age-grid point. Between grid points λ should be interpreted as a piecewise polynomial function (natural spline) of degree `splines_degree`; to evaluate this function at arbitrary ages use the `castor` routine `evaluate_spline`.

fitted_mu   Numeric vector of size Ngrid, listing fitted or fixed extinction rates μ at each age-grid point. Between grid points μ should be interpreted as a piecewise polynomial function (natural spline) of degree `splines_degree`; to evaluate this function at arbitrary ages use the `castor` routine `evaluate_spline`.

fitted_rho   Numeric, specifying the fitted or fixed sampling fraction ρ.

guess_lambda	Numeric vector of size Ngrid, specifying the initial guess for λ at each age-grid point.
guess_mu	Numeric vector of size Ngrid, specifying the initial guess for μ at each age-grid point.
guess_rho0	Numeric, specifying the initial guess for ρ.
age_grid	The age-grid on which λ and μ are defined. This will be the same as the provided age_grid, unless the latter was NULL or of length <=1.
NFP	Integer, number of free (i.e., independently) fitted parameters. If none of the λ, μ and ρ were fixed, and const_lambda=FALSE and const_mu=FALSE, then NFP will be equal to 2*Ngrid+1.
AIC	The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.
BIC	The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of branching times), and L is the maximized likelihood.
condition	Character, specifying what conditioning was root for the likelihood (e.g. "crown" or "stem").
converged	Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase iter.max and/or eval.max).
Niterations	Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.
Nevaluations	Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.

Author(s)

Stilianos Louca

References

T. Stadler (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. Journal of Theoretical Biology. 261:58-66.

T. Stadler (2013). How can we improve accuracy of macroevolutionary rate estimates? Systematic Biology. 62:321-329.

H. Morlon, T. L. Parsons, J. B. Plotkin (2011). Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences. 108:16327-16332.

S. Louca et al. (2018). Bacterial diversification through geological time. Nature Ecology & Evolution. 2:1458-1467.

See Also

simulate_deterministic_hbd

loglikelihood_hbd

fit_hbd_model_parametric

fit_hbd_pdr_on_grid

fit_hbd_pdr_parametric

fit_hbd_psr_on_grid

Examples

```
## Not run:
# Generate a random tree with exponentially varying lambda & mu
Ntips    = 10000
rho      = 0.5 # sampling fraction
time_grid = seq(from=0, to=100, by=0.01)
lambdas  = 2*exp(0.1*time_grid)
mus      = 1.5*exp(0.09*time_grid)
sim      = generate_random_tree( parameters  = list(rarefaction=rho),
                                 max_tips    = Ntips/rho,
                                 coalescent  = TRUE,
                                 added_rates_times    = time_grid,
                                 added_birth_rates_pc = lambdas,
                                 added_death_rates_pc = mus)
tree = sim$tree
root_age = castor::get_tree_span(tree)$max_distance
cat(sprintf("Tree has %d tips, spans %g Myr\n",length(tree$tip.label),root_age))


# Fit mu on grid
# Assume that lambda & rho are known
Ngrid    = 5
age_grid  = seq(from=0,to=root_age,length.out=Ngrid)
fit = fit_hbd_model_on_grid(tree,
          age_grid    = age_grid,
          max_mu      = 100,
          fixed_lambda= approx(x=time_grid,y=lambdas,xout=sim$final_time-age_grid)$y,
          fixed_rho0  = rho,
          condition   = "crown",
          Ntrials     = 10,# perform 10 fitting trials
          Nthreads    = 2,# use two CPUs
          max_model_runtime = 1)  # limit model evaluation to 1 second
if(!fit$success){
  cat(sprintf("ERROR: Fitting failed: %s\n",fit$error))
}else{
  cat(sprintf("Fitting succeeded:\nLoglikelihood=%g\n",fit$loglikelihood))

  # plot fitted & true mu
  plot( x     = fit$age_grid,
        y     = fit$fitted_mu,
        main  = 'Fitted & true mu',
        xlab  = 'age',
        ylab  = 'mu',
        type  = 'b',
        col   = 'red',
        xlim  = c(root_age,0))
```

```
    lines(x    = sim$final_time-time_grid,
          y    = mus,
          type = 'l',
          col  = 'blue');

  # get fitted mu as a function of age
  mu_fun = approxfun(x=fit$age_grid, y=fit$fitted_mu)
}

## End(Not run)
```

fit_hbd_model_parametric

Fit a parametric homogenous birth-death model to a timetree.

Description

Given an ultrametric timetree, fit a homogenous birth-death (HBD) model in which speciation and extinction rates (λ and μ) are given as parameterized functions of time before present. "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction rates (in the literature this is sometimes referred to simply as "birth-death model"). Every HBD model is defined based on the values that λ and μ take over time as well as the sampling fraction ρ (fraction of extant species sampled); in turn, λ, μ and ρ can be parameterized by a finite set of parameters. This function estimates these parameters by maximizing the likelihood (Morlon et al. 2011) of the timetree under the resulting HBD model.

Usage

```
fit_hbd_model_parametric( tree,
                          param_values,
                          param_guess          = NULL,
                          param_min            = -Inf,
                          param_max            = +Inf,
                          param_scale          = NULL,
                          oldest_age           = NULL,
                          age0                 = 0,
                          lambda,
                          mu                   = 0,
                          rho0                 = 1,
                          age_grid             = NULL,
                          condition            = "auto",
                          relative_dt          = 1e-3,
                          Ntrials              = 1,
                          max_start_attempts   = 1,
                          Nthreads             = 1,
                          max_model_runtime    = NULL,
                          Nbootstraps          = 0,
                          Ntrials_per_bootstrap = NULL,
```

```
fit_algorithm       = "nlminb",
fit_control         = list(),
focal_param_values  = NULL,
verbose             = FALSE,
diagnostics         = FALSE,
verbose_prefix      = "")
```

Arguments

tree
: An ultrametric timetree of class "phylo", representing the time-calibrated reconstructed phylogeny of a set of extant species.

param_values
: Numeric vector, specifying fixed values for a some or all model parameters. For fitted (i.e., non-fixed) parameters, use NaN or NA. For example, the vector c(1.5,NA,40) specifies that the 1st and 3rd model parameters are fixed at the values 1.5 and 40, respectively, while the 2nd parameter is to be fitted. The length of this vector defines the total number of model parameters. If entries in this vector are named, the names are taken as parameter names. Names should be included if you'd like returned parameter vectors to have named entries, or if the functions lambda, mu or rho query parameter values by name (as opposed to numeric index).

param_guess
: Numeric vector of size NP, specifying a first guess for the value of each model parameter. For fixed parameters, guess values are ignored. Can be NULL only if all model parameters are fixed.

param_min
: Optional numeric vector of size NP, specifying lower bounds for model parameters. If of size 1, the same lower bound is applied to all parameters. Use -Inf to omit a lower bound for a parameter. If NULL, no lower bounds are applied. For fixed parameters, lower bounds are ignored.

param_max
: Optional numeric vector of size NP, specifying upper bounds for model parameters. If of size 1, the same upper bound is applied to all parameters. Use +Inf to omit an upper bound for a parameter. If NULL, no upper bounds are applied. For fixed parameters, upper bounds are ignored.

param_scale
: Optional numeric vector of size NP, specifying typical scales for model parameters. If of size 1, the same scale is assumed for all parameters. If NULL, scales are determined automatically. For fixed parameters, scales are ignored. It is strongly advised to provide reasonable scales, as this facilitates the numeric optimization algorithm.

oldest_age
: Strictly positive numeric, specifying the oldest time before present ("age") to consider when calculating the likelihood. If this is equal to or greater than the root age, then oldest_age is taken as the stem age, and the classical formula by Morlon et al. (2011) is used. If oldest_age is less than the root age, the tree is split into multiple subtrees at that age by treating every edge crossing that age as the stem of a subtree, and each subtree is considered an independent realization of the HBD model stemming at that age. This can be useful for avoiding points in the tree close to the root, where estimation uncertainty is generally higher. If oldest_age==NULL, it is automatically set to the root age.

age0
: Non-negative numeric, specifying the youngest age (time before present) to consider for fitting, and with respect to which rho is defined. If age0>0, then rho0

refers to the sampling fraction at age age0, i.e. the fraction of lineages extant at age0 that are included in the tree. See below for more details.

lambda	Function specifying the speciation rate at any given age (time before present) and for any given parameter values. This function must take exactly two arguments, the 1st one being a numeric vector (one or more ages) and the 2nd one being a numeric vector of size NP (parameter values), and return a numeric vector of the same size as the 1st argument with strictly positive entries. Can also be a single number (i.e., lambda is fixed).
mu	Function specifying the extinction rate at any given age and for any given parameter values. This function must take exactly two arguments, the 1st one being a numeric vector (one or more ages) and the 2nd one being a numeric vector of size NP (parameter values), and return a numeric vector of the same size as the 1st argument with non-negative entries. Can also be a single number (i.e., mu is fixed).
rho0	Function specifying the sampling fraction (fraction of extant species sampled at age0) for any given parameter values. This function must take exactly one argument, a numeric vector of size NP (parameter values), and return a numeric between 0 (exclusive) and 1 (inclusive). Can also be a single number (i.e., rho0 is fixed).
age_grid	Numeric vector, specifying ages at which the lambda and mu functionals should be evaluated. This age grid must be fine enough to capture the possible variation in λ and μ over time, within the permissible parameter range. If of size 1, then lambda & mu are assumed to be time-independent. Listed ages must be strictly increasing, and must cover at least the full considered age interval (from 0 to oldest_age). Can also be NULL or a vector of size 1, in which case the speciation rate and extinction rate is assumed to be time-independent.
condition	Character, either "crown", "stem", "none" or "auto", specifying on what to condition the likelihood. If "crown", the likelihood is conditioned on the survival of the two daughter lineages branching off at the root. If "stem", the likelihood is conditioned on the survival of the stem lineage. Note that "crown" really only makes sense when oldest_age is equal to the root age, while "stem" is recommended if oldest_age differs from the root age. "none" is usually not recommended. If "auto", the condition is chosen according to the recommendations mentioned earlier.
relative_dt	Strictly positive numeric (unitless), specifying the maximum relative time step allowed for integration over time, when calculating the likelihood. Smaller values increase integration accuracy but increase computation time. Typical values are 0.0001-0.001. The default is usually sufficient.
Ntrials	Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing Ntrials reduces the risk of reaching a non-global local maximum in the fitting objective.
max_start_attempts	
	Integer, specifying the number of times to attempt finding a valid start point (per trial) before giving up on that trial. Randomly choosen extreme start parameters may occasionally result in Inf/undefined likelihoods, so this option allows the algorithm to keep looking for valid starting points.

Nthreads
: Integer, specifying the number of parallel threads to use for performing multiple fitting trials simultaneously. This should generally not exceed the number of available CPUs on your machine. Parallel computing is not available on the Windows platform.

max_model_runtime
: Optional numeric, specifying the maximum number of seconds to allow for each evaluation of the likelihood function. Use this to abort fitting trials leading to parameter regions where the likelihood takes a long time to evaluate (these are often unlikely parameter regions).

Nbootstraps
: Integer, specifying the number of parametric bootstraps to perform for estimating standard errors and confidence intervals of estimated model parameters. Set to 0 for no bootstrapping.

Ntrials_per_bootstrap
: Integer, specifying the number of fitting trials to perform for each bootstrap sampling. If `NULL`, this is set equal to `max(1,Ntrials)`. Decreasing `Ntrials_per_bootstrap` will reduce computation time, at the expense of potentially inflating the estimated confidence intervals; in some cases (e.g., for very large trees) this may be useful if fitting takes a long time and confidence intervals are very narrow anyway. Only relevant if `Nbootstraps>0`.

fit_algorithm
: Character, specifying which optimization algorithm to use. Either "nlminb" or "subplex" are allowed.

fit_control
: Named list containing options for the `nlminb` or `subplex` optimization routine, depending on the choice of `fit_algorithm`. For example, for "nlminb" commonly modified options are `iter.max`, `eval.max` or `rel.tol`. For a complete list of options and default values see the documentation of `nlminb` in the `stats` package or of `nloptr` in the `nloptr` package.

focal_param_values
: Optional numeric matrix having NP columns and an arbitrary number of rows, listing combinations of parameter values of particular interest and for which the log-likelihoods should be returned. This may be used for diagnostic purposes, e.g., to examine the shape of the likelihood function.

verbose
: Logical, specifying whether to print progress reports and warnings to the screen. Note that errors always cause a return of the function (see return values `success` and `error`).

diagnostics
: Logical, specifying whether to print detailed information (such as model likelihoods) at every iteration of the fitting routine. For debugging purposes mainly.

verbose_prefix
: Character, specifying the line prefix for printing progress reports to the screen.

Details

This function is designed to estimate a finite set of scalar parameters $(p_1, .., p_n \in \mathsf{R})$ that determine the speciation rate λ, the extinction rate μ and the sampling fraction ρ, by maximizing the likelihood of observing a given timetree under the HBD model. For example, the investigator may assume that both λ and μ vary exponentially over time, i.e. they can be described by $\lambda(t) = \lambda_o \cdot e^{-\alpha t}$ and $\mu(t) = \mu_o \cdot e^{-\beta t}$ (where λ_o, μ_o are unknown present-day rates and α, β are unknown factors, and t is time before present), and that the sampling fraction ρ is known. In this case the model has 4 free parameters, $p_1 = \lambda_o$, $p_2 = \mu_o$, $p_3 = \alpha$ and $p_4 = \beta$, each of which may be fitted to the tree.

It is generally advised to provide as much information to the function `fit_hbd_model_parametric` as possible, including reasonable lower and upper bounds (`param_min` and `param_max`), a reasonable parameter guess (`param_guess`) and reasonable parameter scales `param_scale`. If some model parameters can vary over multiple orders of magnitude, it is advised to transform them so that they vary across fewer orders of magnitude (e.g., via log-transformation). It is also important that the `age_grid` is sufficiently fine to capture the variation of `lambda` and `mu` over time, since the likelihood is calculated under the assumption that both vary linearly between grid points.

Value

A list with the following elements:

`success`	Logical, indicating whether model fitting succeeded. If `FALSE`, the returned list will include an additional "error" element (character) providing a description of the error; in that case all other return variables may be undefined.
`objective_value`	
	The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.
`objective_name`	The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".
`param_fitted`	Numeric vector of size NP (number of model parameters), listing all fitted or fixed model parameters in their standard order (see details above). If `param_names` was provided, elements in `fitted_params` will be named.
`param_guess`	Numeric vector of size NP, listing guessed or fixed values for all model parameters in their standard order. If `param_names` was provided, elements in `param_guess` will be named.
`loglikelihood`	The log-likelihood of the fitted model for the given timetree.
`NFP`	Integer, number of fitted (i.e., non-fixed) model parameters.
`AIC`	The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.
`BIC`	The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of branching times), and L is the maximized likelihood.
`condition`	Character, specifying what conditioning was root for the likelihood (e.g. `"crown"` or `"stem"`).
`converged`	Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase `iter.max` and/or `eval.max`).
`Niterations`	Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.
`Nevaluations`	Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.

trial_start_objectives

> Numeric vector of size Ntrials, listing the initial objective values (e.g., log-likelihoods) for each fitting trial, i.e. at the start parameter values.

trial_objective_values

> Numeric vector of size Ntrials, listing the final maximized objective values (e.g., loglikelihoods) for each fitting trial.

trial_Nstart_attempts

> Integer vector of size Ntrials, listing the number of start attempts for each fitting trial, until a starting point with valid likelihood was found.

trial_Niterations

> Integer vector of size Ntrials, listing the number of iterations needed for each fitting trial.

trial_Nevaluations

> Integer vector of size Ntrials, listing the number of likelihood evaluations needed for each fitting trial.

standard_errors

> Numeric vector of size NP, estimated standard error of the parameters, based on parametric bootstrapping. Only returned if Nbootstraps>0.

medians   Numeric vector of size NP, median the estimated parameters across parametric bootstraps. Only returned if Nbootstraps>0.

CI50lower   Numeric vector of size NP, lower bound of the 50% confidence interval (25-75% percentile) for the parameters, based on parametric bootstrapping. Only returned if Nbootstraps>0.

CI50upper   Numeric vector of size NP, upper bound of the 50% confidence interval for the parameters, based on parametric bootstrapping. Only returned if Nbootstraps>0.

CI95lower   Numeric vector of size NP, lower bound of the 95% confidence interval (2.5-97.5% percentile) for the parameters, based on parametric bootstrapping. Only returned if Nbootstraps>0.

CI95upper   Numeric vector of size NP, upper bound of the 95% confidence interval for the parameters, based on parametric bootstrapping. Only returned if Nbootstraps>0.

consistency   Numeric between 0 and 1, estimated consistency of the data with the fitted model (Lindholm et al. 2019). See the documentation of `fit_hbds_model_on_grid` for an explanation.

Author(s)

Stilianos Louca

References

H. Morlon, T. L. Parsons, J. B. Plotkin (2011). Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences. 108:16327-16332.

S. Louca et al. (2018). Bacterial diversification through geological time. Nature Ecology & Evolution. 2:1458-1467.

A. Lindholm, D. Zachariah, P. Stoica, T. B. Schoen (2019). Data consistency approach to model validation. IEEE Access. 7:59788-59796.

See Also

  simulate_deterministic_hbd

  loglikelihood_hbd

  fit_hbd_model_on_grid

  fit_hbd_pdr_on_grid

  fit_hbd_pdr_parametric

Examples

```
## Not run:
# Generate a random tree with exponentially varying lambda & mu
Ntips     = 10000
rho       = 0.5 # sampling fraction
time_grid = seq(from=0, to=100, by=0.01)
lambdas   = 2*exp(0.1*time_grid)
mus       = 1.5*exp(0.09*time_grid)
tree      = generate_random_tree( parameters  = list(rarefaction=rho),
                                  max_tips     = Ntips/rho,
                                  coalescent   = TRUE,
                                  added_rates_times     = time_grid,
                                  added_birth_rates_pc  = lambdas,
                                  added_death_rates_pc  = mus)$tree
root_age = castor::get_tree_span(tree)$max_distance
cat(sprintf("Tree has %d tips, spans %g Myr\n",length(tree$tip.label),root_age))


# Define a parametric HBD model, with exponentially varying lambda & mu
# Assume that the sampling fraction is known
# The model thus has 4 parameters: lambda0, mu0, alpha, beta
lambda_function = function(ages,params){
return(params['lambda0']*exp(-params['alpha']*ages));
}
mu_function = function(ages,params){
return(params['mu0']*exp(-params['beta']*ages));
}
rho_function = function(params){
return(rho) # rho does not depend on any of the parameters
}

# Define an age grid on which lambda_function & mu_function shall be evaluated
# Should be sufficiently fine to capture the variation in lambda & mu
age_grid = seq(from=0,to=100,by=0.01)

# Perform fitting
# Lets suppose extinction rates are already known
cat(sprintf("Fitting model to tree..\n"))
fit = fit_hbd_model_parametric( tree,
                        param_values  = c(lambda0=NA, mu0=3, alpha=NA, beta=-0.09),
                        param_guess   = c(1,1,0,0),
                        param_min     = c(0,0,-1,-1),
                        param_max     = c(10,10,1,1),
```

```
                        param_scale    = 1, # all params are in the order of 1
                        lambda         = lambda_function,
                        mu             = mu_function,
                        rho0           = rho_function,
                        age_grid       = age_grid,
                        Ntrials        = 10,   # perform 10 fitting trials
                        Nthreads       = 2,    # use 2 CPUs
                        max_model_runtime = 1, # limit model evaluation to 1 second
                        fit_control    = list(rel.tol=1e-6))
if(!fit$success){
cat(sprintf("ERROR: Fitting failed: %s\n",fit$error))
}else{
cat(sprintf("Fitting succeeded:\nLoglikelihood=%g\n",fit$loglikelihood))
print(fit)
}

## End(Not run)
```

---

fit_hbd_pdr_on_grid   *Fit pulled diversification rates of birth-death models on a time grid.*

Description

Given an ultrametric timetree, estimate the pulled diversification rate of homogenous birth-death (HBD) models that best explains the tree via maximum likelihood. Every HBD model is defined by some speciation and extinction rates (λ and μ) over time, as well as the sampling fraction ρ (fraction of extant species sampled). "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction rates. For any given HBD model there exists an infinite number of alternative HBD models that predict the same deterministic lineages-through-time curve and yield the same likelihood for any given reconstructed timetree; these "congruent" models cannot be distinguished from one another solely based on the tree.

Each congruence class is uniquely described by the "pulled diversification rate" (PDR; Louca et al 2018), defined as $PDR = \lambda - \mu + \lambda^{-1} d\lambda/d\tau$ (where τ is time before present) as well as the product $\rho\lambda_o$ (where λ_o is the present-day speciation rate). That is, two HBD models are congruent if and only if they have the same PDR and the same product $\rho\lambda_o$. This function is designed to estimate the generating congruence class for the tree, by fitting the PDR on a grid of discrete times as well as the product $\rho\lambda_o$.

Usage

```
fit_hbd_pdr_on_grid( tree,
                     oldest_age          = NULL,
                     age0                = 0,
                     age_grid            = NULL,
                     min_PDR             = -Inf,
                     max_PDR             = +Inf,
                     min_rholambda0      = 1e-10,
                     max_rholambda0      = +Inf,
```

```
guess_PDR             = NULL,
guess_rholambda0      = NULL,
fixed_PDR             = NULL,
fixed_rholambda0      = NULL,
splines_degree        = 1,
condition             = "auto",
relative_dt           = 1e-3,
Ntrials               = 1,
Nbootstraps           = 0,
Ntrials_per_bootstrap = NULL,
Nthreads              = 1,
max_model_runtime     = NULL,
fit_control           = list(),
verbose               = FALSE,
verbose_prefix        = "")
```

Arguments

tree   An ultrametric timetree of class "phylo", representing the time-calibrated phy-
                logeny of a set of extant species.

oldest_age   Strictly positive numeric, specifying the oldest time before present ("age") to
                consider when calculating the likelihood. If this is equal to or greater than the
                root age, then oldest_age is taken as the stem age, and the classical formula by
                Morlon et al. (2011) is used. If oldest_age is less than the root age, the tree is
                split into multiple subtrees at that age by treating every edge crossing that age as
                the stem of a subtree, and each subtree is considered an independent realization
                of the HBD model stemming at that age. This can be useful for avoiding points
                in the tree close to the root, where estimation uncertainty is generally higher. If
                oldest_age==NULL, it is automatically set to the root age.

age0   Non-negative numeric, specifying the youngest age (time before present) to con-
                sider for fitting, and with respect to which rholambda0 is defined. If age0>0,
                then rholambda0 refers to the product of the sampling fraction at age age0 and
                the speciation rate at age age0. See below for more details.

age_grid   Numeric vector, listing ages in ascending order at which the PDR is allowed to
                vary independently. This grid must cover at least the age range from age0 to
                oldest_age. If NULL or of length <=1 (regardless of value), then the PDR is
                assumed to be time-independent.

min_PDR   Numeric vector of length Ngrid (=max(1,length(age_grid))), or a single nu-
                meric, specifying lower bounds for the fitted PDR at each point in the age grid.
                If a single numeric, the same lower bound applies at all ages. Use -Inf to omit
                lower bounds.

max_PDR   Numeric vector of length Ngrid, or a single numeric, specifying upper bounds
                for the fitted PDR at each point in the age grid. If a single numeric, the same
                upper bound applies at all ages. Use +Inf to omit upper bounds.

min_rholambda0 Strictly positive numeric, specifying the lower bound for the fitted $\rho\lambda_o$ (sam-
                pling fraction times present-day extinction rate).

max_rholambda0   Strictly positive numeric, specifying the upper bound for the fitted $\rho\lambda_o$. Set to +Inf to omit this upper bound.

guess_PDR   Initial guess for the PDR at each age-grid point. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same PDR at all ages) or a numeric vector of size Ngrid specifying a separate guess at each age-grid point. To omit an initial guess for some but not all age-grid points, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.

guess_rholambda0

   Numeric, specifying an initial guess for the product $\rho\lambda_o$. If NULL, a guess will be computed automatically.

fixed_PDR   Optional fixed (i.e. non-fitted) PDR values on one or more age-grid points. Either NULL (PDR is not fixed anywhere), or a single numeric (PDR fixed to the same value at all grid points) or a numeric vector of size Ngrid (PDR fixed at one or more age-grid points, use NA for non-fixed values).

fixed_rholambda0

   Numeric, optionally specifying a fixed value for the product $\rho\lambda_o$. If NULL or NA, the product $\rho\lambda_o$ is estimated.

splines_degree   Integer between 0 and 3 (inclusive), specifying the polynomial degree of the PDR between age-grid points. If 0, then the PDR is considered piecewise constant, if 1 then the PDR is considered piecewise linear, if 2 or 3 then the PDR is considered to be a spline of degree 2 or 3, respectively. The splines_degree influences the analytical properties of the curve, e.g. splines_degree==1 guarantees a continuous curve, splines_degree==2 guarantees a continuous curve and continuous derivative, and so on. A degree of 0 is generally not recommended.

condition   Character, either "crown", "stem" or "auto", specifying on what to condition the likelihood. If "crown", the likelihood is conditioned on the survival of the two daughter lineages branching off at the root. If "stem", the likelihood is conditioned on the survival of the stem lineage. Note that "crown" really only makes sense when oldest_age is equal to the root age, while "stem" is recommended if oldest_age differs from the root age. If "auto", the condition is chosen according to the recommendations mentioned earlier.

relative_dt   Strictly positive numeric (unitless), specifying the maximum relative time step allowed for integration over time, when calculating the likelihood. Smaller values increase integration accuracy but increase computation time. Typical values are 0.0001-0.001. The default is usually sufficient.

Ntrials   Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing Ntrials reduces the risk of reaching a non-global local maximum in the fitting objective.

Nbootstraps   Integer, specifying an optional number of bootstrap samplings to perform, for estimating standard errors and confidence intervals of maximum-likelihood fitted parameters. If 0, no bootstrapping is performed. Typical values are 10-100. At each bootstrap sampling, a random timetree is generated under the birth-death model according to the fitted PDR and $\rho\lambda_o$, the parameters are estimated anew based on the generated tree, and subsequently compared to the original

fitted parameters. Each bootstrap sampling will use roughly the same informa-
tion and similar computational resources as the original maximum-likelihood fit
(e.g., same number of trials, same optimization parameters, same initial guess,
etc).

Ntrials_per_bootstrap
Integer, specifying the number of fitting trials to perform for each bootstrap sam-
pling. If NULL, this is set equal to max(1,Ntrials). Decreasing Ntrials_per_bootstrap
will reduce computation time, at the expense of potentially inflating the esti-
mated confidence intervals; in some cases (e.g., for very large trees) this may
be useful if fitting takes a long time and confidence intervals are very narrow
anyway. Only relevant if Nbootstraps>0.

Nthreads   Integer, specifying the number of parallel threads to use for performing multiple
fitting trials simultaneously. This should generally not exceed the number of
available CPUs on your machine. Parallel computing is not available on the
Windows platform.

max_model_runtime
Optional numeric, specifying the maximum number of seconds to allow for each
evaluation of the likelihood function. Use this to abort fitting trials leading to
parameter regions where the likelihood takes a long time to evaluate (these are
often unlikely parameter regions).

fit_control   Named list containing options for the nlminb optimization routine, such as
iter.max, eval.max or rel.tol. For a complete list of options and default
values see the documentation of nlminb in the stats package.

verbose   Logical, specifying whether to print progress reports and warnings to the screen.
Note that errors always cause a return of the function (see return values success
and error).

verbose_prefix   Character, specifying the line prefix for printing progress reports to the screen.

Details

If age0>0, the input tree is essentially trimmed at age0 (omitting anything younger than age0), and
the PDR and rholambda0 are fitted to this new (shorter) tree, with time shifted appropriately. The
fitted rholambda0 is thus the product of the sampling fraction at age0 and the speciation rate at
age0. Note that the sampling fraction at age0 is simply the fraction of lineages extant at age0 that
are represented in the timetree.

It is generally advised to provide as much information to the function fit_hbd_pdr_on_grid as
possible, including reasonable lower and upper bounds (min_PDR, max_PDR, min_rholambda0 and
max_rholambda0) and a reasonable parameter guess (guess_PDR and guess_rholambda0). It is
also important that the age_grid is sufficiently fine to capture the expected major variations of the
PDR over time, but keep in mind the serious risk of overfitting when age_grid is too fine and/or
the tree is too small.

Value

A list with the following elements:

success	Logical, indicating whether model fitting succeeded. If FALSE, the returned list will include an additional "error" element (character) providing a description of the error; in that case all other return variables may be undefined.
objective_value	The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.
objective_name	The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".
loglikelihood	The log-likelihood of the fitted model for the given timetree.
fitted_PDR	Numeric vector of size Ngrid, listing fitted or fixed pulled diversification rates (PDR) at each age-grid point. Between grid points the fitted PDR should be interpreted as a piecewise polynomial function (natural spline) of degree splines_degree; to evaluate this function at arbitrary ages use the castor routine [evaluate_spline]().
fitted_rholambda0	Numeric, specifying the fitted or fixed product $\rho\lambda(0)$.
guess_PDR	Numeric vector of size Ngrid, specifying the initial guess for the PDR at each age-grid point.
guess_rholambda0	Numeric, specifying the initial guess for $\rho\lambda(0)$.
age_grid	The age-grid on which the PDR is defined. This will be the same as the provided age_grid, unless the latter was NULL or of length <=1.
NFP	Integer, number of fitted (i.e., non-fixed) parameters. If none of the PDRs or $\rho\lambda0$ were fixed, this will be equal to Ngrid+1.
AIC	The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.
BIC	The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of branching times), and L is the maximized likelihood.
converged	Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase iter.max and/or eval.max).
Niterations	Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.
Nevaluations	Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.
bootstrap_estimates	If Nbootstraps>0, this will be a named list containing the elements PDR (numeric matrix of size Nbootstraps x Ngrid, listing the fitted PDR at each grid point and for each bootstrap) and rholambda0 (a numeric vector of size Nbootstraps, listing the fitted $\rho\lambda_o$ for each bootstrap).
standard_errors	If Nbootstraps>0, this will be a named list containing the elements PDR (numeric vector of size Ngrid, listing bootstrap-estimated standard errors for the

fitted PDRs) and rholambda0 (a single numeric, bootstrap-estimated standard error for the fitted $\rho\lambda_o$).

medians   If Nbootstraps>0, this will be a named list containing the elements PDR (numeric vector of size Ngrid, listing median fitted PDRs across bootstraps) and rholambda0 (a single numeric, median fitted $\rho\lambda_o$ across bootstraps).

CI50lower   If Nbootstraps>0, this will be a named list containing the elements PDR (numeric vector of size Ngrid, listing bootstrap-estimated lower bounds of the 50-percent confidence intervals for the fitted PDRs) and rholambda0 (a single numeric, bootstrap-estimated lower bound of the 50-percent confidence intervals for the fitted $\rho\lambda_o$).

CI50upper   Similar to CI50lower, listing upper bounds of 50-percentile confidence intervals.

CI95lower   Similar to CI50lower, listing lower bounds of 95-percentile confidence intervals.

CI95upper   Similar to CI95lower, listing upper bounds of 95-percentile confidence intervals.

Author(s)

Stilianos Louca

References

S. Louca et al. (2018). Bacterial diversification through geological time. Nature Ecology & Evolution. 2:1458-1467.

See Also

simulate_deterministic_hbd

loglikelihood_hbd

fit_hbd_model_parametric

fit_hbd_model_on_grid

fit_hbd_pdr_parametric

Examples

```
## Not run:
# Generate a random tree with exponentially varying lambda & mu
Ntips     = 10000
rho       = 0.5 # sampling fraction
time_grid = seq(from=0, to=100, by=0.01)
lambdas   = 2*exp(0.1*time_grid)
mus       = 1.5*exp(0.09*time_grid)
sim       = generate_random_tree( parameters  = list(rarefaction=rho),
                                  max_tips    = Ntips/rho,
                                  coalescent  = TRUE,
                                  added_rates_times    = time_grid,
```

```
                                         added_birth_rates_pc  = lambdas,
                                         added_death_rates_pc  = mus)
    tree = sim$tree
    root_age = castor::get_tree_span(tree)$max_distance
    cat(sprintf("Tree has %d tips, spans %g Myr\n",length(tree$tip.label),root_age))


    # calculate true PDR
    lambda_slopes = diff(lambdas)/diff(time_grid);
    lambda_slopes = c(lambda_slopes[1],lambda_slopes)
    PDRs = lambdas - mus - (lambda_slopes/lambdas)


    # Fit PDR on grid
    Ngrid     = 10
    age_grid  = seq(from=0,to=root_age,length.out=Ngrid)
    fit = fit_hbd_pdr_on_grid(tree,
              age_grid    = age_grid,
              min_PDR     = -100,
              max_PDR     = +100,
              condition   = "crown",
              Ntrials     = 10,# perform 10 fitting trials
              Nthreads    = 2,# use two CPUs
              max_model_runtime = 1)  # limit model evaluation to 1 second
    if(!fit$success){
      cat(sprintf("ERROR: Fitting failed: %s\n",fit$error))
    }else{
      cat(sprintf("Fitting succeeded:\nLoglikelihood=%g\n",fit$loglikelihood))

      # plot fitted & true PDR
      plot( x      = fit$age_grid,
            y      = fit$fitted_PDR,
            main = 'Fitted & true PDR',
            xlab = 'age',
            ylab = 'PDR',
            type = 'b',
            col  = 'red',
            xlim = c(root_age,0))
      lines(x      = sim$final_time-time_grid,
            y      = PDRs,
            type = 'l',
            col  = 'blue');

      # get fitted PDR as a function of age
      PDR_fun = approxfun(x=fit$age_grid, y=fit$fitted_PDR)
    }

    ## End(Not run)
```

---

fit_hbd_pdr_parametric

Fit parameterized pulled diversification rates of birth-death models.

Description

Given an ultrametric timetree, estimate the pulled diversification rate (PDR) of homogenous birth-death (HBD) models that best explains the tree via maximum likelihood, assuming that the PDR is given as a parameterized function of time before present. Every HBD model is defined by some speciation and extinction rates (λ and μ) over time, as well as the sampling fraction ρ (fraction of extant species sampled). "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction rates. For any given HBD model there exists an infinite number of alternative HBD models that predict the same deterministic lineages-through-time curve and yield the same likelihood for any given reconstructed timetree; these "congruent" models cannot be distinguished from one another solely based on the tree.

Each congruence class is uniquely described by its PDR, defined as $PDR = \lambda - \mu + \lambda^{-1}d\lambda/d\tau$ (where τ is time before present) as well as the product $\rho\lambda_o$ (where λ_o is the present-day speciation rate). That is, two HBD models are congruent if and only if they have the same PDR and the same product $\rho\lambda_o$. This function is designed to estimate the generating congruence class for the tree, by fitting a finite number of parameters defining the PDR and $\rho\lambda_o$.

Usage

```
fit_hbd_pdr_parametric( tree,
                        param_values,
                        param_guess       = NULL,
                        param_min         = -Inf,
                        param_max         = +Inf,
                        param_scale       = NULL,
                        oldest_age        = NULL,
                        age0              = 0,
                        PDR,
                        rholambda0,
                        age_grid          = NULL,
                        condition         = "auto",
                        relative_dt       = 1e-3,
                        Ntrials           = 1,
                        max_start_attempts = 1,
                        Nthreads          = 1,
                        max_model_runtime = NULL,
                        fit_control       = list())
```

Arguments

tree   An ultrametric timetree of class "phylo", representing the time-calibrated phylogeny of a set of extant species.

param_values   Numeric vector, specifying fixed values for a some or all model parameters. For fitted (i.e., non-fixed) parameters, use NaN or NA. For example, the vector c(1.5,NA,40) specifies that the 1st and 3rd model parameters are fixed at the values 1.5 and 40, respectively, while the 2nd parameter is to be fitted. The length of this vector defines the total number of model parameters. If entries in this vector are named, the names are taken as parameter names. Names should be included if you'd like returned parameter vectors to have named entries, or if

	the functions PDR or rho query parameter values by name (as opposed to numeric index).
param_guess	Numeric vector of size NP, specifying a first guess for the value of each model parameter. For fixed parameters, guess values are ignored. Can be NULL only if all model parameters are fixed.
param_min	Optional numeric vector of size NP, specifying lower bounds for model parameters. If of size 1, the same lower bound is applied to all parameters. Use -Inf to omit a lower bound for a parameter. If NULL, no lower bounds are applied. For fixed parameters, lower bounds are ignored.
param_max	Optional numeric vector of size NP, specifying upper bounds for model parameters. If of size 1, the same upper bound is applied to all parameters. Use +Inf to omit an upper bound for a parameter. If NULL, no upper bounds are applied. For fixed parameters, upper bounds are ignored.
param_scale	Optional numeric vector of size NP, specifying typical scales for model parameters. If of size 1, the same scale is assumed for all parameters. If NULL, scales are determined automatically. For fixed parameters, scales are ignored. It is strongly advised to provide reasonable scales, as this facilitates the numeric optimization algorithm.
oldest_age	Strictly positive numeric, specifying the oldest time before present ("age") to consider when calculating the likelihood. If this is equal to or greater than the root age, then oldest_age is taken as the stem age, and the classical formula by Morlon et al. (2011) is used. If oldest_age is less than the root age, the tree is split into multiple subtrees at that age by treating every edge crossing that age as the stem of a subtree, and each subtree is considered an independent realization of the HBD model stemming at that age. This can be useful for avoiding points in the tree close to the root, where estimation uncertainty is generally higher. If oldest_age==NULL, it is automatically set to the root age.
age0	Non-negative numeric, specifying the youngest age (time before present) to consider for fitting, and with respect to which rholambda0 is defined. If age0>0, then rholambda0 refers to the product of the sampling fraction at age age0 and the speciation rate at age age0. See below for more details.
PDR	Function specifying the pulled diversification rate at any given age (time before present) and for any given parameter values. This function must take exactly two arguments, the 1st one being a numeric vector (one or more ages) and the 2nd one being a numeric vector of size NP (parameter values), and return a numeric vector of the same size as the 1st argument. Can also be a single number (i.e., PDR is fixed).
rholambda0	Function specifying the product $\rho\lambda_o$ (sampling fraction times speciation rate at age0) for any given parameter values. This function must take exactly one argument, a numeric vector of size NP (parameter values), and return a strictly positive numeric. Can also be a single number (i.e., rholambda0 is fixed).
age_grid	Numeric vector, specifying ages at which the PDR function should be evaluated. This age grid must be fine enough to capture the possible variation in the PDR over time, within the permissible parameter range. If of size 1, then the PDR is assumed to be time-independent. Listed ages must be strictly increasing, and must cover at least the full considered age interval (from age0 to oldest_age).

Can also be NULL or a vector of size 1, in which case the PDR is assumed to be time-independent.

condition   Character, either "crown", "stem" or "auto", specifying on what to condition the likelihood. If "crown", the likelihood is conditioned on the survival of the two daughter lineages branching off at the root. If "stem", the likelihood is conditioned on the survival of the stem lineage. Note that "crown" really only makes sense when oldest_age is equal to the root age, while "stem" is recommended if oldest_age differs from the root age. If "auto", the condition is chosen according to the recommendations mentioned earlier.

relative_dt   Strictly positive numeric (unitless), specifying the maximum relative time step allowed for integration over time, when calculating the likelihood. Smaller values increase integration accuracy but increase computation time. Typical values are 0.0001-0.001. The default is usually sufficient.

Ntrials   Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing Ntrials reduces the risk of reaching a non-global local maximum in the fitting objective.

max_start_attempts
                     Integer, specifying the number of times to attempt finding a valid start point (per trial) before giving up on that trial. Randomly choosen extreme start parameters may occasionally result in Inf/undefined likelihoods, so this option allows the algorithm to keep looking for valid starting points.

Nthreads   Integer, specifying the number of parallel threads to use for performing multiple fitting trials simultaneously. This should generally not exceed the number of available CPUs on your machine. Parallel computing is not available on the Windows platform.

max_model_runtime
                     Optional numeric, specifying the maximum number of seconds to allow for each evaluation of the likelihood function. Use this to abort fitting trials leading to parameter regions where the likelihood takes a long time to evaluate (these are often unlikely parameter regions).

fit_control   Named list containing options for the nlminb optimization routine, such as iter.max, eval.max or rel.tol. For a complete list of options and default values see the documentation of nlminb in the stats package.

Details

This function is designed to estimate a finite set of scalar parameters ($p_1, .., p_n \in$ R) that determine the PDR and the product $\rho\lambda_o$ (sampling fraction times present-dat extinction rate), by maximizing the likelihood of observing a given timetree under the HBD model. For example, the investigator may assume that the PDR varies exponentially over time, i.e. can be described by $PDR(t) = A \cdot e^{-Bt}$ (where A and B are unknown coefficients and t is time before present), and that the product $\rho\lambda_o$ is unknown. In this case the model has 3 free parameters, $p_1 = A$, $p_2 = B$ and $p_3 = \rho\lambda_o$, each of which may be fitted to the tree.

If age0>0, the input tree is essentially trimmed at age0 (omitting anything younger than age0), and the PDR and rholambda0 are fitted to this new (shorter) tree, with time shifted appropriately. The fitted rholambda0 is thus the product of the sampling fraction at age0 and the speciation rate at

age0. Note that the sampling fraction at age0 is simply the fraction of lineages extant at age0 that are represented in the timetree.

It is generally advised to provide as much information to the function fit_hbd_pdr_parametric as possible, including reasonable lower and upper bounds (param_min and param_max), a reasonable parameter guess (param_guess) and reasonable parameter scales param_scale. If some model parameters can vary over multiple orders of magnitude, it is advised to transform them so that they vary across fewer orders of magnitude (e.g., via log-transformation). It is also important that the age_grid is sufficiently fine to capture the variation of the PDR over time, since the likelihood is calculated under the assumption that both vary linearly between grid points.

Value

A list with the following elements:

success	Logical, indicating whether model fitting succeeded. If FALSE, the returned list will include an additional "error" element (character) providing a description of the error; in that case all other return variables may be undefined.
objective_value	The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.
objective_name	The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".
param_fitted	Numeric vector of size NP (number of model parameters), listing all fitted or fixed model parameters in their standard order (see details above). If param_names was provided, elements in fitted_params will be named.
param_guess	Numeric vector of size NP, listing guessed or fixed values for all model parameters in their standard order.
loglikelihood	The log-likelihood of the fitted model for the given timetree.
NFP	Integer, number of fitted (i.e., non-fixed) model parameters.
AIC	The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.
BIC	The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of branching times), and L is the maximized likelihood.
converged	Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase iter.max and/or eval.max).
Niterations	Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.
Nevaluations	Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.
trial_start_objectives	Numeric vector of size Ntrials, listing the initial objective values (e.g., log-likelihoods) for each fitting trial, i.e. at the start parameter values.

trial_objective_values

> Numeric vector of size `Ntrials`, listing the final maximized objective values (e.g., loglikelihoods) for each fitting trial.

trial_Nstart_attempts

> Integer vector of size `Ntrials`, listing the number of start attempts for each fitting trial, until a starting point with valid likelihood was found.

trial_Niterations

> Integer vector of size `Ntrials`, listing the number of iterations needed for each fitting trial.

trial_Nevaluations

> Integer vector of size `Ntrials`, listing the number of likelihood evaluations needed for each fitting trial.

Author(s)

Stilianos Louca

References

H. Morlon, T. L. Parsons, J. B. Plotkin (2011). Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences. 108:16327-16332.

S. Louca et al. (2018). Bacterial diversification through geological time. Nature Ecology & Evolution. 2:1458-1467.

See Also

[simulate_deterministic_hbd](#)

[loglikelihood_hbd](#)

[fit_hbd_model_on_grid](#)

[fit_hbd_model_parametric](#)

[fit_hbd_pdr_on_grid](#)

Examples

```
## Not run:
# Generate a random tree with exponentially varying lambda & mu
Ntips     = 10000
rho       = 0.5 # sampling fraction
time_grid = seq(from=0, to=100, by=0.01)
lambdas   = 2*exp(0.1*time_grid)
mus       = 1.5*exp(0.09*time_grid)
tree      = generate_random_tree( parameters  = list(rarefaction=rho),
                                  max_tips     = Ntips/rho,
                                  coalescent   = TRUE,
                                  added_rates_times    = time_grid,
                                  added_birth_rates_pc = lambdas,
                                  added_death_rates_pc = mus)$tree
root_age = castor::get_tree_span(tree)$max_distance
cat(sprintf("Tree has %d tips, spans %g Myr\n",length(tree$tip.label),root_age))
```

```
# Define a parametric HBD congruence class, with exponentially varying PDR
# The model thus has 3 parameters
PDR_function = function(ages,params){
return(params['A']*exp(-params['B']*ages));
}
rholambda0_function = function(params){
return(params['rholambda0'])
}

# Define an age grid on which lambda_function & mu_function shall be evaluated
# Should be sufficiently fine to capture the variation in the PDR
age_grid = seq(from=0,to=100,by=0.01)

# Perform fitting
# Lets suppose extinction rates are already known
cat(sprintf("Fitting class to tree..\n"))
fit = fit_hbd_pdr_parametric( tree,
                    param_values  = c(A=NA, B=NA, rholambda0=NA),
                    param_guess   = c(1,0,1),
                    param_min     = c(-10,-10,0),
                    param_max     = c(10,10,10),
                    param_scale   = 1, # all params are in the order of 1
                    PDR           = PDR_function,
                    rholambda0    = rholambda0_function,
                    age_grid      = age_grid,
                    Ntrials       = 10,    # perform 10 fitting trials
                    Nthreads      = 2,     # use 2 CPUs
                    max_model_runtime = 1, # limit model evaluation to 1 second
                    fit_control       = list(rel.tol=1e-6))
if(!fit$success){
cat(sprintf("ERROR: Fitting failed: %s\n",fit$error))
}else{
cat(sprintf("Fitting succeeded:\nLoglikelihood=%g\n",fit$loglikelihood))
print(fit)
}

## End(Not run)
```

---

fit_hbd_psr_on_grid   *Fit pulled speciation rates of birth-death models on a time grid.*

Description

Given an ultrametric timetree, estimate the pulled speciation rate of homogenous birth-death (HBD) models that best explains the tree via maximum likelihood. Every HBD model is defined by some speciation and extinction rates (λ and μ) over time, as well as the sampling fraction ρ (fraction of extant species sampled). "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction rates. For any given HBD model there exists an

infinite number of alternative HBD models that predict the same deterministic lineages-through-time curve and yield the same likelihood for any given reconstructed timetree; these "congruent" models cannot be distinguished from one another solely based on the tree.

Each congruence class is uniquely described by the "pulled speciation rate" (PSR), defined as the relative slope of the deterministic LTT over time, $PSR = -M^{-1}dM/d\tau$ (where τ is time before present). In other words, two HBD models are congruent if and only if they have the same PSR. This function is designed to estimate the generating congruence class for the tree, by fitting the PSR on a discrete time grid.

Usage

```
fit_hbd_psr_on_grid(  tree,
                      oldest_age            = NULL,
                      age0                  = 0,
                      age_grid              = NULL,
                      min_PSR               = 0,
                      max_PSR               = +Inf,
                      guess_PSR             = NULL,
                      fixed_PSR             = NULL,
                      splines_degree        = 1,
                      condition             = "auto",
                      relative_dt           = 1e-3,
                      Ntrials               = 1,
                      Nbootstraps           = 0,
                      Ntrials_per_bootstrap = NULL,
                      Nthreads              = 1,
                      max_model_runtime     = NULL,
                      fit_control           = list(),
                      verbose               = FALSE,
                      verbose_prefix        = "")
```

Arguments

tree	An ultrametric timetree of class "phylo", representing the time-calibrated phylogeny of a set of extant species.
oldest_age	Strictly positive numeric, specifying the oldest time before present ("age") to consider when calculating the likelihood. If this is equal to or greater than the root age, then oldest_age is taken as the stem age, and the classical formula by Morlon et al. (2011) is used. If oldest_age is less than the root age, the tree is split into multiple subtrees at that age by treating every edge crossing that age as the stem of a subtree, and each subtree is considered an independent realization of the HBD model stemming at that age. This can be useful for avoiding points in the tree close to the root, where estimation uncertainty is generally higher. If oldest_age==NULL, it is automatically set to the root age.
age0	Non-negative numeric, specifying the youngest age (time before present) to consider for fitting. If age0>0, the tree essentially is trimmed at age0, omitting anything younger than age0, and the PSR is fitted to the trimmed tree while shifting time appropriately.

age_grid   Numeric vector, listing ages in ascending order at which the PSR is allowed to vary independently. This grid must cover at least the age range from age0 to oldest_age. If NULL or of length <=1 (regardless of value), then the PSR is assumed to be time-independent.

min_PSR   Numeric vector of length Ngrid (=max(1,length(age_grid))), or a single numeric, specifying lower bounds for the fitted PSR at each point in the age grid. If a single numeric, the same lower bound applies at all ages. Note that the PSR is never negative.

max_PSR   Numeric vector of length Ngrid, or a single numeric, specifying upper bounds for the fitted PSR at each point in the age grid. If a single numeric, the same upper bound applies at all ages. Use +Inf to omit upper bounds.

guess_PSR   Initial guess for the PSR at each age-grid point. Either NULL (an initial guess will be computed automatically), or a single numeric (guessing the same PSR at all ages) or a numeric vector of size Ngrid specifying a separate guess at each age-grid point. To omit an initial guess for some but not all age-grid points, set their guess values to NA. Guess values are ignored for non-fitted (i.e., fixed) parameters.

fixed_PSR   Optional fixed (i.e. non-fitted) PSR values on one or more age-grid points. Either NULL (PSR is not fixed anywhere), or a single numeric (PSR fixed to the same value at all grid points) or a numeric vector of size Ngrid (PSR fixed at one or more age-grid points, use NA for non-fixed values).

splines_degree Integer between 0 and 3 (inclusive), specifying the polynomial degree of the PSR between age-grid points. If 0, then the PSR is considered piecewise constant, if 1 then the PSR is considered piecewise linear, if 2 or 3 then the PSR is considered to be a spline of degree 2 or 3, respectively. The splines_degree influences the analytical properties of the curve, e.g. splines_degree==1 guarantees a continuous curve, splines_degree==2 guarantees a continuous curve and continuous derivative, and so on. A degree of 0 is generally not recommended.

condition   Character, either "crown", "stem" or "auto", specifying on what to condition the likelihood. If "crown", the likelihood is conditioned on the survival of the two daughter lineages branching off at the root. If "stem", the likelihood is conditioned on the survival of the stem lineage. Note that "crown" really only makes sense when oldest_age is equal to the root age, while "stem" is recommended if oldest_age differs from the root age. If "auto", the condition is chosen according to the recommendations mentioned earlier.

relative_dt   Strictly positive numeric (unitless), specifying the maximum relative time step allowed for integration over time, when calculating the likelihood. Smaller values increase integration accuracy but increase computation time. Typical values are 0.0001-0.001. The default is usually sufficient.

Ntrials   Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing Ntrials reduces the risk of reaching a non-global local maximum in the fitting objective.

Nbootstraps   Integer, specifying an optional number of bootstrap samplings to perform, for estimating standard errors and confidence intervals of maximum-likelihood fitted parameters. If 0, no bootstrapping is performed. Typical values are 10-100. At

each bootstrap sampling, a random timetree is generated under the birth-death model according to the fitted PSR, the parameters are estimated anew based on the generated tree, and subsequently compared to the original fitted parameters. Each bootstrap sampling will use roughly the same information and similar computational resources as the original maximum-likelihood fit (e.g., same number of trials, same optimization parameters, same initial guess, etc).

Ntrials_per_bootstrap

Integer, specifying the number of fitting trials to perform for each bootstrap sampling. If `NULL`, this is set equal to `max(1,Ntrials)`. Decreasing `Ntrials_per_bootstrap` will reduce computation time, at the expense of potentially inflating the estimated confidence intervals; in some cases (e.g., for very large trees) this may be useful if fitting takes a long time and confidence intervals are very narrow anyway. Only relevant if `Nbootstraps>0`.

Nthreads   Integer, specifying the number of parallel threads to use for performing multiple fitting trials simultaneously. This should generally not exceed the number of available CPUs on your machine. Parallel computing is not available on the Windows platform.

max_model_runtime

Optional numeric, specifying the maximum number of seconds to allow for each evaluation of the likelihood function. Use this to abort fitting trials leading to parameter regions where the likelihood takes a long time to evaluate (these are often unlikely parameter regions).

fit_control   Named list containing options for the `nlminb` optimization routine, such as `iter.max`, `eval.max` or `rel.tol`. For a complete list of options and default values see the documentation of `nlminb` in the `stats` package.

verbose   Logical, specifying whether to print progress reports and warnings to the screen. Note that errors always cause a return of the function (see return values `success` and `error`).

verbose_prefix   Character, specifying the line prefix for printing progress reports to the screen.

Details

It is generally advised to provide as much information to the function `fit_hbd_psr_on_grid` as possible, including reasonable lower and upper bounds (`min_PSR` and `max_PSR`) and a reasonable parameter guess (`guess_PSR`). It is also important that the `age_grid` is sufficiently fine to capture the expected major variations of the PSR over time, but keep in mind the serious risk of overfitting when `age_grid` is too fine and/or the tree is too small.

Value

A list with the following elements:

success   Logical, indicating whether model fitting succeeded. If `FALSE`, the returned list will include an additional "error" element (character) providing a description of the error; in that case all other return variables may be undefined.

objective_value

The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.

objective_name   The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".

loglikelihood   The log-likelihood of the fitted model for the given timetree.

fitted_PSR   Numeric vector of size Ngrid, listing fitted or fixed pulled speciation rates (PSR) at each age-grid point. Between grid points the fitted PSR should be interpreted as a piecewise polynomial function (natural spline) of degree splines_degree; to evaluate this function at arbitrary ages use the castor routine evaluate_spline.

guess_PSR   Numeric vector of size Ngrid, specifying the initial guess for the PSR at each age-grid point.

age_grid   The age-grid on which the PSR is defined. This will be the same as the provided age_grid, unless the latter was NULL or of length <=1.

NFP   Integer, number of fitted (i.e., non-fixed) parameters. If none of the PSRs were fixed, this will be equal to Ngrid.

AIC   The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters, and L is the maximized likelihood.

BIC   The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of branching times), and L is the maximized likelihood.

converged   Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase iter.max and/or eval.max).

Niterations   Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.

Nevaluations   Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.

bootstrap_estimates
   If Nbootstraps>0, this will be a numeric matrix of size Nbootstraps x Ngrid, listing the fitted PSR at each grid point and for each bootstrap.

standard_errors
   If Nbootstraps>0, this will be a numeric vector of size NGrid, listing bootstrap-estimated standard errors for the fitted PSR at each grid point.

CI50lower   If Nbootstraps>0, this will be a numeric vector of size Ngrid, listing bootstrap-estimated lower bounds of the 50-percent confidence intervals for the fitted PSR at each grid point.

CI50upper   Similar to CI50lower, listing upper bounds of 50-percentile confidence intervals.

CI95lower   Similar to CI50lower, listing lower bounds of 95-percentile confidence intervals.

CI95upper   Similar to CI95lower, listing upper bounds of 95-percentile confidence intervals.

Author(s)

Stilianos Louca

References

S. Louca et al. (2018). Bacterial diversification through geological time. Nature Ecology & Evolution. 2:1458-1467.

See Also

[simulate_deterministic_hbd](#)

[loglikelihood_hbd](#)

[fit_hbd_model_parametric](#)

[fit_hbd_model_on_grid](#)

[fit_hbd_pdr_parametric](#)

[fit_hbd_pdr_on_grid](#)

Examples

```
## Not run:
# Generate a random tree with exponentially varying lambda & mu
Ntips     = 10000
rho       = 0.5 # sampling fraction
time_grid = seq(from=0, to=100, by=0.01)
lambdas   = 2*exp(0.1*time_grid)
mus       = 1.5*exp(0.09*time_grid)
sim       = generate_random_tree( parameters  = list(rarefaction=rho),
                                  max_tips    = Ntips/rho,
                                  coalescent  = TRUE,
                                  added_rates_times    = time_grid,
                                  added_birth_rates_pc = lambdas,
                                  added_death_rates_pc = mus)
tree = sim$tree
root_age = castor::get_tree_span(tree)$max_distance
cat(sprintf("Tree has %d tips, spans %g Myr\n",length(tree$tip.label),root_age))

# Fit PSR on grid
oldest_age=root_age/2 # only consider recent times when fitting
Ngrid     = 10
age_grid  = seq(from=0,to=oldest_age,length.out=Ngrid)
fit = fit_hbd_psr_on_grid(tree,
        oldest_age  = oldest_age,
        age_grid    = age_grid,
        min_PSR     = 0,
        max_PSR     = +100,
        condition   = "crown",
        Ntrials     = 10,# perform 10 fitting trials
        Nthreads    = 10,# use two CPUs
        max_model_runtime = 1)  # limit model evaluation to 1 second
if(!fit$success){
```

```
      cat(sprintf("ERROR: Fitting failed: %s\n",fit$error))
   }else{
      cat(sprintf("Fitting succeeded:\nLoglikelihood=%g\n",fit$loglikelihood))
      # plot fitted PSR
      plot( x    = fit$age_grid,
            y    = fit$fitted_PSR,
            main = 'Fitted PSR',
            xlab = 'age',
            ylab = 'PSR',
            type = 'b',
            xlim = c(root_age,0))

      # plot deterministic LTT of fitted model
      plot( x    = fit$age_grid,
            y    = fit$fitted_LTT,
            main = 'Fitted dLTT',
            xlab = 'age',
            ylab = 'lineages',
            type = 'b',
            log  = 'y',
            xlim = c(root_age,0))

      # get fitted PSR as a function of age
      PSR_fun = approxfun(x=fit$age_grid, y=fit$fitted_PSR)
   }

   ## End(Not run)
```

---

fit_mk   *Fit a Markov (Mk) model for discrete trait evolution.*

Description

Estimate the transition rate matrix of a continuous-time Markov model for discrete trait evolution
("Mk model") via maximum-likelihood, based on one or more phylogenetic trees and its tips' states.

Usage

```
fit_mk( trees,
        Nstates,
        tip_states              = NULL,
        tip_priors              = NULL,
        rate_model              = "ER",
        root_prior              = "auto",
        oldest_ages             = NULL,
        guess_transition_matrix = NULL,
        Ntrials                 = 1,
        max_model_runtime       = NULL,
        optim_algorithm         = "nlminb",
```

```
           optim_max_iterations    = 200,
           optim_rel_tol           = 1e-8,
           check_input             = TRUE,
           Nthreads                = 1,
           Nbootstraps             = 0,
           Ntrials_per_bootstrap   = NULL,
           verbose                 = FALSE,
           verbose_prefix          = "")
```

Arguments

trees
: Either a single phylogenetic tree of class "phylo", or a list of phylogenetic trees. Edge lengths should correspond (or be analogous) to time. The trees don't need to be ultrametric.

Nstates
: Integer, specifying the number of possible discrete states that the trait can have.

tip_states
: Either an integer vector of size Ntips (only permitted if trees[] is a single tree) or a list containing Ntrees such integer vectors (if trees[] is a list of trees), listing the state of each tip in each tree. Note that tip_states cannot include NAs or NaNs; if the states of some tips are uncertain, you should use the option tip_priors instead. Can also be NULL, in which case tip_priors must be provided.

tip_priors
: Either a numeric matrix of size Ntips x Nstates (only permitted if trees[] is a single tree), or a list containing Ntrees such matrixes (if trees[] is a list of trees), listing the likelihood of each state at each tip in each tree. Can also be NULL, in which case tip_states must be provided. Hence, tip_priors[t][i,s] is the likelihood of the observed state of tip i in tree t, if the tip's true state was in state s. For example, if you know for certain that a tip is in state k, then set tip_priors[t][i,s]=1 for s=k and tip_priors[t][i,s]=0 for all other s.

rate_model
: Rate model to be used for the transition rate matrix. Can be "ER" (all rates equal), "SYM" (transition rate i–>j is equal to transition rate j–>i), "ARD" (all rates can be different), "SUEDE" (only stepwise transitions i–>i+1 and i–>i-1 allowed, all 'up' transitions are equal, all 'down' transitions are equal) or "SRD" (only stepwise transitions i–>i+1 and i–>i-1 allowed, and each rate can be different). Can also be an index matrix that maps entries of the transition matrix to the corresponding independent rate parameter to be fitted. Diagonal entries should map to 0, since diagonal entries are not treated as independent rate parameters but are calculated from the remaining entries in the transition rate matrix. All other entries that map to 0 represent a transition rate of zero. The format of this index matrix is similar to the format used by the ace function in the ape package. rate_model is only relevant if transition_matrix==NULL.

root_prior
: Prior probability distribution of the root's states, used to calculate the model's overall likelihood from the root's marginal ancestral state likelihoods. Can be "flat" (all states equal), "empirical" (empirical probability distribution of states across the tree's tips), "stationary" (stationary probability distribution of the transition matrix), "likelihoods" (use the root's state likelihoods as prior), "max_likelihood" (put all weight onto the state with maximum likelihood) or "auto" (will be chosen automatically based on some internal logic). If

                  "stationary" and transition_matrix==NULL, then a transition matrix is first fitted using a flat root prior, and then used to calculate the stationary distribution. root_prior can also be a non-negative numeric vector of size Nstates and with total sum equal to 1.

oldest_ages Optional numeric or numeric vector of size Ntrees, specifying the oldest age (time before present) for each tree to consider when fitting the Mk model. If NULL, the entire trees are considered from the present all the way to their root. If non-NULL, then each tree is "cut" at the corresponding oldest age, yielding multiple subtrees, each of which is assumed to be an independent realization of the Mk process. If oldest_ages is a single numeric, then all trees are cut at the same oldest age. This option may be useful if temporal variation is suspected in the Mk rates, and only data near the present are to be used for fitting to avoid violating the assumptions of a constant-rates Mk model.

guess_transition_matrix

                  Optional 2D numeric matrix, specifying a reasonable first guess for the transition rate matrix. May contain NA. May also be NULL, in which case a reasonable first guess is automatically generated.

Ntrials Number of trials (starting points) for fitting the transition rate matrix. A higher number may reduce the risk of landing in a local non-global optimum of the likelihood function, but will increase computation time during fitting.

max_model_runtime

                  Optional positive numeric, specifying the maximum time (in seconds) allowed for a single evaluation of the likelihood function. If a specific Mk model takes longer than this threshold to evaluate, then its likelihood is set to -Inf. This option can be used to avoid badly parameterized models during fitting and can thus reduce fitting time. If NULL or <=0, this option is ignored.

optim_algorithm

                  Either "optim" or "nlminb", specifying which optimization algorithm to use for maximum-likelihood estimation of the transition matrix.

optim_max_iterations

                  Maximum number of iterations (per fitting trial) allowed for optimizing the likelihood function.

optim_rel_tol Relative tolerance (stop criterion) for optimizing the likelihood function.

check_input Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Nthreads Number of parallel threads to use for running multiple fitting trials simultaneously. This only makes sense if your computer has multiple cores/CPUs and if Ntrials>1. This option is ignored on Windows, because Windows does not support forking.

Nbootstraps Integer, specifying the number of parametric bootstraps to perform for estimating standard errors and confidence intervals of estimated rate parameters. Set to 0 for no bootstrapping.

Ntrials_per_bootstrap

                  Integer, specifying the number of fitting trials to perform for each bootstrap sampling. If NULL, this is set equal to max(1,Ntrials). Decreasing Ntrials_per_bootstrap

will reduce computation time, at the expense of potentially inflating the estimated confidence intervals; in some cases (e.g., for very large trees) this may be useful if fitting takes a long time and confidence intervals are very narrow anyway. Only relevant if Nbootstraps>0.

verbose   Logical, specifying whether to print progress reports and warnings to the screen.

verbose_prefix Character, specifying the line prefix for printing progress reports to the screen.

Details

The trait's states must be represented by integers within 1,..,Nstates, where Nstates is the total number of possible states. If the states are originally in some other format (e.g. characters or factors), you should map them to a set of integers 1,..,Nstates. The order of states (if relevant) should be reflected in their integer representation. For example, if your original states are "small", "medium" and "large" and rate_model=="SUEDE", it is advised to represent these states as integers 1,2,3. You can easily map any set of discrete states to integers using the function [map_to_state_space](#).

This function allows the specification of the precise tip states (if these are known) using the vector tip_states. Alternatively, if some tip states are not fully known, you can pass the state likelihoods using the matrix tip_priors. Note that exactly one of the two arguments, tip_states or tip_priors, must be non-NULL.

Tips must be represented in tip_states or tip_priors in the same order as in tree$tip.label. None of the input vectors or matrixes need include row or column names; if they do, however, they are checked for consistency (if check_input==TRUE).

The tree is either assumed to be complete (i.e. include all possible species), or to represent a random subset of species chosen independently of their states. If the tree is not complete and tips are not chosen independently of their states, then this method will not be valid.

fit_Mk uses maximum-likelihood to estimate each free parameter of the transition rate matrix. The number of free parameters depends on the rate_model considered; for example, ER implies a single free parameter, while ARD implies Nstates x (Nstates-1) free parameters. If multiple trees are provided as input, the likelihood is the product of likelihoods for each tree, i.e. as if each tree was an independent realization of the same Markov process.

This function is similar to [asr_mk_model](#), but focused solely on fitting the transition rate matrix (i.e., without estimating ancestral states) and with the ability to utilize multiple trees at once.

Value

A named list with the following elements:

success   Logical, indicating whether the fitting was successful. If FALSE, an additional element error (of type character) is included containing an explanation of the error; in that case the value of any of the other elements is undetermined.

Nstates   Integer, the number of states assumed for the model.

transition_matrix

                A matrix of size Nstates x Nstates, the fitted transition rate matrix of the model. The [r,c]-th entry is the transition rate from state r to state c.

loglikelihood   Numeric, the log-likelihood of the observed tip states under the fitted model.

Niterations	Integer, the number of iterations required to reach the maximum log-likelihood. Depending on the optimization algorithm used (see optim_algorithm), this may be NA.
Nevaluations	Integer, the number of evaluations of the likelihood function required to reach the maximum log-likelihood. Depending on the optimization algorithm used (see optim_algorithm), this may be NA.
converged	Logical, indicating whether the fitting algorithm converged. Note that fit_Mk may return successfully even if convergence was not achieved; if this happens, the fitted transition matrix may not be reasonable. In that case it is recommended to change the optimization options, for example increasing optim_max_iterations.
guess_rate	Numeric, the initial guess used for the average transition rate, prior to fitting.
AIC	Numeric, the Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of independent fitted parameters and L is the maximized likelihood.
standard_errors	
	Numeric matrix of size Nstates x Nstates, estimated standard error of the fitted transition rates, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI50lower	Numeric matrix of size Nstates x Nstates, lower bounds of the 50% confidence intervals (25-75% percentile) for the fitted transition rates, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI50upper	Numeric matrix of size Nstates x Nstates, upper bounds of the 50% confidence intervals for the fitted transition rates, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI95lower	Numeric matrix of size Nstates x Nstates, lower bounds of the 95% confidence intervals (2.5-97.5% percentile) for the fitted transition rates, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI95upper	Numeric matrix of size Nstates x Nstates, upper bounds of the 95% confidence intervals for the fitted transition rates, based on parametric bootstrapping. Only returned if Nbootstraps>0.

Author(s)

Stilianos Louca

References

Z. Yang, S. Kumar and M. Nei (1995). A new method for inference of ancestral nucleotide and amino acid sequences. Genetics. 141:1641-1650.

M. Pagel (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London B: Biological Sciences. 255:37-45.

See Also

asr_mk_model, simulate_mk_model, fit_musse

Examples

```
## Not run:
# generate random tree
Ntips = 1000
tree  = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# create random transition matrix
Nstates = 5
Q = get_random_mk_transition_matrix(Nstates, rate_model="ER", max_rate=0.01)
cat(sprintf("Simulated ER transition rate=%g\n",Q[1,2]))

# simulate the trait's evolution
simulation = simulate_mk_model(tree, Q)
tip_states = simulation$tip_states

# fit Mk transition matrix
results = fit_mk(tree, Nstates, tip_states, rate_model="ER", Ntrials=2)

# print Mk model fitting summary
cat(sprintf("Mk model: log-likelihood=%g\n",results$loglikelihood))
cat(sprintf("Fitted ER transition rate=%g\n",results$transition_matrix[1,2]))

## End(Not run)
```

---

fit_musse   *Fit a discrete-state-dependent diversification model via maximum-*
                                likelihood.

Description

The Binary State Speciation and Extinction (BiSSE) model (Maddison et al. 2007) and its extension
to Multiple State Speciation Extinction (MuSSE) models (FitzJohn et al. 2009, 2012), Hidden State
Speciation Extinction (HiSSE) models (Beaulieu and O'meara, 2016) or Several Examined and
Concealed States-dependent Speciation and Extinction (SecSSE) models (van Els et al. 2018),
describe a Poissonian cladogenic process whose birth/death (speciation/extinction) rates depend on
the states of an evolving discrete trait. Specifically, extant tips either go extinct or split continuously
in time at Poissonian rates, and birth/death rates at each extant tip depend on the current state of the
tip; lineages tansition stochastically between states acccording to a continuous-time Markov process
with fixed transition rates. At the end of the simulation (i.e., at "present-day"), extant lineages are
sampled according to some state-dependent probability ("sampling_fraction"), which may depend
on proxy state. Optionally, tips may also be sampled continuously over time according to some
Poissonian rate (which may depend on proxy state), in which case the resulting tree may not be
ultrametric.

This function takes as main input a phylogenetic tree (ultrametric unless Poissonian sampling is
included) and a list of tip proxy states, and fits the parameters of a BiSSE/MuSSE/HiSSE/SecSSE
model to the data via maximum-likelihood. Tips can have missing (unknown) proxy states, and the
function can account for biases in species sampling and biases in the identification of proxy states.
The likelihood is calculated using a mathematically equivalent, but computationally more efficient

variant, of the classical postorder-traversal BiSSE/MuSSE/HiSSE/SecSSE algorithm, as described by Louca (2019). This function has been optimized for large phylogenetic trees, with a relatively small number of states (i.e. Nstates«Ntips); its time complexity scales roughly linearly with Ntips.

Usage

```
fit_musse(tree,
          Nstates,
          NPstates              = NULL,
          proxy_map             = NULL,
          state_names           = NULL,
          tip_pstates           = NULL,
          tip_priors            = NULL,
          sampling_fractions    = 1,
          reveal_fractions      = 1,
          sampling_rates        = 0,
          transition_rate_model = "ARD",
          birth_rate_model      = "ARD",
          death_rate_model      = "ARD",
          transition_matrix     = NULL,
          birth_rates           = NULL,
          death_rates           = NULL,
          first_guess           = NULL,
          lower                 = NULL,
          upper                 = NULL,
          root_prior            = "auto",
          root_conditioning     = "auto",
          oldest_age            = NULL,
          Ntrials               = 1,
          max_start_attempts    = 10,
          optim_algorithm       = "subplex",
          optim_max_iterations  = 10000,
          optim_max_evaluations = NULL,
          optim_rel_tol         = 1e-6,
          check_input           = TRUE,
          include_ancestral_likelihoods = FALSE,
          Nthreads              = 1,
          Nbootstraps           = 0,
          Ntrials_per_bootstrap = NULL,
          max_condition_number  = 1e4,
          relative_ODE_step     = 0.1,
          E_value_step          = 1e-4,
          D_temporal_resolution = 100,
          max_model_runtime     = NULL,
          verbose               = TRUE,
          diagnostics           = FALSE,
          verbose_prefix        = "")
```

Arguments

tree	Phylogenetic tree of class "phylo", representing the evolutionary relationships between sampled species/lineages. Unless Poissonian sampling is included in the model (option `sampling_rates`), this tree should be ultrametric.
Nstates	Integer, specifying the number of possible discrete states a tip can have, influencing speciation/extinction rates. For example, if Nstates==2 then this corresponds to the common Binary State Speciation and Extinction (BiSSE) model (Maddison et al., 2007). In the case of a HiSSE/SecSSE model, Nstates refers to the total number of diversification rate categories. For example, in the case of the HiSSE model described by Beaulieu and O'meara (2016), Nstates=4.
NPstates	Integer, optionally specifying a number of "proxy-states" that are observed instead of the underlying speciation/extinction-modulating states. To fit a HiSSE/SecSSE model, NPstates should be smaller than Nstates. Each state corresponds to a different proxy-state, as defined using the variable `proxy_map` (see below). For BiSSE/MuSSE with no hidden states, NPstates can be set to either NULL or equal to Nstates; in either case, NPstates will be considered equal to Nstates. For example, in the case of the HiSSE model described by Beaulieu and O'meara (2016), NPstates=2.
proxy_map	Integer vector of size Nstates and with values in 1,..NPstates, specifying the correspondence between states (i.e. diversification-rate categories) and proxy-states, in a HiSSE/SecSSE model. Specifically, `proxy_map[s]` indicates which proxy-state the state s is represented by. Each proxy-state can represent multiple states (i.e. proxies are ambiguous), but each state must be represented by exactly one proxy-state. For example, to setup the HiSSE model described by Beaulieu and O'meara (2016), use `proxy_map=c(1,2,1,2)`. For non-HiSSE models, set this to NULL or to `c(1:Nstates)`. See below for more details.
state_names	Optional character vector of size Nstates, specifying a name/description for each state. This does not influence any of the calculations. It is merely used to add human-readable row/column names (rather than integers) to the returned vectors/matrices. If NULL, no row/column names are added.
tip_pstates	Integer vector of size Ntips, listing the proxy state at each tip, in the same order as tips are indexed in the tree. The vector may (but need not) include names; if it does, these are checked for consistency with the tree (if check_input==TRUE). Values must range from 1 to NPstates (which is assumed equal to Nstates in the case of BiSSE/MuSSE). States may also be NA, corresponding to unknown tip proxy states (no information available).
tip_priors	Numeric matrix of size Ntips x Nstates (or of size Ntips x NPstates), listing prior likelihoods of each state (or each proxy-state) at each tip. Can be provided as an alternative to tip_pstates. Thus, `tip_priors[i,s]` is the likelihood of observing the data if the tip i was at state s (or proxy-state s). Either tip_pstates or tip_priors must be non-NULL, but not both.
sampling_fractions	
	Numeric vector of size NPstates, with values between 0 and 1, listing the sampling fractions of extant species depending on proxy-state. That is, `sampling_fractions[p]` is the probability that an extant species, having proxy state p, is included in the phylogeny at present-day. If all extant species are included in the tree with the

same probability (i.e., independent of state), this can also be a single number. If NULL (default), all extant species are assumed to be included in the tree.

reveal_fractions

Numeric vector of size NPstates, with values between 0 and 1, listing the probabilities of proxy-state identification depending on proxy-state. That is, reveal_fractions[p] is the probability that a species with proxy-state p will have a known ("revealed") state, conditional upon being included in the tree. This can be used to incorporate reveal biases for tips, depending on their proxy state. Can also be NULL or a single number (in which case reveal fractions are assumed to be independent of proxy-state). Note that only the relative values in reveal_fractions matter, for example c(1,2,1) has the same effect as c(0.5,1,0.5), because reveal_fractions is normalized internally anyway.

sampling_rates   Numeric vector of size NPstates, listing Poissonian per-lineage sampling rates over time. Hence, sampling_rates[p] is the rate at which lineages are sampled over time when they are in proxy state p. Can also be a single numeric, in which case sampling rates are the same for all proxy states. If NULL, Poissonian sampling is assumed to not occur. Note that earlier MuSSE/HiSSE models (e.g., by Beaulieu and O'Meara, 2016) do not include Poissonian sampling (i.e., all tips are assumed to have been sampled at present-day). Poissonian sampling through time is common in epidemiological models but uncommon in macroevolution models.

transition_rate_model

Either a character or a 2D integer matrix of size Nstates x Nstates, specifying the model for the transition rates between states. This option controls the parametric complexity of the state transition model, i.e. the number of independent rates and the correspondence between independent and dependent rates. If a character, then it must be one of "ER", "SYM", "ARD", "SUEDE" or "SRD", as used for Mk models (see the function [asr_mk_model](#) for details). For example, "ARD" (all rates different) specifies that all transition rates should be considered as independent parameters with potentially different values.

If an integer matrix, then it defines a custom parametric structure for the transition rates, by mapping entries of the transition matrix to a set of independent transition-rate parameters (numbered 1,2, and so on), similarly to the option rate_model in the function [asr_mk_model](#), and as returned for example by the function [get_transition_index_matrix](#). Entries must be between 1 and Nstates, however 0 may also be used to denote a fixed value of zero. For example, if transition_rate_model[1,2]=transition_rate_model[2,1], then the transition rates 1->2 and 2->1 are assumed to be equal. Entries on the diagonal are ignored, since the diagonal elements are always adjusted to ensure a valid Markov transition matrix. To construct a custom matrix with the proper structure, it may be convenient to first generate an "ARD" matrix using [get_transition_index_matrix](#), and then modify individual entries to reduce the number of independent rates.

birth_rate_model

Either a character or an integer vector of length Nstates, specifying the model for the various birth (speciation) rates. This option controls the parametric complexity of the possible birth rates, i.e. the number of independent birth rates and the correspondence between independent and dependent birth rates. If a char-

acter, then it must be either "ER" (equal rates) or "ARD" (all rates different). If an integer vector, it must map each state to an indepedent birth-rate parameter (indexed 1,2,..). For example, the vector `c(1,2,1)` specifies that the birth-rates λ_1 and λ_3 must be the same, but λ_2 is independent.

death_rate_model

    Either a character or an integer vector of length Nstates, specifying the model for the various death (extinction) rates. Similar to `birth_rate_model`.

transition_matrix

    Either `NULL` or a 2D matrix of size Nstates x Nstates, specifying known (and thus fixed) transition rates between states. For example, setting some elements to 0 specifies that these transitions cannot occur directly. May also contain `NA`, indicating rates that are to be fitted. If `NULL` or empty, all rates are considered unknown and are therefore fitted. Note that, unless `transition_rate_model=="ARD"`, values in `transition_matrix` are assumed to be consistent with the rate model, that is, rates specified to be equal under the transition rate model are expected to also have equal values in `transition_matrix`.

birth_rates   Either `NULL`, or a single number, or a numeric vector of length Nstates, specifying known (and thus fixed) birth rates for each state. May contain `NA`, indicating rates that are to be fitted. For example, the vector `c(5,0,NA)` specifies that $\lambda_1 = 5$, $\lambda_2 = 0$ and that λ_3 is to be fitted. If `NULL` or empty, all birth rates are considered unknown and are therefore fitted. If a single number, all birth rates are considered fixed at that given value.

death_rates   Either `NULL`, or a single number, or a numeric vector of length Nstates, specifying known (and thus fixed) death rates for each state. Similar to `birth_rates`.

first_guess   Either `NULL`, or a named list containing optional initial suggestions for various model parameters, i.e. start values for fitting. The list can contain any or all of the following elements:

- `transition_matrix`: A single number or a 2D numeric matrix of size Nstates x Nstates, specifying suggested start values for the transition rates. May contain NA, indicating rates that should be guessed automatically by the function. If a single number, then that value is used as a start value for all transition rates.
- `birth_rates`: A single number or a numeric vector of size Nstates, specifying suggested start values for the birth rates. May contain NA, indicating rates that should be guessed automatically by the function (by fitting a simple birth-death model, see `fit_tree_model`).
- `death_rates`: A single number or a numeric vector of size Nstates, specifying suggested start values for the death rates. May contain NA, indicating rates that should be guessed automatically by the function (by fitting a simple birth-death model, see `fit_tree_model`).

Start values are only relevant for fitted (i.e., non-fixed) parameters.

lower   Either `NULL` or a named list containing optional lower bounds for various model parameters. The list can contain any or all of the elements `transition_matrix`, `birth_rates` and `death_rates`, structured similarly to `first_guess`. For example, `list(transition_matrix=0.1,birth_rates=c(5,NA,NA))` specifies that all transition rates between states must be 0.1 or greater, that the birth rate

λ_1 must be 5 or greater, and that all other model parameters have unspecified lower bound. For parameters with unspecified lower bounds, zero is used as a lower bound. Lower bounds only apply to fitted (i.e., non-fixed) parameters.

upper   Either NULL or a named list containing optional upper bounds for various model parameters. The list can contain any or all of the elements `transition_matrix`, `birth_rates` and `death_rates`, structured similarly to upper. For example, `list(transition_matrix=2,birth_rates=c(10,NA,NA))` specifies that all transition rates between states must be 2 or less, that the birth rate λ_1 must be 10 or less, and that all other model parameters have unspecified upper bound. For parameters with unspecified upper bounds, infinity is used as an upper bound. Upper bounds only apply to fitted (i.e., non-fixed) parameters.

root_prior   Either a character or a numeric vector of size Nstates, specifying the prior probabilities of states for the root, i.e. the weights for obtaining a single model likelihood by averaging the root's state likelihoods. If a character, then it must be one of "flat", "empirical", "likelihoods", "max_likelihood" or "auto". "empirical" means the root's prior is set to the proportions of (estimated) extant species in each state (correcting for sampling fractions and reveal fractions, if applicable). "likelihoods" means that the computed state-likelihoods of the root are used, after normalizing to obtain a probability distribution; this is the approach used in the package `hisse::hisse` v1.8.9 under the option `root.p=NULL`, and the approach in the package `diversitree::find.mle` v0.9-10 under the option `root=ROOT.OBS`. If "max_likelihood", then the root's prior is set to a Dirac distribution, with full weight given to the maximum-likelihood state at the root (after applying the conditioning). If a numeric vector, `root_prior` specifies custom probabilities (weights) for each state. Note that if `root_conditioning` is "madfitz" or "herr_als" (see below), then the prior is set before the conditioning and not updated afterwards for consistency with other R packages.

root_conditioning

     Character, specifying an optional modification to be applied to the root's state likelihoods prior to averaging. Can be "none" (no modification), "madfitz", "herr_als", "crown" or "stem". "madfitz" and "herr_als" (after van Els, Etiene and Herrera-Alsina 2018) are the options implemented in the package `hisse` v1.8.9, conditioning the root's state-likelihoods based on the birth-rates and the computed extinction probability (after or before averaging, respectively). See van Els (2018) for a comparison between "madfitz" and "herr_als". The option "stem" conditions the state likelihoods on the probability that the stem lineage would survive until the present. The option "crown" conditions the state likelihoods on the probability that a split occurred at `oldest_age` and that the two child lineages survived until the present; this option is only recommended if `oldest_age` is equal to the root age.

oldest_age   Strictly positive numeric, specifying the oldest age (time before present) to consider for fitting. If this is smaller than the tree's root age, then the tree is split into multiple subtrees at `oldest_age`, and each subtree is considered as an independent realization of the same diversification/evolution process whose parameters are to be estimated. The `root_conditioning` and `root_prior` are applied separately to each subtree, prior to calculating the joint (product) likelihood of all subtrees. This option can be used to restrict the fitting to a small (recent) time interval, during which the MuSSE/BiSSE assumptions (e.g., time-independent

speciation/extinction/transition rates) are more likely to hold. If `oldest_age` is `NULL`, it is automatically set to the root age. In principle `oldest_age` may also be older than the root age.

Ntrials   Non-negative integer, specifying the number of trials for fitting the model, using alternative (randomized) starting parameters at each trial. A larger `Ntrials` reduces the risk of landing on a local non-global optimum of the likelihood function, and thus increases the chances of finding the truly best fit. If 0, then no fitting is performed, and only the first-guess (i.e., provided or guessed start params) is evaluated and returned. Hence, setting `Ntrials=0` can be used to obtain a reasonable set of start parameters for subsequent fitting or for Markov Chain Monte Carlo.

max_start_attempts
                    Integer, specifying the number of times to attempt finding a valid start point (per trial) before giving up on that trial. Randomly chosen extreme start parameters may occasionally result in Inf/undefined likelihoods, so this option allows the algorithm to keep looking for valid starting points. For complex models (e.g., with >5 states), setting this to 10, 100 or even 1000 may be beneficial.

optim_algorithm
                    Character, specifying the optimization algorithm for fitting. Must be one of either "optim", "nlminb" or "subplex".

optim_max_iterations
                    Integer, maximum number of iterations allowed for fitting. Only relevant for "optim" and "nlminb".

optim_max_evaluations
                    Integer, maximum number of function evaluations allowed for fitting. Only relevant for "nlminb" and "subplex".

optim_rel_tol   Numeric, relative tolerance for the fitted log-likelihood.

check_input   Logical, specifying whether to check the validity of input variables. If you are certain that all input variables are valid, you can set this to `FALSE` to reduce computation.

include_ancestral_likelihoods
                    Logical, specifying whether to include the state likelihoods for each node, in the returned variables. These are the "D" variables calculated as part of the likelihood based on the subtree descending from each node, and may be used for "local" ancestral state reconstructions.

Nthreads   Integer, specifying the number of threads for running multiple fitting trials in parallel. Only relevant if `Ntrials>1`. Should generally not exceed the number of CPU cores on a machine. Must be a least 1.

Nbootstraps   Integer, specifying an optional number of bootstrap samplings to perform, for estimating standard errors and confidence intervals of maximum-likelihood fitted parameters. If 0, no bootstrapping is performed. Typical values are 10-100. At each bootstrap sampling, a simulation of the fitted MuSSE/HiSSE model is performed, the parameters are estimated anew based on the simulation, and subsequently compared to the original fitted parameters. Each bootstrap sampling will thus use roughly as many computational resources as the original maximum-likelihood fit (e.g., same number of trials, same optimization parameters etc).

Ntrials_per_bootstrap

> Integer, specifying the number of fitting trials to perform for each bootstrap sampling. If NULL, this is set equal to max(1,Ntrials). Decreasing Ntrials_per_bootstrap will reduce computation time, at the expense of potentially inflating the estimated confidence intervals; in some cases (e.g., for very large trees) this may be useful if fitting takes a long time and confidence intervals are very narrow anyway. Only relevant if Nbootstraps>0.

max_condition_number

> Positive unitless number, specifying the maximum permissible condition number for the "G" matrix computed for the log-likelihood. A higher condition number leads to faster computation (roughly on a log-scale) especially for large trees, at the potential expense of lower accuracy. Typical values are 1e2-1e5. See Louca (2019) for further details on the condition number of the G matrix.

relative_ODE_step

> Positive unitless number, specifying the default relative time step for the ordinary differential equation solvers.

E_value_step   Positive unitless number, specifying the relative difference between subsequent recorded and interpolated E-values, in the ODE solver for the extinction probabilities E (Louca 2019). Typical values are 1e-2 to 1e-5. A smaller E_value_step increases interpolation accuracy, but also increases memory requirements and adds runtime (scaling with the tree's age span, not Ntips).

D_temporal_resolution

> Positive unitless number, specifying the relative resolution for interpolating G-map over time (Louca 2019). This is relative to the typical time scales at which G-map varies. For example, a resolution of 10 means that within a typical time scale there will be 10 interpolation points. Typical values are 1-1000. A greater resolution increases interpolation accuracy, but also increases memory requirements and adds runtime (scaling with the tree's age span, not Ntips).

max_model_runtime

> Numeric, optional maximum number of seconds for evaluating the likelihood of a model, prior to cancelling the calculation and returning Inf. This may be useful if extreme model parameters (e.g., reached transiently during fitting) require excessive calculation time. Parameters for which the calculation of the likelihood exceed this threshold, will be considered invalid and thus avoided during fitting. For example, for trees with 1000 tips a time limit of 10 seconds may be reasonable. If 0, no time limit is imposed.

verbose   Logical, specifying whether to print progress reports and warnings to the screen. In any case, fatal errors are always reported.

diagnostics   Logical, specifying whether to print detailed information (such as model likelihoods) at every iteration of the fitting routine. For debugging purposes mainly.

verbose_prefix   Character, specifying the line prefix for printing progress reports, warnings and errors to the screen.

Details

HiSSE/SecSSE models include two discrete traits, one trait that defines the rate categories of diversification rates (as in BiSSE/MuSSE), and one trait that does not itself influence diversification but

whose states (here called "proxy states") each represent one or more of the diversity-modulating states. HiSSE models (Beaulieu and O'meara, 2016) and SecSSE models (van Els et al., 2018) are closely related to BiSSE/MuSSE models, the main difference being the fact that the actual diversification-modulating states are not directly observed. In essence, a HiSSE/SecSSE model is a BiSSE/MuSSE model, where the final tip states are replaced by their proxy states, thus "masking" the underlying diversity-modulating trait. This function is able to fit HiSSE/SecSSE models with appropriate choice of the input variables Nstates, NPstates and proxy_map. Note that the terminology and setup of HiSSE/SecSSE models followed here differs from their description in the original papers by Beaulieu and O'meara (2016) and van Els et al. (2018), in order to achieve what we think is a more intuitive unification of BiSSE/MuSSE/HiSSE/SecSSE. For ease of terminology, when considering a BiSSE/MuSSE model, here we use the terms "states" and "proxy-states" interchangeably, since under BiSSE/MuSSE the proxy trait can be considered identical to the diversification-modulating trait. A distinction between "states" and "proxy-states" is only relevant for HiSSE/SecSSE models.

As an example of a HiSSE model, Nstates=4, NPstates=2 and proxy_map=c(1,2,1,2) specifies that states 1 and 3 are represented by proxy-state 1, and states 2 and 4 are represented by proxy-state 2. This is the original case described by Beaulieu and O'Meara (2016); in their terminology, there would be 2 "hidden"" states ("0" and "1") and 2 "observed" states ("A" and "B"), and the 4 diversification rate categories (Nstates=4) would be called "0A", "1A", "0B" and "1B". The somewhat different terminology used here allows for easier generalization to an arbitrary number of diversification-modulating states and an arbitrary number of proxy states. For example, if there are 6 diversification modulating states, represented by 3 proxy-states as 1->A, 2->A, 3->B, 4->C, 5->C, 6->C, then one would set Nstates=6, NPstates=3 and proxy_map=c(1,1,2,3,3,3).

The run time of this function scales asymptotically linearly with tree size (Ntips), although run times can vary substantially depending on model parameters. As a rule of thumb, the higher the birth/death/transition rates are compared to the tree's overall time span, the slower the calculation becomes.

The following arguments control the tradeoff between accuracy and computational efficiency:

- max_condition_number: A smaller value means greater accuracy, at longer runtime and more memory.
- relative_ODE_step: A smaller value means greater accuracy, at longer runtime.
- E_value_step: A smaller value means greater accuracy, at longer runtime and more memory.
- D_temporal_resolution: A greater value means greater accuracy, at longer runtime and more memory.

Typically, the default values for these arguments should be fine. For smaller trees, where cladogenic and sampling stochasticity is the main source of uncertainty, these parameters can probably be made less stringent (i.e., leading to lower accuracy and faster computation), but then again for small trees computational efficiency may not be an issue anyway.

Value

A named list with the following elements:

success   Logical, indicating whether the fitting was successful. If FALSE, an additional element error (of type character) is included containing an explanation of the error; in that case the value of any of the other elements is undetermined.

Nstates	Integer, the number of states assumed for the model.
NPstates	Integer, the number of proxy states assumed for the model. Note that in the case of a BiSSE/MuSSE model, this will be the same as Nstates.
root_prior	Character, or numeric vector of length Nstates, specifying the root prior used.
parameters	Named list containing the final maximum-likelihood fitted model parameters. If Ntrials>1, then this contains the fitted parameters yielding the highest likelihood. Will contain the following elements:

- transition_matrix: 2D numeric matrix of size Nstates x Nstates, listing the fitted transition rates between states.
- birth_rates: Numeric vector of length Nstates, listing the fitted state-dependent birth rates.
- death_rates: Numeric vector of length Nstates, listing the fitted state-dependent death rates.

start_parameters	
	Named list containing the default start parameter values for the fitting. Structured similarly to parameters. Note that if Ntrials>1, only the first trial will have used these start values, all other trials will have used randomized start values. Will be defined even if Ntrials==0, and can thus be used to obtain a reasonable guess for the start parameters without actually fitting the model.
loglikelihood	Numeric, the maximized log-likelihood of the model, if fitting succeeded.
AIC	Numeric, the Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.
Niterations	The number of iterations needed for the best fit. Only relevant if the optimization method was "optim" or "nlminb".
Nevaluations	Integer, the number of function evaluations needed for the best fit. Only relevant if the optimization method was "nlminb" or "subplex".
converged	Logical, indicating whether convergence was successful during fitting. If convergence was not achieved, and the fitting was stopped due to one of the stopping criteria optim_max_iterations or optim_max_evaluations, the final likelihood will still be returned, but the fitted parameters may not be reasonable.
warnings	Character vector, listing any warnings encountered during evaluation of the likelihood function at the fitted parameter values. For example, this vector may contain warnings regarding the differential equation solvers or regarding the rank of the G-matrix (Louca, 2019).
subroots	Integer vector, listing indices of tips/nodes in the tree that were considered as starting points of independent MuSSE processes. If oldest_age was equal to or greater than the root age, then subroots will simply list the tree's root.
ML_subroot_states	
	Integer vector, with values between 1 and Nstates, giving the maximum-likelihood estimate of each subroot's state.
ML_substem_states	
	Integer vector, with values between 1 and Nstates, giving the maximum-likelihood estimate of the state at each subroot's stem (i.e., exactly at oldest_age).

trial_start_loglikelihoods

Numeric vector of length `Ntrials`, listing the initial loglikelihoods (i.e., at the starting parameter values) for each fitting trial.

trial_loglikelihoods

Numeric vector of length `Ntrials`, listing the maximized loglikelihoods for each fitting trial. These may be used for diagnosing the robustness of maximum-likelihood estimates and the assessing the needed for increasing `Ntrials`.

trial_Nstart_attempts

Integer vector of length `Ntrials`, listing the number of random start attempts for each trial (see option `max_start_attempts`).

trial_Niterations

Integer vector of length `Ntrials`, listing the number of iterations of each trial. Depending on the fitting algorithm used (option `optim_algorithm`), these may be NA (not available).

trial_Nevaluations

Integer vector of length `Ntrials`, listing the number of likelihood evaluations of each trial. Depending on the fitting algorithm used (option `optim_algorithm`), these may be NA (not available).

standard_errors

Named list containing the elements "transition_matrix" (numeric matrix of size Nstates x Nstates), "birth_rates" (numeric vector of size Nstates) and "death_rates" (numeric vector of size Nstates), listing standard errors of all model parameters estimated using parametric bootstrapping. Only included if Nbootstraps>0. Note that the standard errors of non-fitted (i.e., fixed) parameters will be zero.

CI50lower   Named list containing the elements "transition_matrix" (numeric matrix of size Nstates x Nstates), "birth_rates" (numeric vector of size Nstates) and "death_rates" (numeric vector of size Nstates), listing the lower end of the 50% confidence interval (i.e. the 25% quantile) for each model parameter, estimated using parametric bootstrapping. Only included if Nbootstraps>0.

CI50upper   Similar to `CI50lower`, but listing the upper end of the 50% confidence interval (i.e. the 75% quantile) for each model parameter. For example, the confidence interval for he birth-rate λ_1 will be between `CI50lower$birth_rates[1]` and `CI50upper$birth_rates[1]`. Only included if Nbootstraps>0.

CI95lower   Similar to `CI50lower`, but listing the lower end of the 95% confidence interval (i.e. the 2.5% quantile) for each model parameter. Only included if Nbootstraps>0.

CI95upper   Similar to `CI50upper`, but listing the upper end of the 95% confidence interval (i.e. the 97.5% quantile) for each model parameter. Only included if Nbootstraps>0.

CI   2D numeric matrix, listing maximum-likelihood estimates, standard errors and confidence intervals for all model parameters (one row per parameter, one column for ML-estimates, one column for standard errors, two columns per confidence interval). Standard errors and confidence intervals are as estimated using parametric bootstrapping. This matrix contains the same information as `parameters`, `standard_errors`, `CI50lower`, `CI50upper`, `CI95lower` and `CI95upper`, but in a more compact format. Only included if Nbootstraps>0.

ancestral_likelihoods

> 2D matrix of size Nnodes x Nstates, listing the computed state-likelihoods for each node in the tree. These may be used for "local" ancestral state reconstructions, based on the information contained in the subtree descending from each node. Note that for each node the ancestral likelihoods have been normalized for numerical reasons, however they should not be interpreted as actual probabilities. For each node n and state s, `ancestral_likelihoods[n,s]` is proportional to the likelihood of observing the descending subtree and associated tip proxy states, if node n was at state s. Only included if `include_ancestral_likelihoods==TRUE`.

Author(s)

Stilianos Louca

References

W. P. Maddison, P. E. Midford, S. P. Otto (2007). Estimating a binary character's effect on speciation and extinction. Systematic Biology. 56:701-710.

R. G. FitzJohn, W. P. Maddison, S. P. Otto (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology. 58:595-611

R. G. FitzJohn (2012). Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution. 3:1084-1092

J. M. Beaulieu and B. C. O'Meara (2016). Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology. 65:583-601.

D. Kuehnert, T. Stadler, T. G. Vaughan, A. J. Drummond (2016). Phylodynamics with migration: A computational framework to quantify population structure from genomic data. Molecular Biology and Evolution. 33:2102-2116.

P. van Els, R. S. Etiene, L. Herrera-Alsina (2018). Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. Systematic Biology. syy057.

S. Louca and M. W. Pennell (2020). A general and efficient algorithm for the likelihood of diversification and discrete-trait evolutionary models. Systematic Biology. 69:545-556.

See Also

simulate_dsse, asr_mk_model, fit_tree_model

Examples

```
# EXAMPLE 1: BiSSE model
# - - - - - - - - - - - - - - -
# Choose random BiSSE model parameters
Nstates = 2
Q = get_random_mk_transition_matrix(Nstates, rate_model="ARD", max_rate=0.1)
parameters = list(birth_rates     = runif(Nstates,5,10),
                  death_rates     = runif(Nstates,0,5),
                  transition_matrix = Q)
rarefaction = 0.5 # randomly omit half of the tips


# Simulate a tree under the BiSSE model
```

```
simulation = simulate_musse(Nstates,
                            parameters        = parameters,
                            max_tips          = 1000,
                            sampling_fractions = rarefaction)
tree      = simulation$tree
tip_states = simulation$tip_states

## Not run:
# fit BiSSE model to tree & tip data
fit = fit_musse(tree,
                Nstates           = Nstates,
                tip_pstates       = tip_states,
                sampling_fractions = rarefaction)
if(!fit$success){
  cat(sprintf("ERROR: Fitting failed"))
}else{
  # compare fitted birth rates to true values
  errors = (fit$parameters$birth_rates - parameters$birth_rates)
  relative_errors = errors/parameters$birth_rates
  cat(sprintf("BiSSE relative birth-rate errors:\n"))
  print(relative_errors)
}

## End(Not run)


# EXAMPLE 2: HiSSE model, with bootstrapping
# - - - - - - - - - - - - - - - -
# Choose random HiSSE model parameters
Nstates  = 4
NPstates = 2
Q = get_random_mk_transition_matrix(Nstates, rate_model="ARD", max_rate=0.1)
rarefaction = 0.5 # randomly omit half of the tips
parameters = list(birth_rates       = runif(Nstates,5,10),
                  death_rates       = runif(Nstates,0,5),
                  transition_matrix = Q)

# reveal the state of 30% & 60% of tips (in state 1 & 2, respectively)
reveal_fractions = c(0.3,0.6)

# use proxy map corresponding to Beaulieu and O'Meara (2016)
proxy_map = c(1,2,1,2)

# Simulate a tree under the HiSSE model
simulation = simulate_musse(Nstates,
                            NPstates          = NPstates,
                            proxy_map         = proxy_map,
                            parameters        = parameters,
                            max_tips          = 1000,
                            sampling_fractions = rarefaction,
                            reveal_fractions  = reveal_fractions)
tree      = simulation$tree
tip_states = simulation$tip_proxy_states
```

```
## Not run:
# fit HiSSE model to tree & tip data
# run multiple trials to ensure global optimum
# also estimate confidence intervals via bootstrapping
fit = fit_musse(tree,
                Nstates           = Nstates,
                NPstates          = NPstates,
                proxy_map         = proxy_map,
                tip_pstates       = tip_states,
                sampling_fractions = rarefaction,
                reveal_fractions  = reveal_fractions,
                Ntrials           = 5,
                Nbootstraps       = 10,
                max_model_runtime = 0.1)
if(!fit$success){
  cat(sprintf("ERROR: Fitting failed"))
}else{
  # compare fitted birth rates to true values
  errors = (fit$parameters$birth_rates - parameters$birth_rates)
  relative_errors = errors/parameters$birth_rates
  cat(sprintf("HiSSE relative birth-rate errors:\n"))
  print(relative_errors)

  # print 95%-confidence interval for first birth rate
  cat(sprintf("CI95 for lambda1: %g-%g",
              fit$CI95lower$birth_rates[1],
              fit$CI95upper$birth_rates[1]))
}

## End(Not run)
```

---

fit_sbm_const   *Fit a phylogeographic Spherical Brownian Motion model.*

Description

Given one or more rooted phylogenetic trees and geographic coordinates (latitudes & longitudes) for the tips of each tree, this function estimates the diffusivity of a Spherical Brownian Motion (SBM) model for the evolution of geographic location along lineages (Perrin 1928; Brillinger 2012). Estimation is done via maximum-likelihood and using independent contrasts between sister lineages.

Usage

```
fit_sbm_const(trees,
        tip_latitudes,
        tip_longitudes,
        radius,
        phylodistance_matrixes  = NULL,
```

```
        clade_states            = NULL,
        planar_approximation    = FALSE,
        only_basal_tip_pairs    = FALSE,
        only_distant_tip_pairs  = FALSE,
        min_MRCA_time           = 0,
        max_MRCA_age            = Inf,
        max_phylodistance       = Inf,
        no_state_transitions    = FALSE,
        only_state              = NULL,
        min_diffusivity         = NULL,
        max_diffusivity         = NULL,
        Nbootstraps             = 0,
        NQQ                     = 0,
        SBM_PD_functor          = NULL,
        focal_diffusivities     = NULL)
```

Arguments

trees   Either a single rooted tree or a list of rooted trees, of class "phylo". The root of
                        each tree is assumed to be the unique node with no incoming edge. Edge lengths
                        are assumed to represent time intervals or a similarly interpretable phylogenetic
                        distance. When multiple trees are provided, it is either assumed that their roots
                        coincide in time (if align_trees_at_root=TRUE) or that each tree's youngest
                        tip was sampled at present day (if align_trees_at_root=FALSE).

tip_latitudes   Numeric vector of length Ntips, or a list of vectors, listing latitudes of tips in
                        decimal degrees (from -90 to 90). If trees is a list of trees, then tip_latitudes
                        should be a list of vectors of the same length as trees, listing tip latitudes for
                        each of the input trees.

tip_longitudes   Numeric vector of length Ntips, or a list of vectors, listing longitudes of tips in
                        decimal degrees (from -180 to 180). If trees is a list of trees, then tip_longitudes
                        should be a list of vectors of the same length as trees, listing tip longitudes for
                        each of the input trees.

radius   Strictly positive numeric, specifying the radius of the sphere. For Earth, the
                        mean radius is 6371 km.

phylodistance_matrixes

                        Numeric matrix, or a list of numeric matrixes, listing phylogenetic distances be-
                        tween tips for each tree. If trees is a list of trees, then phylodistance_matrixes
                        should be a list of the same length as trees, whose n-th element should be
                        a numeric matrix comprising as many rows and columns as there are tips in
                        the n-th tree; the entry phylodistance_matrixes[[n]][i,j] is the phyloge-
                        netic distance between tips i and j in tree n. If trees is a single tree, then
                        phylodistance_matrixes can be a single numeric matrix. If NULL (default),
                        phylogenetic distances between tips are calculated based on the provided trees,
                        otherwise phylogenetic distances are taken from phylodistance_matrixes; in
                        the latter case the trees are only used for the topology (determining tip pairs for
                        independent contrasts), but not for calculating phylogenetic distances.

clade_states   Either NULL, or an integer vector of length Ntips+Nnodes, or a list of integer
                        vectors, listing discrete states of every tip and node in the tree. If trees is a

list of trees, then clade_states should be a list of vectors of the same length as trees, listing tip and node states for each of the input trees. For example, clade_states[[2]][10] specifies the state of the 10-th tip or node in the 2nd tree. States may be, for example, geographic regions, sub-types, discrete traits etc, and can be used to restrict independent contrasts to tip pairs within the same state (see option no_state_transitions).

planar_approximation

Logical, specifying whether to estimate the diffusivity based on a planar approximation of the SBM model, i.e. by assuming that geographic distances between tips are as if tips are distributed on a 2D cartesian plane. This approximation is only accurate if geographical distances between tips are small compared to the sphere's radius.

only_basal_tip_pairs

Logical, specifying whether to only compare immediate sister tips, i.e., tips connected through a single parental node.

only_distant_tip_pairs

Logical, specifying whether to only compare tips at distinct geographic locations.

min_MRCA_time   Numeric, specifying the minimum allowed time (distance from root) of the most recent common ancestor (MRCA) of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at least this distance from the root. Set min_MRCA_time<=0 to disable this filter.

max_MRCA_age   Numeric, specifying the maximum allowed age (distance from youngest tip) of the MRCA of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at most this age (time to present). Set max_MRCA_age=Inf to disable this filter.

max_phylodistance

Numeric, maximum allowed geodistance for an independent contrast to be included in the SBM fitting. Set max_phylodistance=Inf to disable this filter.

no_state_transitions

Logical, specifying whether to omit independent contrasts between tips whose shortest connecting paths include state transitions. If TRUE, only tips within the same state and with no transitions between them (as specified in clade_states) are compared. If TRUE, then clade_states must be provided.

only_state   Optional integer, specifying the state in which tip pairs (and their connecting ancestral nodes) must be in order to be considered. If specified, then clade_states must be provided.

min_diffusivity

Non-negative numeric, specifying the minimum possible diffusivity. If NULL, this is automatically chosen.

max_diffusivity

Non-negative numeric, specifying the maximum possible diffusivity. If NULL, this is automatically chosen.

Nbootstraps   Non-negative integer, specifying an optional number of parametric bootstraps to performs for estimating standard errors and confidence intervals.

NQQ   Integer, optional number of simulations to perform for creating QQ plots of the
                   theoretically expected distribution of geodistances vs. the empirical distribution
                   of geodistances (across independent contrasts). The resolution of the returned
                   QQ plot will be equal to the number of independent contrasts used for fitting. If
                   <=0, no QQ plots will be calculated.

SBM_PD_functor   SBM probability density functor object. Used internally for efficiency and for
                   debugging purposes, and should be kept at its default value NULL.

focal_diffusivities

                   Optional numeric vector, listing diffusivities of particular interest and for which
                   the log-likelihoods should be returned. This may be used e.g. for diagnostic pur-
                   poses, e.g. to see how "sharp" the likelihood peak is at the maximum-likelihood
                   estimate.

Details

For short expected transition distances this function uses the approximation formula by Ghosh et
al. (2012). For longer expected transition distances the function uses a truncated approximation of
the series representation of SBM transition densities (Perrin 1928). It is assumed that tips are sam-
pled randomly without any biases for certain geographic regions. If you suspect strong geographic
sampling biases, consider using the function `fit_sbm_geobiased_const`.

This function can use multiple trees to fit the diffusivity under the assumption that each tree is an
independent realization of the same SBM process, i.e. all lineages in all trees dispersed with the
same diffusivity.

If edge.length is missing from one of the input trees, each edge in the tree is assumed to have
length 1. The tree may include multifurcations as well as monofurcations, however multifurcations
are internally expanded into bifurcations by adding dummy nodes.

Value

A list with the following elements:

success   Logical, indicating whether the fitting was successful. If FALSE, then an addi-
                   tional return variable, error, will contain a description of the error; in that case
                   all other return variables may be undefined.

diffusivity   Numeric, the estimated diffusivity, in units distance^2/time. Distance units are
                   the same as used for the radius, and time units are the same as the tree's edge
                   lengths. For example, if the radius was specified in km and edge lengths are in
                   Myr, then the estimated diffusivity will be in km^2/Myr.

loglikelihood   Numeric, the log-likelihood of the data at the estimated diffusivity.

Ncontrasts   Integer, number of independent contrasts (i.e., tip pairs) used to estimate the
                   diffusivity. This is the number of independent data points used.

phylodistances   Numeric vector of length Ncontrasts, listing the phylogenetic distances of the
                   independent contrasts used in the fitting.

geodistances   Numeric vector of length Ncontrasts, listing the geographical distances of the
                   independent contrasts used in the fitting.

focal_loglikelihoods

> Numeric vector of the same length as `focal_diffusivities`, listing the log-likelihoods for the diffusivities provided in `focal_diffusivities`.

standard_error Numeric, estimated standard error of the estimated diffusivity, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI50lower Numeric, lower bound of the 50% confidence interval for the estimated diffusivity (25-75% percentile), based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI50upper Numeric, upper bound of the 50% confidence interval for the estimated diffusivity, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI95lower Numeric, lower bound of the 95% confidence interval for the estimated diffusivity (2.5-97.5% percentile), based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI95upper Numeric, upper bound of the 95% confidence interval for the estimated diffusivity, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

consistency Numeric between 0 and 1, estimated consistency of the data with the fitted model. If L denotes the loglikelihood of new data generated by the fitted model (under the same model) and M denotes the expectation of L, then `consistency` is the probability that $|L - M|$ will be greater or equal to $|X - M|$, where X is the loglikelihood of the original data under the fitted model. Only returned if `Nbootstraps>0`. A low consistency (e.g., <0.05) indicates that the fitted model is a poor description of the data. See Lindholm et al. (2019) for background.

QQplot Numeric matrix of size Ncontrasts x 2, listing the computed QQ-plot. The first column lists quantiles of geodistances in the original dataset, the 2nd column lists quantiles of hypothetical geodistances simulated based on the fitted model.

SBM_PD_functor SBM probability density functor object. Used internally for efficiency and for debugging purposes.

Author(s)

Stilianos Louca

References

F. Perrin (1928). Etude mathematique du mouvement Brownien de rotation. 45:1-51.

D. R. Brillinger (2012). A particle migrating randomly on a sphere. in Selected Works of David Brillinger. Springer.

A. Ghosh, J. Samuel, S. Sinha (2012). A Gaussian for diffusion on the sphere. Europhysics Letters. 98:30003.

A. Lindholm, D. Zachariah, P. Stoica, T. B. Schoen (2019). Data consistency approach to model validation. IEEE Access. 7:59788-59796.

S. Louca (2021). Phylogeographic estimation and simulation of global diffusive dispersal. Systematic Biology. 70:340-359.

See Also

[fit_sbm_geobiased_const](), [simulate_sbm](), [fit_sbm_parametric](), [fit_sbm_linear](), [fit_sbm_on_grid]()

Examples

```
## Not run:
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=500)$tree

# simulate SBM on the tree
D = 1e4
simulation = simulate_sbm(tree, radius=6371, diffusivity=D)

# fit SBM on the tree
fit = fit_sbm_const(tree,simulation$tip_latitudes,simulation$tip_longitudes,radius=6371)
cat(sprintf('True D=%g, fitted D=%g\n',D,fit$diffusivity))

## End(Not run)
```

---

fit_sbm_geobiased_const

> *Fit a phylogeographic Spherical Brownian Motion model with geographic sampling bias.*

Description

Given one or more rooted phylogenetic trees and geographic coordinates (latitudes & longitudes) for the tips of each tree, this function estimates the diffusivity of a Spherical Brownian Motion (SBM) model for the evolution of geographic location along lineages (Perrin 1928; Brillinger 2012), while correcting for geographic sampling biases. Estimation is done via maximum-likelihood and using independent contrasts between sister lineages, while correction for geographic sampling bias is done through an iterative simulation+fitting process until convergence.

Usage

```
fit_sbm_geobiased_const(trees,
                        tip_latitudes,
                        tip_longitudes,
                        radius,
                        reference_latitudes   = NULL,
                        reference_longitudes  = NULL,
                        only_basal_tip_pairs  = FALSE,
                        only_distant_tip_pairs = FALSE,
                        min_MRCA_time         = 0,
                        max_MRCA_age          = Inf,
                        max_phylodistance     = Inf,
                        min_diffusivity       = NULL,
                        max_diffusivity       = NULL,
                        rarefaction           = 0.1,
                        Nsims                 = 100,
                        max_iterations        = 100,
```

```
                                Nbootstraps        = 0,
                                NQQ                = 0,
                                Nthreads           = 1,
                                SBM_PD_functor     = NULL,
                                verbose            = FALSE,
                                verbose_prefix     = "")
```

Arguments

trees
: Either a single rooted tree or a list of rooted trees, of class "phylo". The root of each tree is assumed to be the unique node with no incoming edge. Edge lengths are assumed to represent time intervals or a similarly interpretable phylogenetic distance. When multiple trees are provided, it is either assumed that their roots coincide in time (if align_trees_at_root=TRUE) or that each tree's youngest tip was sampled at present day (if align_trees_at_root=FALSE).

tip_latitudes
: Numeric vector of length Ntips, or a list of vectors, listing latitudes of tips in decimal degrees (from -90 to 90). If trees is a list of trees, then tip_latitudes should be a list of vectors of the same length as trees, listing tip latitudes for each of the input trees.

tip_longitudes
: Numeric vector of length Ntips, or a list of vectors, listing longitudes of tips in decimal degrees (from -180 to 180). If trees is a list of trees, then tip_longitudes should be a list of vectors of the same length as trees, listing tip longitudes for each of the input trees.

radius
: Strictly positive numeric, specifying the radius of the sphere. For Earth, the mean radius is 6371 km.

reference_latitudes
: Optional numeric vector, listing latitudes of reference coordinates based on which to calculate the geographic sampling density. If NULL, the geographic sampling density is estimated based on tip_latitudes and tip_longitudes.

reference_longitudes
: Optional numeric vector of the same length as reference_latitudes, listing latitudes of reference coordinates based on which to calculate the geographic sampling density. If NULL, the geographic sampling density is estimated based on tip_latitudes and tip_longitudes.

only_basal_tip_pairs
: Logical, specifying whether to only compare immediate sister tips, i.e., tips connected through a single parental node.

only_distant_tip_pairs
: Logical, specifying whether to only compare tips at distinct geographic locations.

min_MRCA_time
: Numeric, specifying the minimum allowed time (distance from root) of the most recent common ancestor (MRCA) of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at least this distance from the root. Set min_MRCA_time<=0 to disable this filter.

max_MRCA_age
: Numeric, specifying the maximum allowed age (distance from youngest tip) of the MRCA of sister tips considered in the fitting. In other words, an independent

contrast is only considered if the two sister tips' MRCA has at most this age (time to present). Set `max_MRCA_age=Inf` to disable this filter.

max_phylodistance

Numeric, maximum allowed geodistance for an independent contrast to be included in the SBM fitting. Set `max_phylodistance=Inf` to disable this filter.

min_diffusivity

Non-negative numeric, specifying the minimum possible diffusivity. If `NULL`, this is automatically chosen.

max_diffusivity

Non-negative numeric, specifying the maximum possible diffusivity. If `NULL`, this is automatically chosen.

rarefaction   Numeric, between 00 and 1, specifying the fraction of extant lineages to sample from the simulated trees. Should be strictly smaller than 1, in order for geographic bias correction to have an effect. Note that regardless of `rarefaction`, the simulated trees will have the same size as the original trees.

Nsims   Integer, number of SBM simulatons to perform per iteration for assessing the effects of geographic bias. Smaller trees require larger `Nsims` (due to higher stochasticity). This must be at least 2, although values of 100-1000 are recommended.

max_iterations   Integer, maximum number of iterations (correction steps) to perform before giving up.

Nbootstraps   Non-negative integer, specifying an optional number of parametric bootstraps to performs for estimating standard errors and confidence intervals.

NQQ   Integer, optional number of simulations to perform for creating QQ plots of the theoretically expected distribution of geodistances vs. the empirical distribution of geodistances (across independent contrasts). The resolution of the returned QQ plot will be equal to the number of independent contrasts used for fitting. If <=0, no QQ plots will be calculated.

Nthreads   Integer, number of parallel threads to use. Ignored on Windows machines.

SBM_PD_functor   SBM probability density functor object. Used internally for efficiency and for debugging purposes, and should be kept at its default value `NULL`.

verbose   Logical, specifying whether to print progress reports and warnings to the screen.

verbose_prefix   Character, specifying the line prefix for printing progress reports to the screen.

Details

This function tries to estimate the true spherical diffusivity of an SBM model of geographic diffusive dispersal, while correcting for geographic sampling biases. This is done using an iterative refinement approach, by which trees and tip locations are repeatedly simulated under the current true diffusivity estimate and the diffusivity estimated from those simulated data are compared to the originally uncorrected diffusivity estimate. Trees are simulated according to a birth-death model with constant rates, fitted to the original input trees (a congruent birth-death model is chosen to match the requested `rarefaction`). Simulated trees are subsampled (rarefied) to match the original input tree sizes, with sampled lineages chosen randomly but in a geographically biased way that resembles the original geographic sampling density (e.g., as inferred from the `reference_latitudes`

and reference_longitudes). Internally, this function repeatedly applies [fit_sbm_const](#) and [simulate_sbm](#). If the true sampling fraction of the input trees is unknown, then it is advised to perform the analysis with a few alternative rarefaction values (e.g., 0.01 and 0.1) to verify the robustness of the estimates.

If edge.length is missing from one of the input trees, each edge in the tree is assumed to have length 1. The tree may include multifurcations as well as monofurcations, however multifurcations are internally expanded into bifurcations by adding dummy nodes.

Value

A list with the following elements:

success	Logical, indicating whether the fitting was successful. If FALSE, then an additional return variable, error, will contain a description of the error; in that case all other return variables may be undefined.
Nlat	Integer, number of latitude-tiles used for building a map of the geographic sampling biases.
Nlon	Integer, number of longitude-tiles used for building a map of the geographic sampling biases.
diffusivity	Numeric, the estimated true diffusivity, i.e. accounting for geographic sampling biases, in units distance^2/time. Distance units are the same as used for the radius, and time units are the same as the tree's edge lengths. For example, if the radius was specified in km and edge lengths are in Myr, then the estimated diffusivity will be in km^2/Myr.
correction_factor	Numeric, estimated ratio between the true diffusivity and the original (uncorrected) diffusivity estimate.
Niterations	Integer, the number of iterations performed until convergence.
stopping_criterion	Character, a short description of the criterion by which the iteration was eventually halted.
uncorrected_fit_diffusivity	Numeric, the originally estimated (uncorrected) diffusivity.
last_sim_fit_diffusivity	Numeric, the mean uncorrected diffuvity estimated from the simulated data in the last iteration. Convergence means that last_sim_fit_diffusivity came close to uncorrected_fit_diffusivity.
all_correction_factors	Numeric vector of length Niterations, listing the estimated correction factors in each iteration.
all_diffusivity_estimates	Numeric vector of length Niterations, listing the mean uncorrected diffusivity estimated from the simulated data in each iteration.
Ntrees	Integer, number of trees considered for the simulations. This might have smaller than length(trees), if for some trees fitting a birth-death model was not possible.

lambda	Numeric vector of length Ntrees, listing the birth rates used to simulate the trees.
mu	Numeric vector of length Ntrees, listing the death rates used to simulate the trees.
rarefaction	Numeric vector of length Ntrees, listing the rarefactions (sampling fractions) used to simulate the trees. These will typically be equal to the rarefaction provided by the function caller, but may differ for example if the congruence class did not include a birth-death model with the requested rarefaction.
Ncontrasts	Integer, number of independent contrasts (i.e., tip pairs) used to estimate the diffusivity. This is the number of independent data points used.
standard_error	Numeric, estimated standard error of the estimated true diffusivity, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI50lower	Numeric, lower bound of the 50% confidence interval for the estimated true diffusivity (25-75% percentile), based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI50upper	Numeric, upper bound of the 50% confidence interval for the estimated true diffusivity, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI95lower	Numeric, lower bound of the 95% confidence interval for the estimated true diffusivity (2.5-97.5% percentile), based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI95upper	Numeric, upper bound of the 95% confidence interval for the estimated true diffusivity, based on parametric bootstrapping. Only returned if Nbootstraps>0.
QQplot	Numeric matrix of size Ncontrasts x 2, listing the computed QQ-plot. The first column lists quantiles of geodistances in the original dataset, the 2nd column lists quantiles of hypothetical geodistances simulated based on the estimated true diffusivity.
SBM_PD_functor	SBM probability density functor object. Used internally for efficiency and for debugging purposes.

Author(s)

Stilianos Louca

References

F. Perrin (1928). Etude mathematique du mouvement Brownien de rotation. 45:1-51.

D. R. Brillinger (2012). A particle migrating randomly on a sphere. in Selected Works of David Brillinger. Springer.

A. Ghosh, J. Samuel, S. Sinha (2012). A Gaussian for diffusion on the sphere. Europhysics Letters. 98:30003.

S. Louca (2021). Phylogeographic estimation and simulation of global diffusive dispersal. Systematic Biology. 70:340-359.

S. Louca (in review as of 2021). The rates of global microbial dispersal.

See Also

simulate_sbm, fit_sbm_parametric, fit_sbm_linear, fit_sbm_on_grid

Examples

```
## Not run:
NFullTips   = 10000
diffusivity = 1
radius      = 6371

# generate tree and run SBM on it
cat(sprintf("Generating tree and simulating SBM (true D=%g)..\n",diffusivity))
tree = castor::generate_tree_hbd_reverse(Ntips = NFullTips,
                                         lambda = 5e-7,
                                         mu     = 2e-7,
                                         rho    = 1)$trees[[1]]
SBMsim = simulate_sbm(tree = tree, radius = radius, diffusivity = diffusivity)

# select subset of tips only found in certain geographic regions
min_abs_lat = 30
max_abs_lat = 80
min_lon     = 0
max_lon     = 90
keep_tips   = which((abs(SBMsim$tip_latitudes)<=max_abs_lat)
                    & (abs(SBMsim$tip_latitudes)>=min_abs_lat)
                    & (SBMsim$tip_longitudes<=max_lon)
                    & (SBMsim$tip_longitudes>=min_lon))
rarefaction    = castor::get_subtree_with_tips(tree, only_tips = keep_tips)
tree           = rarefaction$subtree
tip_latitudes  = SBMsim$tip_latitudes[rarefaction$new2old_tip]
tip_longitudes = SBMsim$tip_longitudes[rarefaction$new2old_tip]
Ntips          = length(tree$tip.label)
rarefaction    = Ntips/NFullTips

# fit SBM while correcting for geographic sampling biases
fit = castor:::fit_sbm_geobiased_const(trees          = tree,
                                       tip_latitudes  = tip_latitudes,
                                       tip_longitudes = tip_longitudes,
                                       radius         = radius,
                                       rarefaction    = Ntips/NFullTips,
                                       Nsims          = 10,
                                       Nthreads       = 4,
                                       verbose        = TRUE,
                                       verbose_prefix = "   ")
if(!fit$success){
    cat(sprintf("ERROR: %s\n",fit$error))
}else{
    cat(sprintf("Estimated true D = %g\n",fit$diffusivity))
}

## End(Not run)
```

---

fit_sbm_linear   *Fit a phylogeographic Spherical Brownian Motion model with linearly*
                              varying diffusivity.

Description

Given a rooted phylogenetic tree and geographic coordinates (latitudes & longitudes) for its tips, this
function estimates the diffusivity of a Spherical Brownian Motion (SBM) model for the evolution
of geographic location along lineages (Perrin 1928; Brillinger 2012), assuming that the diffusivity
varies linearly over time. Estimation is done via maximum-likelihood and using independent con-
trasts between sister lineages. This function is designed to estimate the diffusivity over time, by
fitting two parameters defining the diffusivity as a linear function of time. For fitting more general
functional forms see `fit_sbm_parametric`.

Usage

```
fit_sbm_linear(tree,
               tip_latitudes,
               tip_longitudes,
               radius,
               clade_states         = NULL,
               planar_approximation = FALSE,
               only_basal_tip_pairs = FALSE,
               only_distant_tip_pairs= FALSE,
               min_MRCA_time        = 0,
               max_MRCA_age         = Inf,
               max_phylodistance    = Inf,
               no_state_transitions = FALSE,
               only_state           = NULL,
               time1                = 0,
               time2                = NULL,
               Ntrials              = 1,
               Nthreads             = 1,
               Nbootstraps          = 0,
               Ntrials_per_bootstrap = NULL,
               Nsignificance        = 0,
               NQQ                  = 0,
               fit_control          = list(),
               SBM_PD_functor       = NULL,
               verbose              = FALSE,
               verbose_prefix       = "")
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with
                no incoming edge. Edge lengths are assumed to represent time intervals or a
                similarly interpretable phylogenetic distance.

tip_latitudes   Numeric vector of length Ntips, listing latitudes of tips in decimal degrees (from -90 to 90). The order of entries must correspond to the order of tips in the tree (i.e., as listed in `tree$tip.label`).

tip_longitudes   Numeric vector of length Ntips, listing longitudes of tips in decimal degrees (from -180 to 180). The order of entries must correspond to the order of tips in the tree (i.e., as listed in `tree$tip.label`).

radius   Strictly positive numeric, specifying the radius of the sphere. For Earth, the mean radius is 6371 km.

clade_states   Optional integer vector of length Ntips+Nnodes, listing discrete states of every tip and node in the tree. The order of entries must match the order of tips and nodes in the tree. States may be, for example, geographic regions, sub-types, discrete traits etc, and can be used to restrict independent contrasts to tip pairs within the same state (see option `no_state_transitions`).

planar_approximation

    Logical, specifying whether to estimate the diffusivity based on a planar approximation of the SBM model, i.e. by assuming that geographic distances between tips are as if tips are distributed on a 2D cartesian plane. This approximation is only accurate if geographical distances between tips are small compared to the sphere's radius.

only_basal_tip_pairs

    Logical, specifying whether to only compare immediate sister tips, i.e., tips connected through a single parental node.

only_distant_tip_pairs

    Logical, specifying whether to only compare tips at distinct geographic locations.

min_MRCA_time   Numeric, specifying the minimum allowed time (distance from root) of the most recent common ancestor (MRCA) of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at least this distance from the root. Set `min_MRCA_time=0` to disable this filter.

max_MRCA_age   Numeric, specifying the maximum allowed age (distance from youngest tip) of the MRCA of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at most this age (time to present). Set `max_MRCA_age=Inf` to disable this filter.

max_phylodistance

    Numeric, maximum allowed geodistance for an independent contrast to be included in the SBM fitting. Set `max_phylodistance=Inf` to disable this filter.

no_state_transitions

    Logical, specifying whether to omit independent contrasts between tips whose shortest connecting paths include state transitions. If `TRUE`, only tips within the same state and with no transitions between them (as specified in `clade_states`) are compared.

only_state   Optional integer, specifying the state in which tip pairs (and their connecting ancestral nodes) must be in order to be considered. If specified, then `clade_states` must be provided.

time1   Optional numeric, specifying the first time point at which to estimate the diffu-
                     sivity. By default this is set to root (i.e., time 0).

time2   Optional numeric, specifying the first time point at which to estimate the diffu-
                     sivity. By default this is set to the present day (i.e., the maximum distance of
                     any tip from the root).

Ntrials   Integer, specifying the number of independent fitting trials to perform, each
                     starting from a random choice of model parameters. Increasing `Ntrials` re-
                     duces the risk of reaching a non-global local maximum in the fitting objective.

Nthreads   Integer, specifying the number of parallel threads to use for performing multiple
                     fitting trials simultaneously. This should generally not exceed the number of
                     available CPUs on your machine. Parallel computing is not available on the
                     Windows platform.

Nbootstraps   Integer, specifying the number of parametric bootstraps to perform for estimat-
                     ing standard errors and confidence intervals of estimated model parameters. Set
                     to 0 for no bootstrapping.

Ntrials_per_bootstrap
                     Integer, specifying the number of fitting trials to perform for each bootstrap sam-
                     pling. If `NULL`, this is set equal to `max(1,Ntrials)`. Decreasing `Ntrials_per_bootstrap`
                     will reduce computation time, at the expense of potentially inflating the esti-
                     mated confidence intervals; in some cases (e.g., for very large trees) this may
                     be useful if fitting takes a long time and confidence intervals are very narrow
                     anyway. Only relevant if `Nbootstraps>0`.

Nsignificance   Integer, specifying the number of simulations to perform under a const-diffusivity
                     model for assessing the statistical significance of the fitted slope. Set to 0 to not
                     calculate the significance of the slope.

NQQ   Integer, optional number of simulations to perform for creating QQ plots of the
                     theoretically expected distribution of geodistances vs. the empirical distribution
                     of geodistances (across independent contrasts). The resolution of the returned
                     QQ plot will be equal to the number of independent contrasts used for fitting. If
                     <=0, no QQ plots will be calculated.

fit_control   Named list containing options for the `nlminb` optimization routine, such as
                     `iter.max`, `eval.max` or `rel.tol`. For a complete list of options and default
                     values see the documentation of `nlminb` in the `stats` package.

SBM_PD_functor   SBM probability density functor object. Used internally for efficiency and for
                     debugging purposes, and should be kept at its default value `NULL`.

verbose   Logical, specifying whether to print progress reports and warnings to the screen.

verbose_prefix   Character, specifying the line prefix for printing progress reports to the screen.

Details

This function is essentially a wrapper for the more general function [`fit_sbm_parametric`](), with the
addition that it can estimate the statistical significance of the fitted linear slope.

The statistical significance of the slope is the probability that a constant-diffusivity SBM model
would generate data that would yield a fitted linear slope equal to or greater than the one fitted to
the original data; the significance is estimated by simulating `Nsignificance` constant-diffusivity

models and then fitting a linear-diffusivity model. The constant diffusivity assumed in these simulations is the maximum-likelihood diffusivity fitted internally using `fit_sbm_const`.

Note that estimation of diffusivity at older times is only possible if the timetree includes extinct tips or tips sampled at older times (e.g., as is often the case in viral phylogenies). If tips are only sampled once at present-day, i.e. the timetree is ultrametric, reliable diffusivity estimates can only be achieved near present times.

For short expected transition distances this function uses the approximation formula by Ghosh et al. (2012) to calculate the probability density of geographical transitions along edges. For longer expected transition distances the function uses a truncated approximation of the series representation of SBM transition densities (Perrin 1928).

If edge.length is missing from one of the input trees, each edge in the tree is assumed to have length 1. The tree may include multifurcations as well as monofurcations, however multifurcations are internally expanded into bifurcations by adding dummy nodes.

Value

A list with the following elements:

success	Logical, indicating whether the fitting was successful. If FALSE, then an additional return variable, error, will contain a description of the error; in that case all other return variables may be undefined.
objective_value	The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.
objective_name	The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".
times	Numeric vector of size 2, listing the two time points at which the diffusivity was estimated (time1 and time2).
diffusivities	Numeric vector of size 2, listing the fitted diffusivity at time1 and time2. The fitted model assumes that the diffusivity varied linearly between those two time points.
loglikelihood	The log-likelihood of the fitted linear model for the given data.
NFP	Integer, number of fitted (i.e., non-fixed) model parameters. Will always be 2.
Ncontrasts	Integer, number of independent contrasts used for fitting.
AIC	The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.
BIC	The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of independent contrasts), and L is the maximized likelihood.
converged	Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase iter.max and/or eval.max).

Niterations	Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.
Nevaluations	Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.
trial_start_objectives	
	Numeric vector of size Ntrials, listing the initial objective values (e.g., log-likelihoods) for each fitting trial, i.e. at the start parameter values.
trial_objective_values	
	Numeric vector of size Ntrials, listing the final maximized objective values (e.g., loglikelihoods) for each fitting trial.
trial_Nstart_attempts	
	Integer vector of size Ntrials, listing the number of start attempts for each fitting trial, until a starting point with valid likelihood was found.
trial_Niterations	
	Integer vector of size Ntrials, listing the number of iterations needed for each fitting trial.
trial_Nevaluations	
	Integer vector of size Ntrials, listing the number of likelihood evaluations needed for each fitting trial.
standard_errors	
	Numeric vector of size 2, estimated standard error of the fitted diffusivity at the root and present, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI50lower	Numeric vector of size 2, lower bound of the 50% confidence interval (25-75% percentile) for the fitted diffusivity at the root and present, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI50upper	Numeric vector of size 2, upper bound of the 50% confidence interval for the fitted diffusivity at the root and present, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI95lower	Numeric vector of size 2, lower bound of the 95% confidence interval (2.5-97.5% percentile) for the fitted diffusivity at the root and present, based on parametric bootstrapping. Only returned if Nbootstraps>0.
CI95upper	Numeric vector of size 2, upper bound of the 95% confidence interval for the fitted diffusivity at the root and present, based on parametric bootstrapping. Only returned if Nbootstraps>0.
consistency	Numeric between 0 and 1, estimated consistency of the data with the fitted model. See the documentation of fit_sbm_const for an explanation. Only returned if Nbootstraps>0.
significance	Numeric between 0 and 1, estimate statistical significance of the fitted linear slope. Only returned if Nsignificance>0.
QQplot	Numeric matrix of size Ncontrasts x 2, listing the computed QQ-plot. The first column lists quantiles of geodistances in the original dataset, the 2nd column lists quantiles of hypothetical geodistances simulated based on the fitted model.
SBM_PD_functor	SBM probability density functor object. Used internally for efficiency and for debugging purposes.

Author(s)

Stilianos Louca

References

F. Perrin (1928). Etude mathematique du mouvement Brownien de rotation. 45:1-51.

D. R. Brillinger (2012). A particle migrating randomly on a sphere. in Selected Works of David Brillinger. Springer.

A. Ghosh, J. Samuel, S. Sinha (2012). A Gaussian for diffusion on the sphere. Europhysics Letters. 98:30003.

S. Louca (2021). Phylogeographic estimation and simulation of global diffusive dispersal. Systematic Biology. 70:340-359.

See Also

simulate_sbm, fit_sbm_const, fit_sbm_parametric, fit_sbm_on_grid

Examples

```
## Not run:
# generate a random tree, keeping extinct lineages
tree_params = list(birth_rate_factor=1, death_rate_factor=0.95)
tree = generate_random_tree(tree_params,max_tips=1000,coalescent=FALSE)$tree

# calculate max distance of any tip from the root
max_time = get_tree_span(tree)$max_distance

# simulate time-dependent SBM on the tree
# we assume that diffusivity varies linearly with time
# in this example we measure distances in Earth radii
radius = 1
diffusivity_functor = function(times, params){
    return(params[1] + (times/max_time)*(params[2]-params[1]))
}
true_params = c(1, 2)
time_grid   = seq(0,max_time,length.out=2)
simulation  = simulate_sbm(tree,
                    radius = radius,
                    diffusivity = diffusivity_functor(time_grid,true_params),
                    time_grid = time_grid)

# fit time-independent SBM to get a rough estimate
fit_const = fit_sbm_const(tree,simulation$tip_latitudes,simulation$tip_longitudes,radius=radius)

# fit SBM model with linearly varying diffusivity
fit = fit_sbm_linear(tree,
            simulation$tip_latitudes,
            simulation$tip_longitudes,
            radius = radius,
            Ntrials = 10)
```

```
# compare fitted & true params
print(true_params)
print(fit$diffusivities)

## End(Not run)
```

---

fit_sbm_on_grid   *Fit a phylogeographic Spherical Brownian Motion model with*
                               piecewise-linear diffusivity.

Description

Given a rooted phylogenetic tree and geographic coordinates (latitudes & longitudes) for its tips,
this function estimates the diffusivity of a Spherical Brownian Motion (SBM) model with time-
dependent diffusivity for the evolution of geographic location along lineages (Perrin 1928; Brillinger
2012). Estimation is done via maximum-likelihood and using independent contrasts between sister
lineages. This function is designed to estimate the diffusivity over time, approximated as a piece-
wise linear profile, by fitting the diffusivity on a discrete set of time points. The user thus provides a
set of time points (`time_grid`), and `fit_sbm_on_grid` estimates the diffusivity on each time point,
while assuming that the diffusivity varies linearly between time points.

Usage

```
fit_sbm_on_grid(tree,
                tip_latitudes,
                tip_longitudes,
                radius,
                clade_states          = NULL,
                planar_approximation  = FALSE,
                only_basal_tip_pairs  = FALSE,
                only_distant_tip_pairs= FALSE,
                min_MRCA_time         = 0,
                max_MRCA_age          = Inf,
                max_phylodistance     = Inf,
                no_state_transitions  = FALSE,
                only_state            = NULL,
                time_grid             = 0,
                guess_diffusivity     = NULL,
                min_diffusivity       = NULL,
                max_diffusivity       = Inf,
                Ntrials               = 1,
                Nthreads              = 1,
                Nbootstraps           = 0,
                Ntrials_per_bootstrap = NULL,
                NQQ                   = 0,
                fit_control           = list(),
```

```
SBM_PD_functor        = NULL,
verbose               = FALSE,
verbose_prefix        = "")
```

Arguments

tree
: A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge. Edge lengths are assumed to represent time intervals or a similarly interpretable phylogenetic distance.

tip_latitudes
: Numeric vector of length Ntips, listing latitudes of tips in decimal degrees (from -90 to 90). The order of entries must correspond to the order of tips in the tree (i.e., as listed in `tree$tip.label`).

tip_longitudes
: Numeric vector of length Ntips, listing longitudes of tips in decimal degrees (from -180 to 180). The order of entries must correspond to the order of tips in the tree (i.e., as listed in `tree$tip.label`).

radius
: Strictly positive numeric, specifying the radius of the sphere. For Earth, the mean radius is 6371 km.

clade_states
: Optional integer vector of length Ntips+Nnodes, listing discrete states of every tip and node in the tree. The order of entries must match the order of tips and nodes in the tree. States may be, for example, geographic regions, sub-types, discrete traits etc, and can be used to restrict independent contrasts to tip pairs within the same state (see option `no_state_transitions`).

planar_approximation
: Logical, specifying whether to estimate the diffusivity based on a planar approximation of the SBM model, i.e. by assuming that geographic distances between tips are as if tips are distributed on a 2D cartesian plane. This approximation is only accurate if geographical distances between tips are small compared to the sphere's radius.

only_basal_tip_pairs
: Logical, specifying whether to only compare immediate sister tips, i.e., tips connected through a single parental node.

only_distant_tip_pairs
: Logical, specifying whether to only compare tips at distinct geographic locations.

min_MRCA_time
: Numeric, specifying the minimum allowed time (distance from root) of the most recent common ancestor (MRCA) of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at least this distance from the root. Set `min_MRCA_time=0` to disable this filter.

max_MRCA_age
: Numeric, specifying the maximum allowed age (distance from youngest tip) of the MRCA of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at most this age (time to present). Set `max_MRCA_age=Inf` to disable this filter.

max_phylodistance
: Numeric, maximum allowed geodistance for an independent contrast to be included in the SBM fitting. Set `max_phylodistance=Inf` to disable this filter.

no_state_transitions

        Logical, specifying whether to omit independent contrasts between tips whose shortest connecting paths include state transitions. If TRUE, only tips within the same state and with no transitions between them (as specified in clade_states) are compared.

only_state   Optional integer, specifying the state in which tip pairs (and their connecting ancestral nodes) must be in order to be considered. If specified, then clade_states must be provided.

time_grid   Numeric vector, specifying discrete time points (counted since the root) at which the diffusivity should be fitted; between these time points the diffusivity is assumed to vary linearly. This time grid should be fine enough to sufficiently capture the variation in the diffusivity over time, but must not be too big to avoid overfitting. If NULL or of size 1, then the diffusivity is assumed to be time-independent. Listed times must be strictly increasing, and should cover at least the full considered time interval (from 0 to the maximum distance of any tip from the root); otherwise, constant extrapolation is used to cover missing times. Note that time is measured in the same units as the tree's edge lengths.

guess_diffusivity

        Optional numeric vector, specifying a first guess for the diffusivity. Either of size 1 (the same first guess for all time points), or of the same length as time_grid (different first guess for each time point, NA are replaced with an automatically chosen first guess). If NULL, the first guess is chosen automatically.

min_diffusivity

        Optional numeric vector, specifying lower bounds for the fitted diffusivity. Either of size 1 (the same lower bound is assumed for all time points), or of the same length as time_grid (different lower bound for each time point, NA are replaced with an automatically chosen lower bound). If NULL, lower bounds are chosen automatically.

max_diffusivity

        Optional numeric vector, specifying upper bounds for the fitted diffusivity. Either of size 1 (the same upper bound is assumed for all time points), or of the same length as time_grid (different upper bound for each time point, NA are replaced with infinity). If NULL, no upper bound is imposed.

Ntrials   Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing Ntrials reduces the risk of reaching a non-global local maximum in the fitting objective.

Nthreads   Integer, specifying the number of parallel threads to use for performing multiple fitting trials simultaneously. This should generally not exceed the number of available CPUs on your machine. Parallel computing is not available on the Windows platform.

Nbootstraps   Integer, specifying the number of parametric bootstraps to perform for estimating standard errors and confidence intervals of estimated model parameters. Set to 0 for no bootstrapping.

Ntrials_per_bootstrap

        Integer, specifying the number of fitting trials to perform for each bootstrap sampling. If NULL, this is set equal to max(1,Ntrials). Decreasing Ntrials_per_bootstrap

will reduce computation time, at the expense of potentially inflating the estimated confidence intervals; in some cases (e.g., for very large trees) this may be useful if fitting takes a long time and confidence intervals are very narrow anyway. Only relevant if Nbootstraps>0.

NQQ Integer, optional number of simulations to perform for creating QQ plots of the theoretically expected distribution of geodistances vs. the empirical distribution of geodistances (across independent contrasts). The resolution of the returned QQ plot will be equal to the number of independent contrasts used for fitting. If <=0, no QQ plots will be calculated.

fit_control Named list containing options for the nlminb optimization routine, such as iter.max, eval.max or rel.tol. For a complete list of options and default values see the documentation of nlminb in the stats package.

SBM_PD_functor SBM probability density functor object. Used internally for efficiency and for debugging purposes, and should be kept at its default value NULL.

verbose Logical, specifying whether to print progress reports and warnings to the screen. Note that errors always cause a return of the function (see return values success and error).

verbose_prefix Character, specifying the line prefix for printing progress reports to the screen.

Details

This function is designed to estimate the diffusivity profile over time, approximated by a piecewise linear function. Fitting is done by maximizing the likelihood of observing the given tip coordinates under the SBM model. Internally, this function uses `fit_sbm_parametric`.

It is generally advised to provide as much information to the function fit_sbm_on_grid as possible, including reasonable lower and upper bounds (min_diffusivity and max_diffusivity). It is important that the time_grid is sufficiently fine to capture the variation of the true diffusivity over time, since the likelihood is calculated under the assumption that the diffusivity varies linearly between grid points. However, depending on the size of the tree, the grid size must not be too large, since otherwise overfitting becomes very likely. The time_grid does not need to be uniform, i.e., you may want to use a finer grid in regions where there's more data (tips) available.

Note that estimation of diffusivity at older times is only possible if the timetree includes extinct tips or tips sampled at older times (e.g., as is often the case in viral phylogenies). If tips are only sampled once at present-day, i.e. the timetree is ultrametric, reliable diffusivity estimates can only be achieved near present times. If the tree is ultrametric, you should consider using `fit_sbm_const` instead.

If edge.length is missing from one of the input trees, each edge in the tree is assumed to have length 1. The tree may include multifurcations as well as monofurcations, however multifurcations are internally expanded into bifurcations by adding dummy nodes.

Value

A list with the following elements:

success Logical, indicating whether the fitting was successful. If FALSE, then an additional return variable, error, will contain a description of the error; in that case all other return variables may be undefined.

objective_value

        The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.

objective_name   The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".

time_grid   Numeric vector, the time-grid on which the diffusivity was fitted.

diffusivity   Numeric vector of size Ngrid (length of time_grid), listing the fitted diffusivities at the various time-grid points.

loglikelihood   The log-likelihood of the fitted model for the given data.

NFP   Integer, number of fitted (i.e., non-fixed) model parameters.

Ncontrasts   Integer, number of independent contrasts used for fitting.

phylodistances Numeric vector of length Ncontrasts, listing phylogenetic (patristic) distances of the independent contrasts.

geodistances   Numeric vector of length Ncontrasts, listing geographic (great circle) distances of the independent contrasts.

child_times1   Numeric vector of length Ncontrasts, listing the times (distance from root) of the first tip in each independent contrast.

child_times2   Numeric vector of length Ncontrasts, listing the times (distance from root) of the second tip in each independent contrast.

MRCA_times   Numeric vector of length Ncontrasts, listing the times (distance from root) of the MRCA of the two tips in each independent contrast.

AIC   The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.

BIC   The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of independent contrasts), and L is the maximized likelihood.

converged   Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase iter.max and/or eval.max).

Niterations   Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.

Nevaluations   Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.

trial_start_objectives

        Numeric vector of size Ntrials, listing the initial objective values (e.g., loglikelihoods) for each fitting trial, i.e. at the start parameter values.

trial_objective_values

        Numeric vector of size Ntrials, listing the final maximized objective values (e.g., loglikelihoods) for each fitting trial.

trial_Nstart_attempts

        Integer vector of size Ntrials, listing the number of start attempts for each fitting trial, until a starting point with valid likelihood was found.

trial_Niterations

> Integer vector of size `Ntrials`, listing the number of iterations needed for each fitting trial.

trial_Nevaluations

> Integer vector of size `Ntrials`, listing the number of likelihood evaluations needed for each fitting trial.

standard_errors

> Numeric vector of size NP, estimated standard error of the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

medians

> Numeric vector of size NP, median the estimated parameters across parametric bootstraps. Only returned if `Nbootstraps>0`.

CI50lower

> Numeric vector of size NP, lower bound of the 50% confidence interval (25-75% percentile) for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI50upper

> Numeric vector of size NP, upper bound of the 50% confidence interval for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI95lower

> Numeric vector of size NP, lower bound of the 95% confidence interval (2.5-97.5% percentile) for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI95upper

> Numeric vector of size NP, upper bound of the 95% confidence interval for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

consistency

> Numeric between 0 and 1, estimated consistency of the data with the fitted model. See the documentation of [`fit_sbm_const`] for an explanation.

QQplot

> Numeric matrix of size Ncontrasts x 2, listing the computed QQ-plot. The first column lists quantiles of geodistances in the original dataset, the 2nd column lists quantiles of hypothetical geodistances simulated based on the fitted model.

SBM_PD_functor

> SBM probability density functor object. Used internally for efficiency and for debugging purposes.

Author(s)

Stilianos Louca

References

F. Perrin (1928). Etude mathematique du mouvement Brownien de rotation. 45:1-51.

D. R. Brillinger (2012). A particle migrating randomly on a sphere. in Selected Works of David Brillinger. Springer.

A. Ghosh, J. Samuel, S. Sinha (2012). A Gaussian for diffusion on the sphere. Europhysics Letters. 98:30003.

S. Louca (2021). Phylogeographic estimation and simulation of global diffusive dispersal. Systematic Biology. 70:340-359.

See Also

[simulate_sbm], [fit_sbm_const], [fit_sbm_parametric], [fit_sbm_linear]

Examples

```
## Not run:
# generate a random tree, keeping extinct lineages
tree_params = list(birth_rate_factor=1, death_rate_factor=0.95)
tree = generate_random_tree(tree_params,max_tips=2000,coalescent=FALSE)$tree

# calculate max distance of any tip from the root
max_time = get_tree_span(tree)$max_distance

# simulate time-dependent SBM on the tree
# using a diffusivity that varies roughly exponentially with time
# In this example we measure distances in Earth radii
radius = 1
fine_time_grid = seq(from=0, to=max_time, length.out=10)
fine_D = 0.01 + 0.03*exp(-2*fine_time_grid/max_time)
simul = simulate_sbm(tree,
                     radius    = radius,
                     diffusivity= fine_D,
                     time_grid  = fine_time_grid)

# fit time-dependent SBM on a time-grid of size 4
fit = fit_sbm_on_grid(tree,
           simul$tip_latitudes,
           simul$tip_longitudes,
           radius    = radius,
           time_grid = seq(from=0,to=max_time,length.out=4),
           Nthreads  = 3,  # use 3 CPUs
           Ntrials   = 30) # avoid local optima through multiple trials

# visually compare fitted & true params
plot(x       = fine_time_grid,
     y       = fine_D,
     type    = 'l',
     col     = 'black',
     xlab    = 'time',
     ylab    = 'D',
     ylim    = c(0,max(fine_D)))
lines(x      = fit$time_grid,
      y       = fit$diffusivity,
      type  = 'l',
      col   = 'blue')

## End(Not run)
```

---

fit_sbm_parametric   *Fit a time-dependent phylogeographic Spherical Brownian Motion*
                              model.

Description

Given a rooted phylogenetic tree and geographic coordinates (latitudes & longitudes) for its tips, this function estimates the diffusivity of a Spherical Brownian Motion (SBM) model with time-dependent diffusivity for the evolution of geographic location along lineages (Perrin 1928; Brillinger 2012). Estimation is done via maximum-likelihood and using independent contrasts between sister lineages. This function is designed to estimate the diffusivity over time, by fitting a finite number of parameters defining the diffusivity as a function of time. The user thus provides the general functional form of the diffusivity that depends on time and NP parameters, and `fit_sbm_parametric` estimates each of the free parameters.

Usage

```
fit_sbm_parametric(tree,
               tip_latitudes,
               tip_longitudes,
               radius,
               param_values,
               param_guess,
               diffusivity,
               time_grid            = NULL,
               clade_states         = NULL,
               planar_approximation = FALSE,
               only_basal_tip_pairs = FALSE,
               only_distant_tip_pairs= FALSE,
               min_MRCA_time        = 0,
               max_MRCA_age         = Inf,
               max_phylodistance    = Inf,
               no_state_transitions = FALSE,
               only_state           = NULL,
               param_min            = -Inf,
               param_max            = +Inf,
               param_scale          = NULL,
               Ntrials              = 1,
               max_start_attempts   = 1,
               Nthreads             = 1,
               Nbootstraps          = 0,
               Ntrials_per_bootstrap = NULL,
               NQQ                  = 0,
               fit_control          = list(),
               SBM_PD_functor       = NULL,
               focal_param_values   = NULL,
               verbose              = FALSE,
               verbose_prefix       = "")
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge. Edge lengths are assumed to represent time intervals or a

similarly interpretable phylogenetic distance.

tip_latitudes	Numeric vector of length Ntips, listing latitudes of tips in decimal degrees (from -90 to 90). The order of entries must correspond to the order of tips in the tree (i.e., as listed in `tree$tip.label`).
tip_longitudes	Numeric vector of length Ntips, listing longitudes of tips in decimal degrees (from -180 to 180). The order of entries must correspond to the order of tips in the tree (i.e., as listed in `tree$tip.label`).
radius	Strictly positive numeric, specifying the radius of the sphere. For Earth, the mean radius is 6371 km.
param_values	Numeric vector of length NP, specifying fixed values for a some or all model parameters. For fitted (i.e., non-fixed) parameters, use NaN or NA. For example, the vector `c(1.5,NA,40)` specifies that the 1st and 3rd model parameters are fixed at the values 1.5 and 40, respectively, while the 2nd parameter is to be fitted. The length of this vector defines the total number of model parameters. If entries in this vector are named, the names are taken as parameter names. Names should be included if you'd like returned parameter vectors to have named entries, or if the `diffusivity` function queries parameter values by name (as opposed to numeric index).
param_guess	Numeric vector of size NP, specifying a first guess for the value of each model parameter. For fixed parameters, guess values are ignored. Can be NULL only if all model parameters are fixed.
diffusivity	Function specifying the diffusivity at any given time (time since the root) and for any given parameter values. This function must take exactly two arguments, the 1st one being a numeric vector (one or more times) and the 2nd one being a numeric vector of size NP (parameter values), and return a numeric vector of the same size as the 1st argument.
time_grid	Numeric vector, specifying times (counted since the root) at which the `diffusivity` function should be evaluated. This time grid must be fine enough to capture the possible variation in the diffusivity over time, within the permissible parameter range. If of size 1, then the diffusivity is assumed to be time-independent. Listed times must be strictly increasing, and should cover at least the full considered time interval (from 0 to the maximum distance of any tip from the root); otherwise, constant extrapolation is used to cover missing times. Can also be NULL or a vector of size 1, in which case the diffusivity is assumed to be time-independent. Note that time is measured in the same units as the tree's edge lengths.
clade_states	Optional integer vector of length Ntips+Nnodes, listing discrete states of every tip and node in the tree. The order of entries must match the order of tips and nodes in the tree. States may be, for example, geographic regions, sub-types, discrete traits etc, and can be used to restrict independent contrasts to tip pairs within the same state (see option `no_state_transitions`).
planar_approximation	
	Logical, specifying whether to estimate the diffusivity based on a planar approximation of the SBM model, i.e. by assuming that geographic distances between tips are as if tips are distributed on a 2D cartesian plane. This approximation is only accurate if geographical distances between tips are small compared to the sphere's radius.

only_basal_tip_pairs

Logical, specifying whether to only compare immediate sister tips, i.e., tips connected through a single parental node.

only_distant_tip_pairs

Logical, specifying whether to only compare tips at distinct geographic locations.

min_MRCA_time   Numeric, specifying the minimum allowed time (distance from root) of the most recent common ancestor (MRCA) of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at least this distance from the root. Set min_MRCA_time=0 to disable this filter.

max_MRCA_age   Numeric, specifying the maximum allowed age (distance from youngest tip) of the MRCA of sister tips considered in the fitting. In other words, an independent contrast is only considered if the two sister tips' MRCA has at most this age (time to present). Set max_MRCA_age=Inf to disable this filter.

max_phylodistance

Numeric, maximum allowed geodistance for an independent contrast to be included in the SBM fitting. Set max_phylodistance=Inf to disable this filter.

no_state_transitions

Logical, specifying whether to omit independent contrasts between tips whose shortest connecting paths include state transitions. If TRUE, only tips within the same state and with no transitions between them (as specified in clade_states) are compared.

only_state   Optional integer, specifying the state in which tip pairs (and their connecting ancestral nodes) must be in order to be considered. If specified, then clade_states must be provided.

param_min   Optional numeric vector of size NP, specifying lower bounds for model parameters. If of size 1, the same lower bound is applied to all parameters. Use -Inf to omit a lower bound for a parameter. If NULL, no lower bounds are applied. For fixed parameters, lower bounds are ignored.

param_max   Optional numeric vector of size NP, specifying upper bounds for model parameters. If of size 1, the same upper bound is applied to all parameters. Use +Inf to omit an upper bound for a parameter. If NULL, no upper bounds are applied. For fixed parameters, upper bounds are ignored.

param_scale   Optional numeric vector of size NP, specifying typical scales for model parameters. If of size 1, the same scale is assumed for all parameters. If NULL, scales are determined automatically. For fixed parameters, scales are ignored. It is strongly advised to provide reasonable scales, as this facilitates the numeric optimization algorithm.

Ntrials   Integer, specifying the number of independent fitting trials to perform, each starting from a random choice of model parameters. Increasing Ntrials reduces the risk of reaching a non-global local maximum in the fitting objective.

max_start_attempts

Integer, specifying the number of times to attempt finding a valid start point (per trial) before giving up on that trial. Randomly choosen extreme start parameters may occasionally result in Inf/undefined likelihoods, so this option allows the algorithm to keep looking for valid starting points.

Nthreads   Integer, specifying the number of parallel threads to use for performing multiple
                    fitting trials simultaneously. This should generally not exceed the number of
                    available CPUs on your machine. Parallel computing is not available on the
                    Windows platform.

Nbootstraps   Integer, specifying the number of parametric bootstraps to perform for estimat-
                    ing standard errors and confidence intervals of estimated model parameters. Set
                    to 0 for no bootstrapping.

Ntrials_per_bootstrap

                    Integer, specifying the number of fitting trials to perform for each bootstrap sam-
                    pling. If NULL, this is set equal to max(1,Ntrials). Decreasing Ntrials_per_bootstrap
                    will reduce computation time, at the expense of potentially inflating the esti-
                    mated confidence intervals; in some cases (e.g., for very large trees) this may
                    be useful if fitting takes a long time and confidence intervals are very narrow
                    anyway. Only relevant if Nbootstraps>0.

NQQ   Integer, optional number of simulations to perform for creating QQ plots of the
                    theoretically expected distribution of geodistances vs. the empirical distribution
                    of geodistances (across independent contrasts). The resolution of the returned
                    QQ plot will be equal to the number of independent contrasts used for fitting. If
                    <=0, no QQ plots will be calculated.

fit_control   Named list containing options for the nlminb optimization routine, such as
                    iter.max, eval.max or rel.tol. For a complete list of options and default
                    values see the documentation of nlminb in the stats package.

SBM_PD_functor   SBM probability density functor object. Used internally for efficiency and for
                    debugging purposes, and should be kept at its default value NULL.

focal_param_values

                    Optional numeric matrix having NP columns and an arbitrary number of rows,
                    listing combinations of parameter values of particular interest and for which
                    the log-likelihoods should be returned. This may be used e.g. for diagnostic
                    purposes, e.g. to examine the shape of the likelihood function.

verbose   Logical, specifying whether to print progress reports and warnings to the screen.
                    Note that errors always cause a return of the function (see return values success
                    and error).

verbose_prefix   Character, specifying the line prefix for printing progress reports to the screen.

Details

This function is designed to estimate a finite set of scalar parameters ($p_1, .., p_n \in \mathbb{R}$) that determine
the diffusivity over time, by maximizing the likelihood of observing the given tip coordinates under
the SBM model. For example, the investigator may assume that the diffusivity exponentially over
time, i.e. can be described by $D(t) = A \cdot e^{-Bt}$ (where A and B are unknown coefficients and t
is time since the root). In this case the model has 2 free parameters, $p_1 = A$ and $p_2 = B$, each of
which may be fitted to the tree.

It is generally advised to provide as much information to the function fit_sbm_parametric as
possible, including reasonable lower and upper bounds (param_min and param_max), a reasonable
parameter guess (param_guess) and reasonable parameter scales param_scale. If some model
parameters can vary over multiple orders of magnitude, it is advised to transform them so that

they vary across fewer orders of magnitude (e.g., via log-transformation). It is also important that the `time_grid` is sufficiently fine to capture the variation of the diffusivity over time, since the likelihood is calculated under the assumption that the diffusivity varies linearly between grid points.

Estimation of diffusivity at older times is only possible if the timetree includes extinct tips or tips sampled at older times (e.g., as is often the case in viral phylogenies). If tips are only sampled once at present-day, i.e. the timetree is ultrametric, reliable diffusivity estimates can only be achieved near present times. If the tree is ultrametric, you should consider using `fit_sbm_const` instead.

For short expected transition distances this function uses the approximation formula by Ghosh et al. (2012) to calculate the probability density of geographical transitions along edges. For longer expected transition distances the function uses a truncated approximation of the series representation of SBM transition densities (Perrin 1928).

If `edge.length` is missing from one of the input trees, each edge in the tree is assumed to have length 1. The tree may include multifurcations as well as monofurcations, however multifurcations are internally expanded into bifurcations by adding dummy nodes.

Value

A list with the following elements:

success	Logical, indicating whether the fitting was successful. If `FALSE`, then an additional return variable, `error`, will contain a description of the error; in that case all other return variables may be undefined.
objective_value	The maximized fitting objective. Currently, only maximum-likelihood estimation is implemented, and hence this will always be the maximized log-likelihood.
objective_name	The name of the objective that was maximized during fitting. Currently, only maximum-likelihood estimation is implemented, and hence this will always be "loglikelihood".
param_fitted	Numeric vector of size NP (number of model parameters), listing all fitted or fixed model parameters in their standard order (see details above).
loglikelihood	The log-likelihood of the fitted model for the given data.
NFP	Integer, number of fitted (i.e., non-fixed) model parameters.
Ncontrasts	Integer, number of independent contrasts used for fitting.
phylodistances	Numeric vector of length Ncontrasts, listing phylogenetic (patristic) distances of the independent contrasts.
geodistances	Numeric vector of length Ncontrasts, listing geographic (great circle) distances of the independent contrasts.
child_times1	Numeric vector of length Ncontrasts, listing the times (distance from root) of the first tip in each independent contrast.
child_times2	Numeric vector of length Ncontrasts, listing the times (distance from root) of the second tip in each independent contrast.
MRCA_times	Numeric vector of length Ncontrasts, listing the times (distance from root) of the MRCA of the two tips in each independent contrast.
AIC	The Akaike Information Criterion for the fitted model, defined as $2k - 2\log(L)$, where k is the number of fitted parameters and L is the maximized likelihood.

BIC	The Bayesian information criterion for the fitted model, defined as $\log(n)k - 2\log(L)$, where k is the number of fitted parameters, n is the number of data points (number of independent contrasts), and L is the maximized likelihood.
converged	Logical, specifying whether the maximum likelihood was reached after convergence of the optimization algorithm. Note that in some cases the maximum likelihood may have been achieved by an optimization path that did not yet converge (in which case it's advisable to increase `iter.max` and/or `eval.max`).
Niterations	Integer, specifying the number of iterations performed during the optimization path that yielded the maximum likelihood.
Nevaluations	Integer, specifying the number of likelihood evaluations performed during the optimization path that yielded the maximum likelihood.

guess_loglikelihood

> The loglikelihood of the data for the initial parameter guess (`param_guess`).

focal_loglikelihoods

> A numeric vector of the same size as `nrow(focal_param_values)`, listing loglikelihoods for each of the focal parameter conbinations listed in `focal_loglikelihoods`.

trial_start_objectives

> Numeric vector of size `Ntrials`, listing the initial objective values (e.g., loglikelihoods) for each fitting trial, i.e. at the start parameter values.

trial_objective_values

> Numeric vector of size `Ntrials`, listing the final maximized objective values (e.g., loglikelihoods) for each fitting trial.

trial_Nstart_attempts

> Integer vector of size `Ntrials`, listing the number of start attempts for each fitting trial, until a starting point with valid likelihood was found.

trial_Niterations

> Integer vector of size `Ntrials`, listing the number of iterations needed for each fitting trial.

trial_Nevaluations

> Integer vector of size `Ntrials`, listing the number of likelihood evaluations needed for each fitting trial.

standard_errors

> Numeric vector of size NP, estimated standard error of the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

medians	Numeric vector of size NP, median the estimated parameters across parametric bootstraps. Only returned if `Nbootstraps>0`.
CI50lower	Numeric vector of size NP, lower bound of the 50% confidence interval (25-75% percentile) for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.
CI50upper	Numeric vector of size NP, upper bound of the 50% confidence interval for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.
CI95lower	Numeric vector of size NP, lower bound of the 95% confidence interval (2.5-97.5% percentile) for the parameters, based on parametric bootstrapping. Only returned if `Nbootstraps>0`.

CI95upper	Numeric vector of size NP, upper bound of the 95% confidence interval for the parameters, based on parametric bootstrapping. Only returned if Nbootstraps>0.
consistency	Numeric between 0 and 1, estimated consistency of the data with the fitted model. See the documentation of fit_sbm_const for an explanation.
QQplot	Numeric matrix of size Ncontrasts x 2, listing the computed QQ-plot. The first column lists quantiles of geodistances in the original dataset, the 2nd column lists quantiles of hypothetical geodistances simulated based on the fitted model.
SBM_PD_functor	SBM probability density functor object. Used internally for efficiency and for debugging purposes.

Author(s)

Stilianos Louca

References

F. Perrin (1928). Etude mathematique du mouvement Brownien de rotation. 45:1-51.

D. R. Brillinger (2012). A particle migrating randomly on a sphere. in Selected Works of David Brillinger. Springer.

A. Ghosh, J. Samuel, S. Sinha (2012). A Gaussian for diffusion on the sphere. Europhysics Letters. 98:30003.

S. Louca (2021). Phylogeographic estimation and simulation of global diffusive dispersal. Systematic Biology. 70:340-359.

See Also

simulate_sbm, fit_sbm_const, fit_sbm_linear

Examples

```
## Not run:
# generate a random tree, keeping extinct lineages
tree_params = list(birth_rate_factor=1, death_rate_factor=0.95)
tree = generate_random_tree(tree_params,max_tips=1000,coalescent=FALSE)$tree

# calculate max distance of any tip from the root
max_time = get_tree_span(tree)$max_distance

# simulate time-dependent SBM on the tree
# we assume that diffusivity varies linearly with time
# in this example we measure distances in Earth radii
radius = 1
diffusivity_functor = function(times, params){
return(params[1] + (times/max_time)*(params[2]-params[1]))
}
true_params = c(1, 2)
time_grid   = seq(0,max_time,length.out=2)
simulation  = simulate_sbm(tree,
                    radius      = radius,
```

```
                      diffusivity = diffusivity_functor(time_grid,true_params),
                      time_grid   = time_grid)

# fit time-independent SBM to get a rough estimate
fit_const = fit_sbm_const(tree,simulation$tip_latitudes,simulation$tip_longitudes,radius=radius)

# fit time-dependent SBM, i.e. fit the 2 parameters of the linear form
fit = fit_sbm_parametric(tree,
            simulation$tip_latitudes,
            simulation$tip_longitudes,
            radius = radius,
            param_values = c(NA,NA),
            param_guess = c(fit_const$diffusivity,fit_const$diffusivity),
            diffusivity = diffusivity_functor,
            time_grid = time_grid,
            Ntrials = 10)

# compare fitted & true params
print(true_params)
print(fit$param_fitted)

## End(Not run)
```

---

fit_tree_model   *Fit a cladogenic model to an existing tree.*

Description

Fit the parameters of a tree generation model to an existing phylogenetic tree; branch lengths are assumed to be in time units. The fitted model is a stochastic cladogenic process in which speciations (births) and extinctions (deaths) are Poisson processes, as simulated by the function [generate_random_tree](). The birth and death rates of tips can each be constant or power-law functions of the number of extant tips. For example,

$$B = I + F \cdot N^E,$$

where B is the birth rate, I is the intercept, F is the power-law factor, N is the current number of extant tips and E is the power-law exponent. Each of the parameters I, F, E can be fixed or fitted.

Fitting can be performed via maximum-likelihood estimation, based on the waiting times between subsequent speciation and/or extinction events represented in the tree. Alternatively, fitting can be performed using least-squares estimation, based on the number of lineages represented in the tree over time ("diversity-vs-time" curve, a.k.a. "lineages-through-time"" curve). Note that the birth and death rates are NOT per-capita rates, they are absolute rates of species appearance and disappearance per time.

Usage

```
fit_tree_model( tree,
                parameters          = list(),
```

```
first_guess        = list(),
min_age            = 0,
max_age            = 0,
age_centile        = NULL,
Ntrials            = 1,
Nthreads           = 1,
coalescent         = FALSE,
discovery_fraction = NULL,
fit_control        = list(),
min_R2             = -Inf,
min_wR2            = -Inf,
grid_size          = 100,
max_model_runtime  = NULL,
objective          = 'LL')
```

Arguments

tree
: A phylogenetic tree, in which branch lengths are assumed to be in time units. The tree may be a coalescent tree (i.e. only include extant clades) or a tree including extinct clades; the tree type influences what type of models can be fitted with each method.

parameters
: A named list specifying fixed and/or unknown birth-death model parameters, with one or more of the following elements:

  - birth_rate_intercept: Non-negative number. The intercept of the Poissonian rate at which new species (tips) are added. In units 1/time.
  - birth_rate_factor: Non-negative number. The power-law factor of the Poissonian rate at which new species (tips) are added. In units 1/time.
  - birth_rate_exponent: Numeric. The power-law exponent of the Poissonian rate at which new species (tips) are added. Unitless.
  - death_rate_intercept: Non-negative number. The intercept of the Poissonian rate at which extant species (tips) go extinct. In units 1/time.
  - death_rate_factor: Non-negative number. The power-law factor of the Poissonian rate at which extant species (tips) go extinct. In units 1/time.
  - death_rate_exponent: Numeric. The power-law exponent of the Poissonian rate at which extant species (tips) go extinct. Unitless.
  - resolution: Numeric. Resolution at which the tree was collapsed (i.e. every node of age smaller than this resolution replaced by a single tip). In units time. A resolution of 0 means the tree was not collapsed.
  - rarefaction: Numeric. Species sampling fraction, i.e. fraction of extant species represented (as tips) in the tree. A rarefaction of 1, for example, implies that the tree is complete, i.e. includes all extant species. Rarefaction is assumed to have occurred after collapsing.
  - extant_diversity: The current total extant diversity, regardless of the rarefaction and resolution of the tree at hand. For example, if resolution==0 and rarefaction==0.5 and the tree has 1000 tips, then extant_diversity should be 2000. If resolution is fixed at 0 and rarefaction is also fixed, this can be left NULL and will be inferred automatically by the function.

Each of the above elements can also be NULL, in which case the parameter is fitted. Elements can also be vectors of size 2 (specifying constraint intervals), in which case the parameters are fitted and constrained within the intervals specified. For example, to fit death_rate_factor while constraining it to the interval [1,2], set its value to c(1,2).

first_guess   A named list (with entries named as in parameters) specifying starting values for any of the fitted model parameters. Note that if Ntrials>1, then start values may be randomly modified in all but the first trial. For any parameters missing from first_guess, initial values are always randomly chosen. first_guess can also be NULL.

min_age   Numeric. Minimum distance from the tree crown, for a node/tip to be considered in the fitting. If <=0 or NULL, this constraint is ignored. Use this option to omit most recent nodes.

max_age   Numeric. Maximum distance from the tree crown, for a node/tip to be considered in the fitting. If <=0 or NULL, this constraint is ignored. Use this option to omit old nodes, e.g. with highly uncertain placements.

age_centile   Numeric within 0 and 1. Fraction of youngest nodes/tips to consider for the fitting. This can be used as an alternative to max_age. E.g. if set to 0.6, then the 60% youngest nodes/tips are considered. Either age_centile or max_age must be non-NULL, but not both.

Ntrials   Integer. Number of fitting attempts to perform, each time using randomly varied start values for fitted parameters. The returned fitted parameter values will be taken from the trial with greatest achieved fit objective. A larger number of trials will decrease the chance of hitting a local non-global optimum during fitting.

Nthreads   Number of threads to use for parallel execution of multiple fitting trials. On Windows, this option has no effect because Windows does not support forks.

coalescent   Logical, specifying whether the input tree is a coalescent tree (and thus the coalescent version of the model should be fitted). Only available if objective=='R2'.

discovery_fraction
                     Function handle, mapping age to the fraction of discovered lineages in a tree. That is, discovery_fraction(tau) is the probability that a lineage at age tau, that has an extant descendant today, will be represented (discovered) in the coalescent tree. In particular, discovery_fraction(0) equals the fraction of extant lineages represented in the tree. If this is provided, then parameters$rarefaction is fixed to 1, and discovery_fraction is applied after simulation. Only relevant if coalescent==TRUE. Experimental, so leave this NULL if you don't know what it means.

fit_control   Named list containing options for the stats::nlminb optimization routine, such as eval.max (max number of evaluations), iter.max (max number of iterations) and rel.tol (relative tolerance for convergence).

min_R2   Minimum coefficient of determination of the diversity curve (clade counts vs time) of the model when compared to the input tree, for a fitted model to be accepted. For example, if set to 0.5 then only fit trials achieving an R2 of at least 0.5 will be considered. Set this to -Inf to not filter fitted models based on the R2.

min_wR2	Similar to min_R2, but applying to the weighted R2, where squared-error weights are proportional to the inverse squared diversities.
grid_size	Integer. Number of equidistant time points to consider when calculating the R2 of a model's diversity-vs-time curve.
max_model_runtime	
	Numeric. Maximum runtime (in seconds) allowed for each model evaluation during fitting. Use this to escape from badly parameterized models during fitting (this will likely cause the affected fitting trial to fail). If NULL or <=0, this option is ignored.
objective	Character. Objective function to optimize during fitting. Can be either "LL" (log-likelihood of waiting times between speciation events and between extinction events), "R2" (coefficient of determination of diversity-vs-time curve), "wR2" (weighted R2, where weights of squared errors are proportional to the inverse diversities observed in the tree) or "lR2" (logarithmic R2, i.e. R2 calculated for the logarithm of the diversity-vs-time curve). Note that "wR2" will weight errors at lower diversities more strongly than "R2".

Value

A named list with the following elements:

success	Logical, indicating whether the fitting was successful.
objective_value	
	Numeric. The achieved maximum value of the objective function (log-likelihood, R2 or weighted R2).
parameters	A named list listing all model parameters (fixed and fitted).
start_parameters	
	A named list listing the start values of all model parameters. In the case of multiple fitting trials, this will list the initial (non-randomized) guess.
R2	Numeric. The achieved coefficient of determination of the fitted model, based on the diversity-vs-time curve.
wR2	Numeric. The achieved weighted coefficient of determination of the fitted model, based on the diversity-vs-time curve. Weights of squared errors are proportional to the inverse squared diversities observed in the tree.
lR2	Numeric. The achieved coefficient of determination of the fitted model on a log axis, i.e. based on the logarithm of the diversity-vs-time curve.
Nspeciations	Integer. Number of speciation events (=nodes) considered during fitting. This only includes speciations visible in the tree.
Nextinctions	Integer. Number of extinction events (=non-crown tips) considered during fitting. This only includes extinctions visible in the tree, i.e. tips whose distance from the root is lower than the maximum.
grid_times	Numeric vector. Time points considered for the diversity-vs-time curve. Times will be constrained between min_age and max_age if these were specified.
tree_diversities	
	Number of lineages represented in the tree through time, calculated for each of grid_times.

model_diversities

> Number of lineages through time as predicted by the model (in the deterministic limit), calculated for each of grid_times. If coalescent==TRUE then these are the number of lineages expected to be represented in the coalescent tree (this may be lower than the actual number of extant clades at any given time point, if the model includes extinctions).

fitted_parameter_names

> Character vector, listing the names of fitted (i.e. non-fixed) parameters.

locally_fitted_parameters

> Named list of numeric vectors, listing the fitted values for each parameter and for each fitting trial. For example, if birth_rate_factor was fitted, then locally_fitted_parameters$birth_rate_factor will be a numeric vector of size Ntrials (or less, if some trials failed or omitted), listing the locally-optimized values of the parameter for each considered fitting trial. Mainly useful for diagnostic purposes.

objective   Character. The name of the objective function used for fitting ("LL", "R2" or "wR2").

Ntips   The number of tips in the input tree.

Nnodes   The number of nodes in the input tree.

min_age   The minimum age of nodes/tips considered during fitting.

max_age   The maximum age of nodes/tips considered during fitting.

age_centile   Numeric or NULL, equal to the age_centile specified as input to the function.

Author(s)

Stilianos Louca

See Also

[generate_random_tree](), [simulate_diversification_model]() [reconstruct_past_diversification]()

Examples

```
# Generate a tree using a simple speciation model
parameters = list(birth_rate_intercept  = 1,
                  birth_rate_factor     = 0,
                  birth_rate_exponent   = 0,
                  death_rate_intercept  = 0,
                  death_rate_factor     = 0,
                  death_rate_exponent   = 0,
                  resolution            = 0,
                  rarefaction           = 1)
tree = generate_random_tree(parameters, max_tips=100)

# Fit model to the tree
fitting_parameters = parameters
fitting_parameters$birth_rate_intercept = NULL # fit only this parameter
fitting = fit_tree_model(tree,fitting_parameters)
```

```
# compare fitted to true value
T = parameters$birth_rate_intercept
F = fitting$parameters$birth_rate_intercept
cat(sprintf("birth_rate_intercept: true=%g, fitted=%g\n",T,F))
```

---

gamma_statistic	*Calculate the gamma-statistic of a tree.*

Description

Given a rooted ultrametric phylogenetic tree, calculate the gamma-statistic (Pybus and Harevy, 2000).

Usage

```
gamma_statistic(tree)
```

Arguments

tree   A rooted tree of class "phylo". The tree is assumed to be ultrametric; any deviations from ultrametricity are ignored.

Details

The tree may include multifurcations and monofurcations. If edge lengths are missing (i.e. edge.length=NULL), then each edge is assumed to have length 1.

This function is similar to the function gammaStat in the R package ape v5.3.

Value

Numeric, the gamma-statistic of the tree.

Author(s)

Stilianos Louca

References

O. G. Pybus and P. H. Harvey (2000). Testing macro-evolutionary models using incomplete molecular phylogenies. Proceedings of the Royal Society of London. Series B: Biological Sciences. 267:2267-2272.

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# calculate & print gamma statistic
gammastat = gamma_statistic(tree)
cat(sprintf("Tree has gamma-statistic %g\n",gammastat))
```

---

generate_gene_tree_msc

Generate a gene tree based on the multi-species coalescent model.

Description

Generate a random gene tree within a given species timetree, based on the multi-species coalescent (MSC) model. In this implementation of the MSC, every branch of the species tree has a specific effective population size (Ne) and a specific generation time (T), and gene alleles coalesce backward in time according to the Wright-Fisher model. This model does not account for gene duplication/loss, nor for hybridization or horizontal gene transfer. It is only meant to model "incomplete lineage sorting", otherwise known as "deep coalescence", which is one of the many mechanisms that can cause discordance between gene trees and species trees.

Usage

```
generate_gene_tree_msc( species_tree,
                        allele_counts           = 1,
                        population_sizes         = 1,
                        generation_times         = 1,
                        mutation_rates           = 1,
                        gene_edge_unit           = "time",
                        Nsites                   = 1,
                        bottleneck_at_speciation = FALSE,
                        force_coalescence_at_root = FALSE,
                        ploidy                   = 1,
                        gene_tip_labels          = NULL)
```

Arguments

species_tree   Rooted timetree of class "phylo". The tree can include multifurcations and
                monofurcations. The tree need not necessarily be ultrametric, i.e. it may in-
                clude extinct species. Edge lengths are assumed to be in time units.

allele_counts   Integer vector, listing the number of alleles sampled per species. Either NULL (1
                allele per species), or a single integer (same number of alleles per species), or
                a vector of length Ntips listing the numbers of alleles sampled per species. In
                the latter case, the total number of tips in the returned gene tree will be equal to
                the sum of entries in allele_counts. Some entries in allele_counts may be
                zero (no alleles sampled from those species).

population_sizes

Integer vector, listing the effective population size on the edge leading into each tip/node in the species tree. Either NULL (all population sizes are 1), or a single integer (same population sizes for all edges), or a vector of length Ntips+Nnodes, listing population sizes for each clade's incoming edge (including the root). The population size for the root's incoming edge corresponds to the population size at the tree's stem (only relevant if force_coalescence_at_root=FALSE).

generation_times

Numeric vector, listing the generation time along the edge leading into each clade. Either NULL (all generation times are 1), or a single integer (same generation time for all edges) or a vector of length Ntips+Nnodes, listing generation times for each clade's incoming edge (including the root). The generation time for the root's incoming edge corresponds to the generation time at the tree's stem (only relevant if force_coalescence_at_root=FALSE).

mutation_rates Numeric vector, listing the mutation rate (per site and per generation) along the edge leading into each clade. Either NULL (all mutation rates are 1), or a single integer (same mutation rate for all edges) or a vector of length Ntips+Nnodes, listing mutation rates for each clade's incoming edge (including the root). The mutation rate for the root's incoming edge corresponds to the mutation rate at the tree's stem (only relevant if force_coalescence_at_root=FALSE). The value of mutation_rates is only relevant if gene_edge_unit is "mutations_expected" or "mutations_random". Mutation rates represent probabilities, and so they must be between 0 and 1.

gene_edge_unit Character, either "time", "generations", "mutations_expected" (expected mean number of mutations per site), or "mutations_random" (randomly generated mean number of mutations per site), specifying how edge lengths in the gene tree should be measured. By default, gene-tree edges are measured in time, as is the case for the input species tree.

Nsites Integer, specifying the number of sites (nucleotides) in the gene. Only relevant when generating edge lengths in terms of random mutation counts, i.e. if gene_edge_unit=="mutations_random".

bottleneck_at_speciation

Logical. If TRUE, then all but one children at each node are assumed to have emerged from a single mutant individual, and thus all gene lineages within these bottlenecked species lineages must coalesce at a younger or equal age as the speciation event. Only the first child at each node is excluded from this assumption, corresponding to the "resident population" during the speciation event. This option deviates from the classical MSC model, and is experimental.

force_coalescence_at_root

Logical. If TRUE, all remaining orphan gene lineages that haven't coalesced before reaching the species-tree's root, will be combined at the root (via multiple adjacent bifurcations). If FALSE, coalescence events may extend beyond the species-tree's root into the stem lineage, as long as it takes until all gene lineages have coalesced.

ploidy Integer, specifying the assumed genetic ploidy, i.e. number of gene copies per individual. Typically 1 for haploids, or 2 for diploids.

gene_tip_labels

> Character vector specifying tip labels for the gene tree (i.e., for each of the sampled alleles) in the order of the corresponding species tips. Can also be NULL, in which case gene tips will be set to <species_tip_label>.<allele index>.

Details

This function assumes that Kingman's coalescent assumption is met, i.e. that the effective population size is much larger than the number of allele lineages coalescing within any given branch.

The function assumes that the species tree is a time tree, i.e. with edge lengths given in actual time units. To simulate gene trees in coalescence time units, choose `population_sizes` and `generation_times` accordingly (this only makes sense if the product of `population_sizes` × `generation_times` is the same everywhere). If `species_tree` is ultrametric and `gene_edge_unit=="time"`, then the gene tree will be ultrametric as well.

If `gene_edge_unit` is "mutations_random", then the number of generations elapsed along each time segment is translated into a randomly distributed number of accumulated mutations, according to a binomial distribution where the probability of success is equal to the mutation rate and the number of trials is equal to the number of generations multiplied by `Nsites`; this number of mutations is averaged across all sites, i.e. the edge lengths in the returned gene tree always refer to the mean number of mutations per site. In cases where the mutation rate varies across the species tree and a single gene edge spans multiple species edges, the gene edge length will be a sum of multiple binomially distributed mutation counts (again, divided by the number of sites), corresponding to the times spent in each species edge.

Value

A named list with the following elements:

success

> Logical, indicating whether the gene tree was successfully generated. If `FALSE`, the only other value returned is `error`.

tree

> The generated gene tree, of class "phylo". This tree will be rooted and bifurcating. It is only guaranteed to be ultrametric if `species_tree` was ultrametric.

gene_tip2species_tip

> Integer vector of length NGtips (where NGtips is the number of tips in the gene tree), mapping gene-tree tips to species-tree tips.

gene_node2species_edge

> Integer vector of length NGnodes (where NGnodes is the number of internal nodes in the gene tree), mapping gene-tree nodes (=coalescence events) to the species-tree edges where the coalescences took place.

gene_clade_times

> Numeric vector of size NGtips+NGnodes, listing the time (total temporal distance from species root) of each tip and node in the gene tree. The units will be the same as the time units assumed for the species tree. Note that this may include negative values, if some gene lineages coalesce at a greater age than the root.

error

> Character, containing an explanation of the error that occurred. Only included if `success==FALSE`.

Author(s)

Stilianos Louca

References

J. H. Degnan, N. A. Rosenberg (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution. 24:332-340.

B. Rannala, Z. Yang (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645-1656.

See Also

generate_random_tree, generate_gene_tree_msc_hgt_dl

Examples

```
# Simulate a simple species tree
parameters   = list(birth_rate_factor=1)
Nspecies     = 10
species_tree = generate_random_tree(parameters,max_tips=Nspecies)$tree

# Simulate a haploid gene tree within the species tree
# Assume the same population size and generation time everywhere
# Assume the number of alleles samples per species is poisson-distributed
results = generate_gene_tree_msc(species_tree,
                                  allele_counts     = rpois(Nspecies,3),
                                  population_sizes   = 1000,
                                  generation_times  = 1,
                                  ploidy            = 1);
if(!results$success){
    # simulation failed
    cat(sprintf(" ERROR: %s\n",results$error))
}else{
    # simulation succeeded
    gene_tree = results$tree
    cat(sprintf(" Gene tree has %d tips\n",length(gene_tree$tip.label)))
}
```

---

generate_gene_tree_msc_hgt_dl

Generate gene trees based on the multi-species coalescent, horizontal gene transfers and duplications/losses.

Description

Generate a random gene tree within a given species timetree, based on an extension of the multi-species coalescent (MSC) model that includes horizontal gene transfers (HGT, incorporation of non-homologous genes as new loci), gene duplication and gene loss. The simulation consists of

two phases. In the first phase a random "locus tree" is generated in forward time, according to random HGT, duplication and loss events. In the 2nd phase, alleles picked randomly from each locus are coalesced in backward time according to the multispecies coalescent, an extension of the Wright-Fisher model to multiple species. This function does not account for hybridization.

Usage

```
generate_gene_tree_msc_hgt_dl( species_tree,
                       allele_counts           = 1,
                       population_sizes         = 1,
                       generation_times         = 1,
                       mutation_rates           = 1,
                       HGT_rates                = 0,
                       duplication_rates        = 0,
                       loss_rates               = 0,
                       gene_edge_unit           = "time",
                       Nsites                   = 1,
                       bottleneck_at_speciation = FALSE,
                       force_coalescence_at_root = FALSE,
                       ploidy                   = 1,
                       HGT_source_by_locus      = FALSE,
                       HGT_only_to_empty_clades = FALSE,
                       no_loss_before_time      = 0,
                       max_runtime              = NULL,
                       include_event_times      = TRUE)
```

Arguments

species_tree   Rooted timetree of class "phylo". The tree can include multifurcations and monofurcations. The tree need not necessarily be ultrametric, i.e. it may include extinct species. Edge lengths are assumed to be in time units.

allele_counts   Integer vector, listing the number of alleles sampled per species and per locus. This can be interpreted as the number if individual organisms surveyed from each species, assuming that all loci are included once from each individual. The number of tips in the generated gene tree will be equal to the sum of allele counts across all species. allele_counts can either be NULL (1 allele per species), or a single integer (same number of alleles per species), or a vector of length Ntips listing the numbers of alleles sampled per species. In the latter case, the total number of tips in the returned gene tree will be equal to the sum of entries in allele_counts. Some entries in allele_counts may be zero (no alleles sampled from those species).

population_sizes

                Integer vector, listing the effective population size on the edge leading into each tip/node in the species tree. Either NULL (all population sizes are 1), or a single integer (same population sizes for all edges), or a vector of length Ntips+Nnodes, listing population sizes for each clade's incoming edge (including the root). The population size for the root's incoming edge corresponds to the population size at the tree's stem (only relevant if force_coalescence_at_root=FALSE).

generation_times

Numeric vector, listing the generation time along the edge leading into each clade. Either `NULL` (all generation times are 1), or a single integer (same generation time for all edges) or a vector of length Ntips+Nnodes, listing generation times for each clade's incoming edge (including the root). The generation time for the root's incoming edge corresponds to the generation time at the tree's stem (only relevant if `force_coalescence_at_root=FALSE`).

mutation_rates Numeric vector, listing the probability of mutation per site and per generation along the edge leading into each clade. Either `NULL` (all mutation rates are 1), or a single integer (same mutation rate for all edges) or a vector of length Ntips+Nnodes, listing mutation rates for each clade's incoming edge (including the root). The mutation rate for the root's incoming edge corresponds to the mutation rate at the tree's stem (only relevant if `force_coalescence_at_root=FALSE`). The value of `mutation_rates` is only relevant if `gene_edge_unit` is "mutations_expected" or "mutations_random". Mutation rates represent probabilities, and so they must be between 0 and 1.

HGT_rates   Numeric vector, listing horizontal gene transfer rates per lineage per time, along the edge leading into each clade. Either `NULL` (all HGT rates are 0) or a single integer (same HGT rate for all edges) or a vector of length Ntips+Nnodes, listing HGT rates for each clade's incoming edge (including the root).

duplication_rates

Numeric vector, listing gene duplication rates per locus per lineage per time, along the edge leading into each clade. Either `NULL` (all duplication rates are 0) or a single integer (same duplication rate for all edges) or a vector of length Ntips+Nnodes listing duplication rates for each clade's incoming edge (including the root).

loss_rates   Numeric vector, listing gene loss rates per locus per lineage per time, along the edge leading into each clade. Either `NULL` (all loss rates are 0) or a single integer (same loss rate for all edges) or a vector of length Ntips+Nnodes listing loss rates for each clade's incoming edge (including the root).

gene_edge_unit Character, either "time", "generations", "mutations_expected" (expected mean number of mutations per site), or "mutations_random" (randomly generated mean number of mutations per site), specifying how edge lengths in the gene tree should be measured. By default, gene-tree edges are measured in time, as is the case for the input species tree.

Nsites   Integer, specifying the number of sites (nucleotides) in the gene. Only relevant when generating edge lengths in terms of random mutation counts, i.e. if `gene_edge_unit=="mutations_random"`.

bottleneck_at_speciation

Logical. If `TRUE`, then all but one children at each node are assumed to have emerged from a single mutant individual, and thus all gene lineages within these bottlenecked species lineages must coalesce at a younger or equal age as the speciation event. Only the first child at each node is excluded from this assumption, corresponding to the "resident population" during the speciation event. This option deviates from the classical MSC model, and is experimental.

force_coalescence_at_root

Logical. If `TRUE`, all remaining orphan gene lineages that haven't coalesced before reaching the species-tree's root, will be combined at the root (via multiple

adjacent bifurcations). If `FALSE`, coalescence events may extend beyond the species-tree's root into the stem lineage, as long as it takes until all gene lineages have coalesced.

ploidy   Integer, specifying the assumed genetic ploidy, i.e. number of gene copies per individual. Typically 1 for haploids, or 2 for diploids.

HGT_source_by_locus

Logical. If `TRUE`, then at any HGT event, every extant locus is chosen as source locus with the same probability (hence the probability of a lineage to be a source is proportional to the number of current loci in it). If `FALSE`, source lineages are chosen with the same probability (regardless of the number of current loci in them) and the source locus within the source lineage is chosen randomly.

HGT_only_to_empty_clades

Logical, specifying whether HGT transfers only occur into clades with no current loci.

no_loss_before_time

Numeric, optional time since the root during which no gene losses shall occur (even if `loss_rate>0`). This option can be used to reduce the probability of an early extinction of the entire gene tree, by giving the gene tree some "startup time" to spread into various species lineages. If zero, gene losses are possible right from the start of the simulation.

max_runtime   Numeric, optional maximum computation time (in seconds) to allow for the simulation. Use this to avoid occasional explosions of runtimes, for example due to very large generated trees. Aborted simulations will return with the flag `success=FALSE` (i.e., no tree is returned at all).

include_event_times

Logical, specifying whether the times of HGT, duplication and loss events should be returned as well. If these are not needed, then set `include_event_times=FALSE` for efficiency.

Details

This function assumes that the species tree is a time tree, i.e. with edge lengths given in actual time units. If `species_tree` is ultrametric and `gene_edge_unit=="time"`, then the gene tree (but not necessarily the locus tree) will be ultrametric as well. The root of the locus and gene tree coincides with the root of the species tree.

The meaning of `gene_edge_unit` is the same as for the function [generate_gene_tree_msc](#).

Value

A named list with the following elements:

success   Logical, indicating whether the gene tree was successfully generated. If `FALSE`, the only other value returned is `error`.

locus_tree   The generated locus timetree, of class "phylo". The locus tree describes the genealogy of loci due to HGT, duplication and loss events. Each tip and node of the locus tree is embedded within a specific species edge. For example, tips of the locus tree either coincide with tips of the species tree (if the locus persisted

until the species went extinct or until the present) or they correspond to gene loss events. In the absence of any HGT, duplication and loss events, the locus tree will resemble the species tree.

locus_type   Character vector of length NLtips + NLnodes (where NLtips and NLnodes are the number of tips and nodes in the locus tree, respectively), specifying the type/origin of each tip and node in the locus tree. For nodes, type 'h' corresponds to an HGT event, type 'd' to a duplication event, and type 's' to a speciation event. For tips, type 'l' represents a loss event, and type 't' a terminal locus (i.e., coinciding with a species tip). For example, if the input species tree was an ultrametric tree representing only extant species, then the locus tree tips of type 't' are the loci that could potentially be sampled from those extant species.

locus2clade   Integer vector of length NLtips + NLnodes, with values in NStips+NSnodes, specifying for every locus tip or node the correspondng "host" species tip or node.

HGT_times   Numeric vector, listing HGT event times (counted since the root) in ascending order. Only included if include_event_times==TRUE.

HGT_source_clades
               Integer vector of the same length as HGT_times and with values in 1,..,Ntips+Nnodes, listing the "source" species tip/node of each HGT event (in order of occurrence). The source tip/node is the tip/node from whose incoming edge a locus originated at the time of the transfer. Only included if include_event_times==TRUE.

HGT_target_clades
               Integer vector of the same length as HGT_times and with values in 1,..,Ntips+Nnodes, listing the "target" species tip/node of each HGT event (in order of occurrence). The target (aka. recipient) tip/node is the tip/node within whose incoming edge a locus was created by the transfer. Only included if include_event_times==TRUE.

duplication_times
               Numeric vector, listing gene duplication event times (counted since the root) in ascending order. Only included if include_event_times==TRUE.

duplication_clades
               Integer vector of the same length as duplication_times and with values in 1,..,Ntips+Nnodes, listing the species tip/node in whose incoming edge each duplication event occurred (in order of occurrence). Only included if include_event_times==TRUE.

loss_times   Numeric vector, listing gene loss event times (counted since the root) in ascending order. Only included if include_event_times==TRUE.

loss_clades   Integer vector of the same length as loss_times and with values in 1,..,Ntips+Nnodes, listing the species tip/node in whose incoming edge each loss event occurred (in order of occurrence). Only included if include_event_times==TRUE.

gene_tree   The generated gene tree, of type "phylo".

gene_tip2species_tip
               Integer vector of length NGtips (where NGtips is the number of tips in the gene tree) with values in 1,..,Ntips+Nnodes, mapping gene-tree tips to species-tree tips.

gene_tip2locus_tip
               Integer vector of length NGtips with values in 1,..,NLtips, mapping gene-tree tips to locus-tree tips.

gene_node2locus_edge

                Integer vector of length NGnodes with values in 1,..,NLedges, mapping gene-tree nodes to locus-tree edges.

gene_clade_times

                Numeric vector of size NGtips+NGnodes, listing the time (temporal distance from species root) of each tip and node in the gene tree. The units will be the same as the time units of the species tree. Note that this may include negative values, if some gene lineages coalesce at a greater age than the root.

error   Character, containing an explanation of the error that occurred. Only included if success==FALSE.

Author(s)

Stilianos Louca

References

J. H. Degnan, N. A. Rosenberg (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution. 24:332-340.

B. Rannala, Z. Yang (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645-1656.

See Also

generate_random_tree, generate_gene_tree_msc

Examples

```
# Simulate a simple species tree
parameters   = list(birth_rate_factor=1)
Nspecies     = 10
species_tree = generate_random_tree(parameters,max_tips=Nspecies)$tree

# Simulate a haploid gene tree within the species tree, including HGTs and gene loss
# Assume the same population size and generation time everywhere
# Assume the number of alleles samples per species is poisson-distributed
results = generate_gene_tree_msc_hgt_dl(species_tree,
                                        allele_counts    = rpois(Nspecies,3),
                                        population_sizes = 1000,
                                        generation_times = 1,
                                        ploidy           = 1,
                                        HGT_rates        = 0.1,
                                        loss_rates       = 0.05);
if(!results$success){
    # simulation failed
    cat(sprintf("  ERROR: %s\n",results$error))
}else{
    # simulation succeeded
    gene_tree = results$gene_tree
    cat(sprintf("  Gene tree has %d tips\n",length(gene_tree$tip.label)))
}
```

---

generate_random_tree   *Generate a tree using a Poissonian speciation/extinction model.*

Description

Generate a random timetree via simulation of a Poissonian speciation/extinction (birth/death) process. New species are added (born) by splitting of a randomly chosen extant tip. The tree-wide birth and death rates of tips can each be constant or power-law functions of the number of extant tips. For example,

$$B = I + F \cdot N^E,$$

where B is the tree-wide birth rate (species generation rate), I is the intercept, F is the power-law factor, N is the current number of extant tips and E is the power-law exponent. Optionally, the per-capita (tip-specific) birth and death rates can be extended by adding a custom time series provided by the user.

Usage

```
generate_random_tree(parameters          = list(),
                     max_tips             = NULL,
                     max_extant_tips      = NULL,
                     max_time             = NULL,
                     max_time_eq          = NULL,
                     coalescent           = TRUE,
                     as_generations       = FALSE,
                     no_full_extinction   = TRUE,
                     Nsplits              = 2,
                     added_rates_times    = NULL,
                     added_birth_rates_pc = NULL,
                     added_death_rates_pc = NULL,
                     added_periodic       = FALSE,
                     tip_basename         = "",
                     node_basename        = NULL,
                     edge_basename        = NULL,
                     include_birth_times  = FALSE,
                     include_death_times  = FALSE)
```

Arguments

parameters   A named list specifying the birth-death model parameters, with one or more of the following entries:

birth_rate_intercept: Non-negative number. The intercept of the Poissonian rate at which new species (tips) are added. In units 1/time. By default this is 0.

birth_rate_factor: Non-negative number. The power-law factor of the Poissonian rate at which new species (tips) are added. In units 1/time. By default this is 0.

birth_rate_exponent: Numeric. The power-law exponent of the Poissonian rate at which new species (tips) are added. Unitless. By default this is 1.

death_rate_intercept: Non-negative number. The intercept of the Poissonian rate at which extant species (tips) go extinct. In units 1/time. By default this is 0.

death_rate_factor: Non-negative number. The power-law factor of the Poissonian rate at which extant species (tips) go extinct. In units 1/time. By default this is 0.

death_rate_exponent: Numeric. The power-law exponent of the Poissonian rate at which extant species (tips) go extinct. Unitless. By default this is 1.

resolution: Non-negative numeric, specifying the resolution (in time units) at which to collapse the final tree by combining closely related tips. Any node whose age is smaller than this threshold, will be represented by a single tip. Set resolution=0 to not collapse tips (default).

rarefaction: Numeric between 0 and 1. Rarefaction to be applied to the final tree (fraction of random tips kept in the tree). Note that if coalescent==FALSE, rarefaction may remove both extant as well as extinct clades. Set rarefaction=1 to not perform any rarefaction (default).

max_tips   Integer, maximum number of tips of the tree to be generated. If coalescent=TRUE, this refers to the number of extant tips. Otherwise, it refers to the number of extinct + extant tips. If NULL or <=0, this halting condition is ignored.

max_extant_tips
                      Integer, maximum number of extant lineages allowed at any moment during the simulation. If this number is reached, the simulation is halted. If NULL or <=0, this halting condition is ignored.

max_time   Numeric, maximum duration of the simulation. If NULL or <=0, this constraint is ignored.

max_time_eq   Maximum duration of the simulation, counting from the first point at which speciation/extinction equilibrium is reached, i.e. when (birth rate - death rate) changed sign for the first time. If NULL or <0, this constraint is ignored.

coalescent   Logical, specifying whether only the coalescent tree (i.e. the tree spanning the extant tips) should be returned. If coalescent==FALSE and the death rate is non-zero, then the tree may include non-extant tips (i.e. tips whose distance from the root is less than the total time of evolution). In that case, the tree will not be ultrametric.

as_generations   Logical, specifying whether edge lengths should correspond to generations. If FALSE, then edge lengths correspond to time.

no_full_extinction
                      Logical, specifying whether to prevent complete extinction of the tree. Full extinction is prevented by temporarily disabling extinctions whenever the number of extant tips is 1. Note that, strictly speaking, the trees generated do not exactly follow the proper probability distribution when no_full_extinction is TRUE.

Nsplits   Integer greater than 1. Number of child-tips to generate at each diversification event. If set to 2, the generated tree will be bifurcating. If >2, the tree will be multifurcating.

added_rates_times

> Numeric vector, listing time points (in ascending order) for the custom per-capita birth and/or death rates time series (see added_birth_rates_pc and added_death_rates_pc below). Can also be NULL, in which case the custom time series are ignored.

added_birth_rates_pc

> Numeric vector of the same size as added_rates_times, listing per-capita birth rates to be added to the power law part. Can also be NULL, in which case this option is ignored and birth rates are purely described by the power law.

added_death_rates_pc

> Numeric vector of the same size as added_rates_times, listing per-capita death rates to be added to the power law part. Can also be NULL, in which case this option is ignored and death rates are purely described by the power law.

added_periodic   Logical, indicating whether added_birth_rates_pc and added_death_rates_pc should be extended periodically if needed (i.e. if not defined for the entire simulation time). If FALSE, added birth & death rates are extended with zeros.

tip_basename   Character. Prefix to be used for tip labels (e.g. "tip."). If empty (""), then tip labels will be integers "1", "2" and so on.

node_basename   Character. Prefix to be used for node labels (e.g. "node."). If NULL, no node labels will be included in the tree.

edge_basename   Character. Prefix to be used for edge labels (e.g. "edge."). Edge labels (if included) are stored in the character vector edge.label. If NULL, no edge labels will be included in the tree.

include_birth_times

> Logical. If TRUE, then the times of speciation events (in order of occurrence) will also be returned.

include_death_times

> Logical. If TRUE, then the times of extinction events (in order of occurrence) will also be returned.

Details

If max_time==NULL, then the returned tree will always contain max_tips tips. In particular, if at any moment during the simulation the tree only includes a single extant tip, the death rate is temporarily set to zero to prevent the complete extinction of the tree. If max_tips==NULL, then the simulation is ran as long as specified by max_time. If neither max_time nor max_tips is NULL, then the simulation halts as soon as the time exceeds max_time or the number of tips (extant tips if coalescent is TRUE) exceeds max_tips. If max_tips!=NULL and Nsplits>2, then the last diversification even may generate fewer than Nsplits children, in order to keep the total number of tips within the specified limit.

If rarefaction<1 and resolution>0, collapsing of closely related tips (at the resolution specified) takes place prior to rarefaction (i.e., subsampling applies to the already collapsed tips).

Both the per-capita birth and death rates can be made into completely arbitrary functions of time, by setting all power-law coefficients to zero and providing custom time series added_birth_rates_pc and added_death_rates_pc.

Value

A named list with the following elements:

success	Logical, indicating whether the tree was successfully generated. If FALSE, the only other value returned is error.
tree	A rooted bifurcating (if Nsplits==2) or multifurcating (if Nsplits>2) tree of class "phylo", generated according to the specified birth/death model. If coalescent==TRUE or if all death rates are zero, and only if as_generations==FALSE, then the tree will be ultrametric. If as_generations==TRUE and coalescent==FALSE, all edges will have unit length.
root_time	Numeric, giving the time at which the tree's root was first split during the simulation. Note that if coalescent==TRUE, this may be later than the first speciation event during the simulation.
final_time	Numeric, giving the final time at the end of the simulation. Note that if coalescent==TRUE, then this may be greater than the total time span of the tree (since the root of the coalescent tree need not correspond to the first speciation event).
root_age	Numeric, giving the age (time before present) at the tree's root. This is equal to final_time-root_time.
equilibrium_time	
	Numeric, giving the first time where the sign of (death rate - birth rate) changed from the beginning of the simulation, i.e. when speciation/extinction equilibrium was reached. May be infinite if the simulation stoped before reaching this point.
extant_tips	Integer vector, listing indices of extant tips in the tree. If coalescent==TRUE, all tips will be extant.
Nbirths	Total number of birth events (speciations) that occurred during tree growth. This may be lower than the total number of tips in the tree if death rates were non-zero and coalescent==TRUE, or if Nsplits>2.
Ndeaths	Total number of deaths (extinctions) that occurred during tree growth.
Ncollapsed	Number of tips removed from the tree while collapsing at the resolution specified.
Nrarefied	Number of tips removed from the tree due to rarefaction.
birth_times	Numeric vector, listing the times of speciation events during tree growth, in order of occurrence. Note that if coalescent==TRUE, then speciation_times may be greater than the phylogenetic distance to the coalescent root.
death_times	Numeric vector, listing the times of extinction events during tree growth, in order of occurrence. Note that if coalescent==TRUE, then speciation_times may be greater than the phylogenetic distance to the coalescent root.
error	Character, containing an explanation of ther error that occurred. Only included if success==FALSE.

Author(s)

Stilianos Louca

References

D. J. Aldous (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statistical Science. 16:23-34.

M. Steel and A. McKenzie (2001). Properties of phylogenetic trees generated by Yule-type speciation models. Mathematical Biosciences. 170:91-112.

Examples

```
# Simple speciation model
parameters = list(birth_rate_intercept=1)
tree = generate_random_tree(parameters,max_tips=100)$tree

# Exponential growth rate model
parameters = list(birth_rate_factor=1)
tree = generate_random_tree(parameters,max_tips=100)$tree
```

---

generate_tree_hbds   *Generate a tree from a birth-death-sampling model in forward time.*

Description

Generate a random timetree according to a homogenous birth-death-sampling model with arbitrary time-varying speciation/extinction/sampling rates. Lineages split (speciate) or die (go extinct) at Poissonian rates and independently of each other. Lineages are sampled continuously (i.e., at Poissonian rates) in time and/or during concentrated sampling attempts (i.e., at specific time points). Sampled lineages are assumed to continue in the pool of extant lineages at some given "retention probability". The final tree can be restricted to sampled lineages only, but may optionally include extant (non-sampled) as well as extinct lineages. Speciation, extinction and sampling rates as well as retention probabilities may depend on time. This function may be used to simulate trees commonly encountered in viral epidemiology, where sampled patients are assumed to exit the pool of infectious individuals.

Usage

```
generate_tree_hbds( max_sampled_tips          = NULL,
                    max_sampled_nodes         = NULL,
                    max_extant_tips           = NULL,
                    max_extinct_tips          = NULL,
                    max_tips                  = NULL,
                    max_time                  = NULL,
                    include_extant            = FALSE,
                    include_extinct           = FALSE,
                    as_generations            = FALSE,
                    time_grid                 = NULL,
                    lambda                    = NULL,
                    mu                        = NULL,
                    psi                       = NULL,
```

```
                        kappa                       = NULL,
                        splines_degree              = 1,
                        CSA_times                   = NULL,
                        CSA_probs                   = NULL,
                        CSA_kappas                  = NULL,
                        no_full_extinction          = FALSE,
                        max_runtime                 = NULL,
                        tip_basename                = "",
                        node_basename               = NULL,
                        edge_basename               = NULL,
                        include_birth_times         = FALSE,
                        include_death_times         = FALSE)
```

Arguments

max_sampled_tips

Integer, maximum number of sampled tips. The simulation is halted once this number is reached. If NULL or <=0, this halting criterion is ignored.

max_sampled_nodes

Integer, maximum number of sampled nodes, i.e., of lineages that were sampled but kept in the pool of extant lineages. The simulation is halted once this number is reached. If NULL or <=0, this halting criterion is ignored.

max_extant_tips

Integer, maximum number of extant tips. The simulation is halted once the number of concurrently extant tips reaches this threshold. If NULL or <=0, this halting criterion is ignored.

max_extinct_tips

Integer, maximum number of extant tips. The simulation is halted once this number is reached. If NULL or <=0, this halting criterion is ignored.

max_tips   Integer, maximum number of tips (extant+extinct+sampled). The simulation is halted once this number is reached. If NULL or <=0, this halting criterion is ignored.

max_time   Numeric, maximum duration of the simulation. If NULL or <=0, this halting criterion is ignored.

include_extant   Logical, specifying whether to include extant tips (i.e., neither extinct nor sampled) in the final tree.

include_extinct

Logical, specifying whether to include extant tips (i.e., neither extant nor sampled) in the final tree.

as_generations   Logical, specifying whether edge lengths should correspond to generations. If FALSE, then edge lengths correspond to time. If TRUE, then the time between two subsequent events (speciation, extinction, sampling) is counted as "one generation".

time_grid   Numeric vector, specifying time points (in ascending order) on which the rates lambda, mu and psi are provided. Rates are interpolated polynomially between time grid points as needed (according to splines_degree). The time grid should

generally cover the maximum possible simulation time, otherwise it will be polynomially extrapolated as needed.

lambda   Numeric vector, of the same size as time_grid (or size 1 if time_grid==NULL), listing per-lineage speciation (birth) rates (λ, in units 1/time) at the times listed in time_grid. Speciation rates must be non-negative, and are assumed to vary as a spline between grid points (see argument splines_degree). Can also be a single numeric, in which case λ is assumed to be constant over time.

mu   Numeric vector, of the same size as time_grid (or size 1 if time_grid==NULL), listing per-lineage extinction (death) rates (μ, in units 1/time) at the times listed in time_grid. Extinction rates must be non-negative, and are assumed to vary as a spline between grid points (see argument splines_degree). Can also be a single numeric, in which case μ is assumed to be constant over time. If omitted, the extinction rate is assumed to be zero.

psi   Numeric vector, of the same size as time_grid (or size 1 if time_grid==NULL), listing per-lineage sampling rates (ψ, in units 1/time) at the times listed in time_grid. Sampling rates must be non-negative, and are assumed to vary as a spline between grid points (see argument splines_degree). Can also be a single numeric, in which case ψ is assumed to be constant over time. If omitted, the continuous sampling rate is assumed to be zero.

kappa   Numeric vector, of the same size as time_grid (or size 1 if time_grid==NULL), listing retention probabilities (κ, unitless) of continuously (Poissonian) sampled lineages at the times listed in time_grid. Retention probabilities must be true probabilities (i.e., between 0 and 1), and are assumed to vary as a spline between grid points (see argument splines_degree). Can also be a single numeric, in which case κ is assumed to be constant over time. If omitted, the retention probability is assumed to be zero (a common assumption in epidemiology).

splines_degree   Integer, either 0,1,2 or 3, specifying the polynomial degree of the provided lambda, mu and psi between grid points in age_grid. For example, if splines_degree==1, then the provided lambda, mu and psi are interpreted as piecewise-linear curves; if splines_degree==2 the lambda, mu and psi are interpreted as quadratic splines; if splines_degree==3 the lambda, mu and psi is interpreted as cubic splines. If your age_grid is fine enough, then splines_degree=1 is usually sufficient.

CSA_times   Optional numeric vector, listing times of concentrated sampling attempts, in ascending order. Concentrated sampling is performed in addition to any continuous (Poissonian) sampling specified by psi.

CSA_probs   Optional numeric vector of the same size as CSA_times, listing sampling probabilities at each concentrated sampling time. Note that in contrast to the sampling rates psi, the CSA_probs are interpreted as probabilities and must thus be between 0 and 1. CSA_probs must be provided if and only if CSA_times is provided.

CSA_kappas   Optional numeric vector of the same size as CSA_times, listing sampling retention probabilities at each concentrated sampling time, i.e. the probability at which a sampled lineage is kept in the pool of extant lineages. Note that the CSA_kappas are probabilities and must thus be between 0 and 1. CSA_kappas must be provided if and only if CSA_times is provided.

no_full_extinction

Logical, specifying whether to prevent complete extinction of the tree. Full extinction is prevented by temporarily disabling extinctions whenever the number of extant tips is 1. Note that, strictly speaking, the trees generated do not exactly follow the proper probability distribution when no_full_extinction is TRUE.

max_runtime   Numeric, optional maximum computation time (in seconds) to allow for the simulation. Use this to avoid occasional explosions of runtimes, for example due to very large generated trees. Aborted simulations will return with the flag success=FALSE (i.e., no tree is returned at all).

tip_basename   Character. Prefix to be used for tip labels (e.g. "tip."). If empty (""), then tip labels will be integers "1", "2" and so on.

node_basename   Character. Prefix to be used for node labels (e.g. "node."). If NULL, no node labels will be included in the tree.

edge_basename   Character. Prefix to be used for edge labels (e.g. "edge."). Edge labels (if included) are stored in the character vector edge.label. If NULL, no edge labels will be included in the tree.

include_birth_times

Logical. If TRUE, then the times of speciation events (in order of occurrence) will also be returned.

include_death_times

Logical. If TRUE, then the times of extinction events (in order of occurrence) will also be returned.

Details

The simulation proceeds in forward time, starting with a single root. Speciation/extinction and continuous (Poissonian) sampling events are drawn at exponentially distributed time steps, according to the rates specified by lambda, mu and psi. Sampling also occurs at the optional CSA_times. Only extant lineages are sampled at any time point, and sampled lineages are removed from the pool of extant lineages at probability 1-kappa.

The simulation halts as soon as one of the halting criteria are met, as specified by the options max_sampled_tips, max_sampled_nodes, max_extant_tips, max_extinct_tips, max_tips and max_time, or if no extant tips remain, whichever occurs first. Note that in some scenarios (e.g., if extinction rates are very high) the simulation may halt too early and the generated tree may only contain a single tip (i.e., the root lineage); in that case, the simulation will return an error (see return value success).

The function returns a single generated tree, as well as supporting information such as which tips are extant, extinct or sampled.

Value

A named list with the following elements:

success   Logical, indicating whether the simulation was successful. If FALSE, then the returned list includes an additional 'error' element (character) providing a description of the error; all other return variables may be undefined.

tree   The generated timetree, of class "phylo". Note that this tree need not be ultra-metric, for example if sampling occurs at multiple time points.

root_time	Numeric, giving the time at which the tree's root was first split during the simulation. Note that this may be greater than 0, i.e., if the tips of the final tree do not coalesce all the way back to the simulation's start.
final_time	Numeric, giving the final time at the end of the simulation.
root_age	Numeric, giving the age (time before present) at the tree's root. This is equal to final_time-root_time.
Nbirths	Integer, the total number of speciation (birth) events that occured during the simulation.
Ndeaths	Integer, the total number of extinction (death) events that occured during the simulation.
Nsamplings	Integer, the total number of sampling events that occured during the simulation.
Nretentions	Integer, the total number of sampling events that occured during the simulation and for which lineages were kept in the pool of extant lineages.
sampled_clades	Integer vector, specifying indices (from 1 to Ntips+Nnodes) of sampled tips and nodes in the final tree (regardless of whether their lineages were subsequently retained or removed from the pool).
retained_clades	
	Integer vector, specifying indices (from 1 to Ntips+Nnodes) of sampled tips and nodes in the final tree that were retained, i.e., not removed from the pool following sampling.
extant_tips	Integer vector, specifying indices (from 1 to Ntips) of extant (non-sampled and non-extinct) tips in the final tree. Will be empty if include_extant==FALSE.
extinct_tips	Integer vector, specifying indices (from 1 to Ntips) of extinct (non-sampled and non-extant) tips in the final tree. Will be empty if include_extinct==FALSE.

Author(s)

Stilianos Louca

References

T. Stadler (2010). Sampling-through-time in birth–death trees. Journal of Theoretical Biology. 267:396-404.

T. Stadler et al. (2013). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). PNAS. 110:228-233.

See Also

generate_tree_hbd_reverse, generate_gene_tree_msc, generate_random_tree, fit_hbds_model_parametric, simulate_deterministic_hbds

Examples

```
# define time grid on which lambda, mu and psi will be specified
time_grid = seq(0,100,length.out=1000)
```

```
# specify the time-dependent extinction rate mu on the time-grid
mu_grid = 0.5*time_grid/(10+time_grid)

# define additional concentrated sampling attempts
CSA_times  = c(5,7,9)
CSA_probs  = c(0.5, 0.5, 0.5)
CSA_kappas = c(0.2, 0.1, 0.1)

# generate tree with a constant speciation & sampling rate,
# time-variable extinction rate and additional discrete sampling points
# assuming that all continuously sampled lineages are removed from the pool
simul = generate_tree_hbds( max_time       = 10,
                            include_extant  = FALSE,
                            include_extinct = FALSE,
                            time_grid       = time_grid,
                            lambda          = 1,
                            mu              = mu_grid,
                            psi             = 0.1,
                            kappa           = 0,
                            CSA_times       = CSA_times,
                            CSA_probs       = CSA_probs,
                            CSA_kappas      = CSA_kappas);
if(!simul$success){
    cat(sprintf("ERROR: Could not simulate tree: %s\n",simul$error))
}else{
    # simulation succeeded. print some basic info about the generated tree
    tree = simul$tree
    cat(sprintf("Generated tree has %d tips\n",length(tree$tip.label)))
}
```

---

generate_tree_hbd_reverse

Generate a tree from a birth-death model in reverse time.

Description

Generate an ultrametric timetree (comprising only extant lineages) in reverse time (from present back to the root) based on the homogenous birth-death (HBD; Morlon et al., 2011) model, conditional on a specific number of extant species sampled and (optionally) conditional on the crown age or stem age.

The probability distribution of such trees only depends on the congruence class of birth-death models (e.g., as specified by the pulled speciation rate) but not on the precise model within a congruence class (Louca and Pennell, 2019). Hence, in addition to allowing specification of speciation and extinction rates, this function can alternatively simulate trees simply based on some pulled speciation rate (PSR), or based on some pulled diversification rate (PDR) and the product $\rho\lambda_o$ (present-day sampling fraction times present-day speciation rate).

This function can be used to generate bootstrap samples after fitting an HBD model or HBD congruence class to a real timetree.

Usage

```
generate_tree_hbd_reverse( Ntips,
                          stem_age       = NULL,
                          crown_age      = NULL,
                          age_grid       = NULL,
                          lambda         = NULL,
                          mu             = NULL,
                          rho            = NULL,
                          PSR            = NULL,
                          PDR            = NULL,
                          rholambda0     = NULL,
                          force_max_age  = Inf,
                          splines_degree = 1,
                          relative_dt    = 1e-3,
                          Ntrees         = 1,
                          tip_basename   = "",
                          node_basename  = NULL,
                          edge_basename  = NULL)
```

Arguments

Ntips	Number of tips in the tree, i.e. number of extant species sampled at present day.
stem_age	Numeric, optional stem age on which to condition the tree. If NULL or <=0, the tree is not conditioned on the stem age.
crown_age	Numeric, optional crown age (aka. root age or MRCA age) on which to condition the tree. If NULL or <=0, the tree is not conditioned on the crown age. If both stem_age and crown_age are specified, only the crown age is used; in that case for consistency crown_age must not be greater than stem_age.
age_grid	Numeric vector, listing discrete ages (time before present) on which the PSR is specified. Listed ages must be strictly increasing, and should cover at least the present day (age 0) as well as a sufficient duration into the past. If conditioning on the stem or crown age, that age must also be covered by age_grid. When not conditioning on crown nor stem age, and the generated tree ends up extending beyond the last time point in age_grid, the PSR will be extrapolated as a constant (with value equal to the last value in PSR) as necessary. age_grid also be NULL or a vector of size 1, in which case the PSR is assumed to be time-independent.
lambda	Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing speciation rates (λ, in units 1/time) at the ages listed in age_grid. Speciation rates must be non-negative, and are assumed to vary as a spline between grid points (see argument splines_degree). Can also be NULL, in which case either PSR, or PDR and rholambda0, must be provided.
mu	Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing extinction rates (μ, in units 1/time) at the ages listed in age_grid. Extinction rates must be non-negative, and are assumed to vary as a spline between grid points (see argument splines_degree). Can also be NULL, in which case either PSR, or PDR and rholambda0, must be provided.

rho	Numeric, sampling fraction at present day (fraction of extant species included in the tree). Can also be NULL, in which case either PSR, or PDR and rholambda0, must be provided.
PSR	Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing pulled speciation rates (λ_p, in units 1/time) at the ages listed in age_grid. The PSR must be non-negative (and strictly positive almost everywhere), and is assumed to vary as a spline between grid points (see argument splines_degree). Can also be NULL, in which case either lambda and mu and rho, or PDR and rholambda0, must be provided.
PDR	Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing pulled diversification rates (r_p, in units 1/time) at the ages listed in age_grid. The PDR is assumed to vary polynomially between grid points (see argument splines_degree). Can also be NULL, in which case either lambda and mu and rho, or PSR, must be provided.
rholambda0	Strictly positive numeric, specifying the product $\rho\lambda_o$ (present-day species sampling fraction times present-day speciation rate). Can also be NULL, in which case PSR must be provided.
force_max_age	Numeric, specifying an optional maximum allowed age for the tree's root. If the tree ends up expanding past that age, all remaining lineages are forced to coalesce at that age. This is not statistically consistent with the provided HBD model (in fact it corresponds to a modified HBD model with a spike in the PSR at that time). This argument merely provides a way to prevent excessively large trees if the PSR is close to zero at older ages and when not conditioning on the stem nor crown age, while still keeping the original statistical properties at younger ages. To disable this feature set force_max_age to Inf.
splines_degree	Integer, either 0,1,2 or 3, specifying the polynomial degree of the provided rates PSR, PDR, lambda, mu and rho between grid points in age_grid. For example, if splines_degree==1, then the provided rates are interpreted as piecewise-linear curves; if splines_degree==2 the rates are interpreted as quadratic splines; if splines_degree==3 the rates are interpreted as cubic splines. The splines_degree influences the analytical properties of the curve, e.g. splines_degree==1 guarantees a continuous curve, splines_degree==2 guarantees a continuous curve and continuous derivative, and so on. If your age_grid is fine enough, then splines_degree=1 is usually sufficient.
relative_dt	Strictly positive numeric (unitless), specifying the maximum relative time step allowed for integration over time. Smaller values increase integration accuracy but increase computation time. Typical values are 0.0001-0.001. The default is usually sufficient.
Ntrees	Integer, number of trees to generate. The computation time per tree is lower if you generate multiple trees at once.
tip_basename	Character. Prefix to be used for tip labels (e.g. "tip."). If empty (""), then tip labels will be integers "1", "2" and so on.
node_basename	Character. Prefix to be used for node labels (e.g. "node."). If NULL, no node labels will be included in the tree.
edge_basename	Character. Prefix to be used for edge labels (e.g. "edge."). Edge labels (if included) are stored in the character vector edge.label. If NULL, no edge labels will be included in the tree.

Details

This function requires that the BD model, or the BD congruence class (Louca and Pennell, 2019), is specified using one of the following sets of arguments:

- Using the speciation rate λ, the extinctin rate μ, and the present-day sampling fraction ρ.

- Using the pulled diversification rate (PDR) and the product $\rho\lambda(0)$. The PDR is defined as $r_p = \lambda - \mu + \frac{1}{\lambda}\frac{d\lambda}{d\tau}$, where τ is age (time before present), $\lambda(\tau)$ is the speciation rate at age τ and $\mu(\tau)$ is the extinction rate.

- Using the pulled speciation rate (PSR). The PSR (λ_p) is defined as $\lambda_p(\tau) = \lambda(\tau) \cdot \Phi(\tau)$, where and $\Phi(\tau)$ is the probability that a lineage extant at age τ will survive until the present and be represented in the tree.

Concurrently using/combining more than one the above parameterization methods is not supported.

Either the PSR, or the PDR and `rholambda0`, provide sufficient information to fully describe the probability distribution of the tree (Louca and Pennell, 2019). For example, the probability distribution of generated trees only depends on the PSR, and not on the specific speciation rate λ or extinction rate μ (various combinations of λ and μ can yield the same PSR; Louca and Pennell, 2019). To calculate the PSR and PDR for any arbitrary λ, μ and ρ you can use the function `simulate_deterministic_hbd`.

When not conditioning on the crown age, the age of the root of the generated tree will be stochastic (i.e., non-fixed). This function then assumes a uniform prior distribution (in a sufficiently large time interval) for the origin of the forward HBD process that would have generated the tree, based on a generalization of the EBDP algorithm provided by (Stadler, 2011). When conditioning on stem or crown age, this function is based on the algorithm proposed by Hoehna (2013, Eq. 8).

Note that HBD trees can also be generated using the function `generate_random_tree`. That function, however, generates trees in forward time, and hence when conditioning on the final number of tips the total duration of the simulation is unpredictable; consequently, speciation and extinction rates cannot be specified as functions of "age" (time before present). The function presented here provides a means to generate trees with a fixed number of tips, while specifying λ, μ, λ_p or r_p as functions of age (time before present).

Value

A named list with the following elements:

success	Logical, indicating whether the simulation was successful. If `FALSE`, then the returned list includes an additional 'error' element (character) providing a description of the error; all other return variables may be undefined.
trees	A list of length `Ntrees`, listing the generated trees. Each tree will be an ultrametric timetree of class "phylo".

Author(s)

Stilianos Louca

References

H. Morlon, T. L. Parsons, J. B. Plotkin (2011). Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences. 108:16327-16332.

T. Stadler (2011). Simulating trees with a fixed number of extant species. Systematic Biology. 60:676-684.

S. Hoehna (2013). Fast simulation of reconstructed phylogenies under global time-dependent birth-death processes. Bioinformatics. 29:1367-1374.

S. Louca and M. W. Pennell (in review as of 2019). Phylogenies of extant species are consistent with an infinite array of diversification histories.

See Also

loglikelihood_hbd, simulate_deterministic_hbd, generate_random_tree

Examples

```
# EXAMPLE 1: Generate trees based on some speciation and extinction rate
# In this example we assume an exponentially decreasing speciation rate
#   and a temporary mass extinction event

# define parameters
age_grid = seq(0,100,length.out=1000)
lambda   = 0.1 + exp(-0.5*age_grid)
mu       = 0.05 + exp(-(age_grid-5)^2)
rho      = 0.5 # species sampling fraction at present-day

# generate a tree with 100 tips and no specific crown or stem age
sim = generate_tree_hbd_reverse(Ntips       = 100,
                                age_grid    = age_grid,
                                lambda      = lambda,
                                mu          = mu,
                                rho         = rho)
if(!sim$success){
    cat(sprintf("Tree generation failed: %s\n",sim$error))
}else{
    cat(sprintf("Tree generation succeeded\n"))
    tree = sim$trees[[1]]
}


#########################
# EXAMPLE 2: Generate trees based on the pulled speciation rate
# Here we condition the tree on some fixed crown (MRCA) age

# specify the PSR on a sufficiently fine and wide age grid
age_grid  = seq(0,1000,length.out=10000)
PSR       = 0.1+exp(-0.1*age_grid) # exponentially decreasing PSR

# generate a tree with 100 tips and MRCA age 10
sim = generate_tree_hbd_reverse(Ntips       = 100,
```

```
                                      age_grid    = age_grid,
                                      PSR         = PSR,
                                      crown_age   = 10)
if(!sim$success){
    cat(sprintf("Tree generation failed: %s\n",sim$error))
}else{
    cat(sprintf("Tree generation succeeded\n"))
    tree = sim$trees[[1]]
}
```

---

generate_tree_with_evolving_rates

Generate a random tree with evolving speciation/extinction rates.

Description

Generate a random phylogenetic tree via simulation of a Poissonian speciation/extinction (birth/death) process. New species are added (born) by splitting of a randomly chosen extant tip. Per-capita birth and death rates (aka. speciation and extinction rates) evolve under some stochastic process (e.g. Brownian motion) along each edge. Thus, the probability rate of a tip splitting or going extinct depends on the tip, with closely related tips having more similar per-capita birth and death rates.

Usage

```
generate_tree_with_evolving_rates(parameters        = list(),
                                  rate_model         = 'BM',
                                  max_tips           = NULL,
                                  max_time           = NULL,
                                  max_time_eq        = NULL,
                                  coalescent         = TRUE,
                                  as_generations     = FALSE,
                                  tip_basename       = "",
                                  node_basename      = NULL,
                                  include_event_times = FALSE,
                                  include_rates      = FALSE)
```

Arguments

parameters   A named list specifying the model parameters for the evolving birth/death rates. The precise entries expected depend on the chosen rate_model (see details below).

rate_model   Character, specifying the model for the evolving per-capita birth/death rates. Must be one of the following: 'BM' (Brownian motion constrained to a finite interval via reflection), 'Mk' (discrete-state continuous-time Markov chain with fixed transition rates).

max_tips	Maximum number of tips of the tree to be generated. If `coalescent=TRUE`, this refers to the number of extant tips. Otherwise, it refers to the number of extinct + extant tips. If `NULL` or <=0, the number of tips is unlimited (so be careful).
max_time	Maximum duration of the simulation. If `NULL` or <=0, this constraint is ignored.
max_time_eq	Maximum duration of the simulation, counting from the first point at which speciation/extinction equilibrium is reached, i.e. when (birth rate - death rate) changed sign for the first time. If `NULL` or <0, this constraint is ignored.
coalescent	Logical, specifying whether only the coalescent tree (i.e. the tree spanning the extant tips) should be returned. If `coalescent==FALSE` and the death rate is non-zero, then the tree may include non-extant tips (i.e. tips whose distance from the root is less than the total time of evolution). In that case, the tree will not be ultrametric.
as_generations	Logical, specifying whether edge lengths should correspond to generations. If `FALSE`, then edge lengths correspond to time.
tip_basename	Character. Prefix to be used for tip labels (e.g. "tip."). If empty (""), then tip labels will be integers "1", "2" and so on.
node_basename	Character. Prefix to be used for node labels (e.g. "node."). If `NULL`, no node labels will be included in the tree.
include_event_times	
	Logical. If `TRUE`, then the times of speciation and extinction events (each in order of occurrence) will also be returned.
include_rates	Logical. If `TRUE`, then the bper-capita birth & death rates of all tips and nodes will also be returned.

Details

If `max_time==NULL`, then the returned tree will always contain `max_tips` tips. In particular, if at any moment during the simulation the tree only includes a single extant tip, the death rate is temporarily set to zero to prevent the complete extinction of the tree. If `max_tips==NULL`, then the simulation is ran as long as specified by `max_time`. If neither `max_time` nor `max_tips` is `NULL`, then the simulation halts as soon as the time exceeds `max_time` or the number of tips (extant tips if `coalescent` is TRUE) exceeds `max_tips`.

If `rate_model=='BM'`, then per-capita birth rates (speciation rates) and per-capita death rates (extinction rates) evolve according to Brownian Motion, constrained to a finite interval via reflection. Note that speciation and extinction rates are only updated at branching points, i.e. during speciation events, while waiting times until speciation/extinction are based on rates at the previous branching point. The argument `parameters` should be a named list including one or more of the following elements:

- `birth_rate_diffusivity`: Non-negative number. Diffusivity constant for the Brownian motion model of the evolving per-capita birth rate. In units 1/time^3. See [`simulate_bm_model`](simulate_bm_model) for an explanation of the diffusivity parameter.

- `min_birth_rate_pc`: Non-negative number. The minimum allowed per-capita birth rate of a clade. In units 1/time. By default this is 0.

- `max_birth_rate_pc`: Non-negative number. The maximum allowed per-capita birth rate of a clade. In units 1/time. By default this is 1.

- death_rate_diffusivity: Non-negative number. Diffusivity constant for the Brownian motion model of the evolving per-capita death rate. In units 1/time^3. See [simulate_bm_model](#) for an explanation of the diffusivity parameter.

- min_death_rate_pc: Non-negative number. The minimum allowed per-capita death rate of a clade. In units 1/time. By default this is 0.

- max_death_rate_pc: Non-negative number. The maximum allowed per-capita death rate of a clade. In units 1/time. By default this is 1.

- root_birth_rate_pc: Non-negative number, between min_birth_rate_pc and max_birth_rate_pc, specifying the initial per-capita birth rate of the root. If left unspecified, this will be chosen randomly and uniformly within the allowed interval.

- root_death_rate_pc: Non-negative number, between min_death_rate_pc and max_death_rate_pc, specifying the initial per-capita death rate of the root. If left unspecified, this will be chosen randomly and uniformly within the allowed interval.

- rarefaction: Numeric between 0 and 1. Rarefaction to be applied at the end of the simulation (fraction of random tips kept in the tree). Note that if coalescent==FALSE, rarefaction may remove both extant as well as extinct clades. Set rarefaction=1 to not perform any rarefaction.

If rate_model=='Mk', then speciation/extinction rates are determined by a tip's current "state", which evolves according to a continuous-time discrete-state Markov chain (Mk model) with constant transition rates. The argument parameters should be a named list including one or more of the following elements:

- Nstates: Number of possible discrete states a tip can have. For example, if Nstates then this corresponds to the common Binary State Speciation and Extinction (BiSSE) model (Maddison et al., 2007). By default this is 1.

- state_birth_rates: Numeric vector of size Nstates, listing the per-capita birth rate (speciation rate) at each state. Can also be a single number (all states have the same birth rate).

- state_death_rates: Numeric vector of size Nstates, listing the per-capita death rate (extinction rate) at each state. Can also be a single number (all states have the same death rate).

- transition_matrix: 2D numeric matrix of size Nstates x Nstates. Transition rate matrix for the Markov chain model of birth/death rate evolution.

- start_state: Integer within 1,..,Nstates, specifying the initial state of the first created lineage. If left unspecified, this is chosen randomly and uniformly among all possible states.

- rarefaction: Same as when rate_model=='BM'.

Note: The option rate_model=='Mk' is deprecated and included for backward compatibility purposes only. To generate a tree with Markov transitions between states (known as Multiple State Speciation and Extinction model), use the command simulate_dsse instead.

Value

A named list with the following elements:

success   Logical, indicating whether the simulation was successful. If FALSE, an additional element error (of type character) is included containing an explanation of the error; in that case the value of any of the other elements is undetermined.

tree   A rooted bifurcating tree of class "phylo", generated according to the specified
                        birth/death model.

                        If `coalescent==TRUE` or if all death rates are zero, and only if `as_generations==FALSE`,
                        then the tree will be ultrametric. If `as_generations==TRUE` and `coalescent==FALSE`,
                        all edges will have unit length.

root_time   Numeric, giving the time at which the tree's root was first split during the simu-
                        lation. Note that if `coalescent==TRUE`, this may be later than the first speciation
                        event during the simulation.

final_time   Numeric, giving the final time at the end of the simulation. If `coalescent==TRUE`,
                        then this may be greater than the total time span of the tree (since the root of the
                        coalescent tree need not correspond to the first speciation event).

equilibrium_time
                        Numeric, giving the first time where the sign of (death rate - birth rate) changed
                        from the beginning of the simulation, i.e. when speciation/extinction equilib-
                        rium was reached. May be infinite if the simulation stoped before reaching this
                        point.

Nbirths   Total number of birth events (speciations) that occurred during tree growth. This
                        may be lower than the total number of tips in the tree if death rates were non-zero
                        and `coalescent==TRUE`.

Ndeaths   Total number of deaths (extinctions) that occurred during tree growth.

birth_times   Numeric vector, listing the times of speciation events during tree growth, in or-
                        der of occurrence. Note that if `coalescent==TRUE`, then `speciation_times`
                        may be greater than the phylogenetic distance to the coalescent root. Only re-
                        turned if `include_event_times==TRUE`.

death_times   Numeric vector, listing the times of extinction events during tree growth, in order
                        of occurrence. Note that if `coalescent==TRUE`, then `speciation_times` may
                        be greater than the phylogenetic distance to the coalescent root. Only returned
                        if `include_event_times==TRUE`.

birth_rates_pc   Numeric vector, listing the per-capita birth rate of each tip and node in the tree.
                        The length of an edge in the tree was thus drawn from an exponential distribution
                        with rate equal to the per-capita birth rate of the child tip or node.

death_rates_pc   Numeric vector, listing the per-capita death rate of each tip and node in the tree.

states   Integer vector of size Ntips+Nnodes, listing the discrete state of each tip and
                        node in the tree. Only included if `rate_model=="Mk"`.

start_state   Integer, specifying the initial state of the first created lineage (either provided
                        during the function call, or generated randomly). Only included if `rate_model=="Mk"`.

root_birth_rate_pc
                        Numeric, specifying the initial per-capita birth rate of the root (either provided
                        during the function call, or generated randomly). Only included if `rate_model=="BM"`.

root_death_rate_pc
                        Numeric, specifying the initial per-capita death rate of the root (either provided
                        during the function call, or generated randomly). Only included if `rate_model=="BM"`.

Author(s)

Stilianos Louca

References

D. J. Aldous (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statistical Science. 16:23-34.

W. P. Maddison, P. E. Midford, S. P. Otto (2007). Estimating a binary character's effect on speciation and extinction. Systematic Biology. 56:701-710.

See Also

simulate_dsse

Examples

```
# Example 1
# Generate tree, with rates evolving under Brownian motion
parameters = list(birth_rate_diffusivity  = 1,
                  min_birth_rate_pc        = 1,
                  max_birth_rate_pc        = 2,
                  death_rate_diffusivity   = 0.5,
                  min_death_rate_pc        = 0,
                  max_death_rate_pc        = 1)
simulation = generate_tree_with_evolving_rates(parameters,
                                               rate_model='BM',
                                               max_tips=1000,
                                               include_rates=TRUE)
tree  = simulation$tree
Ntips = length(tree$tip.label)

# plot per-capita birth & death rates of tips
plot( x=simulation$birth_rates_pc[1:Ntips],
      y=simulation$death_rates_pc[1:Ntips],
      type='p',
      xlab="pc birth rate",
      ylab="pc death rate",
      main="Per-capita birth & death rates across tips (BM model)",
      las=1)


#######################
# Example 2
# Generate tree, with rates evolving under a binary-state model
Q = get_random_mk_transition_matrix(Nstates=2, rate_model="ER", max_rate=0.1)
parameters = list(Nstates = 2,
                  state_birth_rates = c(1,1.5),
                  state_death_rates = 0.5,
                  transition_matrix = Q)
simulation = generate_tree_with_evolving_rates(parameters,
                                               rate_model='Mk',
                                               max_tips=1000,
                                               include_rates=TRUE)
tree  = simulation$tree
Ntips = length(tree$tip.label)
```

```
# plot distribution of per-capita birth rates of tips
rates = simulation$birth_rates_pc[1:Ntips]
barplot(table(rates)/length(rates),
        xlab="rate",
        main="Distribution of pc birth rates across tips (Mk model)")
```

geographic_acf   *Phylogenetic autocorrelation function of geographic locations.*

Description

Given a rooted phylogenetic tree and geographic coordinates (latitudes & longitudes) of each tip,
calculate the phylogenetic autocorrelation function (ACF) of the geographic locations. The ACF is
a function of phylogenetic distance x, i.e., ACF(x) is the autocorrelation between two tip locations
conditioned on the tips having phylogenetic ("patristic") distance x.

Usage

```
geographic_acf(tree, tip_latitudes, tip_longitudes, Npairs=10000, Nbins=10)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_latitudes	A numeric vector of size Ntips, specifying the latitudes (decimal degrees) of the tips. Note that tip_latitudes[i] (where i is an integer index) must correspond to the i-th tip in the tree, i.e. as listed in tree$tip.label.
tip_longitudes	A numeric vector of size Ntips, specifying the longitudes (decimal degrees) of the tips. Note that tip_longitudes[i] (where i is an integer index) must correspond to the i-th tip in the tree, i.e. as listed in tree$tip.label.
Npairs	Total number of random tip pairs to draw. A greater number of tip pairs will improve the accuracy of the estimated ACF within each distance bin. Tip pairs are drawn randomly with replacement. If Npairs<=0, then every tip pair is included exactly once.
Nbins	Number of distance bins to consider within the range of phylogenetic distances encountered between tip pairs in the tree. A greater number of bins will increase the resolution of the ACF as a function of phylogenetic distance, but will decrease the number of tip pairs falling within each bin (which reduces the accuracy of the estimated ACF).

Details

The autocorrelation between random geographic locations is defined as the expectation of $< X, Y >$,
where $<>$ is the scalar product and X and Y are the unit vectors pointing towards the two random
locations on the sphere. For comparison, for a spherical Brownian Motion model with constant

diffusivity D and radius r the autocorrelation function is given by $ACF(t) = e^{-2Dt/r^2}$ (see e.g. simulate_sbm).

The phylogenetic autocorrelation function (ACF) of the geographic distribution of species can give insight into the dispersal processes shaping species distributions over global scales. An ACF that decays slowly with increasing phylogenetic distance indicates a strong phylogenetic conservatism of the location and thus slow dispersal, whereas a rapidly decaying ACF indicates weak phylogenetic conservatism and thus fast dispersal. Similarly, if the mean distance between two random tips increases with phylogenetic distance, this indicates a phylogenetic autocorrelation of species locations. Here, phylogenetic distance between tips refers to their patristic distance, i.e. the minimum cumulative edge length required to connect the two tips.

Since the phylogenetic distances between all possible tip pairs do not cover a continuoum (as there is only a finite number of tips), this function randomly draws tip pairs from the tree, maps them onto a finite set of equally-sized distance bins and then estimates the ACF for the centroid of each distance bin based on tip pairs in that bin. In practice, as a next step one would usually plot the estimated ACF (returned vector autocorrelations) over the centroids of the distance bins (returned vector distances).

The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). If tree$edge.length is missing, then every edge is assumed to have length 1. The input tree must be rooted at some node for technical reasons (see function root_at_node), but the choice of the root node does not influence the result.

Value

A list with the following elements:

success
: Logical, indicating whether the calculation was successful. If FALSE, an additional element error (character) is returned that provides a brief description of the error that occurred; in that case all other return values may be undefined.

distances
: Numeric vector of size Nbins, storing the centroid phylogenetic distance of each distance bin in increasing order. The first and last distance bin approximately span the full range of phylogenetic distances encountered between any two random tips in the tree.

autocorrelations
: Numeric vector of size Nbins, storing the estimated Pearson autocorrelation of the trait for each distance bin.

mean_geodistances
: Numeric vector of size Nbins, storing the mean geographic distance between tip pairs in each distance bin.

Npairs_per_distance
: Integer vector of size Nbins, storing the number of random tip pairs associated with each distance bin.

Author(s)

Stilianos Louca

See Also

get_trait_depth, get_trait_acf

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=1000)$tree

# simulate spherical Brownian Motion on the tree
simul = simulate_sbm(tree, radius=1, diffusivity=0.1)
tip_latitudes  = simul$tip_latitudes
tip_longitudes = simul$tip_longitudes

# calculate geographical autocorrelation function
ACF = geographic_acf(tree, tip_latitudes, tip_longitudes, Nbins=10)

# plot ACF (autocorrelation vs phylogenetic distance)
plot(ACF$distances, ACF$autocorrelations, type="l", xlab="distance", ylab="ACF")
```

---

get_all_distances_to_root

Get distances of all tips and nodes to the root.

Description

Given a rooted phylogenetic tree, calculate the phylogenetic distance (cumulative branch length) of the root to each tip and node.

Usage

```
get_all_distances_to_root(tree, as_edge_count=FALSE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

as_edge_count   Logical, specifying whether distances should be counted in number of edges, rather than cumulative edge length. This is the same as if all edges had length 1.

Details

If tree$edge.length is missing, then every edge in the tree is assumed to be of length 1. The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child). The asymptotic average time complexity of this function is O(Nedges), where Nedges is the number of edges in the tree.

Value

A numeric vector of size Ntips+Nnodes, with the i-th element being the distance (cumulative branch length) of the i-th tip or node to the root. Tips are indexed 1,..,Ntips and nodes are indexed (Ntips+1),..,(Ntips+Nnodes).

Author(s)

Stilianos Louca

See Also

get_pairwise_distances

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1,
                                 death_rate_intercept=0.5),
                            max_tips=Ntips)$tree

# calculate distances to root
all_distances = get_all_distances_to_root(tree)

# extract distances of nodes to root
node_distances = all_distances[(Ntips+1):(Ntips+tree$Nnode)]

# plot histogram of distances (across all nodes)
hist(node_distances, xlab="distance to root", ylab="# nodes", prob=FALSE);
```

---

get_all_node_depths   *Get the phylogenetic depth of each node in a tree.*

Description

Given a rooted phylogenetic tree, calculate the phylogenetic depth of each node (mean distance to its descending tips).

Usage

```
get_all_node_depths(tree, as_edge_count=FALSE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

as_edge_count   Logical, specifying whether distances should be counted in number of edges, rather than cumulative edge length. This is the same as if all edges had length 1.

Details

If tree$edge.length is missing, then every edge in the tree is assumed to be of length 1. The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child). The asymptotic average time complexity of this function is O(Nedges), where Nedges is the number of edges in the tree.

Value

A numeric vector of size Nnodes, with the i-th element being the mean distance of the i-th node to all of its tips.

Author(s)

Stilianos Louca

See Also

get_all_distances_to_root

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1,
                                 death_rate_intercept=0.5),
                            max_tips=Ntips)$tree

# calculate node phylogenetic depths
node_depths = get_all_node_depths(tree)

# plot histogram of node depths
hist(node_depths, xlab="phylogenetic depth", ylab="# nodes", prob=FALSE);
```

---

get_all_pairwise_distances

Get distances between all pairs of tips and/or nodes.

Description

Calculate phylogenetic ("patristic") distances between all pairs of tips or nodes in the tree, or among a subset of tips/nodes requested.

Usage

```
get_all_pairwise_distances( tree,
                            only_clades   = NULL,
                            as_edge_counts = FALSE,
                            check_input   = TRUE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with
                no incoming edge.

only_clades   Optional integer vector or character vector, listing tips and/or nodes to which to
                   restrict pairwise distance calculations. If an integer vector, it must list indices
                   of tips (from 1 to Ntips) and/or nodes (from Ntips+1 to Ntips+Nnodes). If a
                   character vector, it must list tip and/or node names.

                   For example, if only_clades=c('apple','lemon','pear'), then only the
                   distance between 'apple' and 'lemon', between 'apple' and 'pear', and between
                   'lemon' and 'pear' are calculated. If only_clades==NULL, then this is equiva-
                   lent to only_clades=c(1:(Ntips+Nnodes)).

check_input   Logical, whether to perform basic validations of the input data. If you know for
                   certain that your input is valid, you can set this to FALSE to reduce computation
                   time.

as_edge_counts   Logical, specifying whether distances should be calculated in terms of edge
                   counts, rather than cumulative edge lengths. This is the same as if all edges
                   had length 1.

Details

The "patristic distance" between two tips and/or nodes is the shortest cumulative branch length that
must be traversed along the tree in order to reach one tip/node from the other.This function returns
a square distance matrix, containing the patristic distance between all possible pairs of tips/nodes
in the tree (or among the ones provided in only_clades).

If tree$edge.length is missing, then each edge is assumed to be of length 1; this is the same as
setting as_edge_counts=TRUE. The tree may include multi-furcations as well as mono-furcations
(i.e. nodes with only one child). The input tree must be rooted at some node for technical reasons
(see function [root_at_node](#)), but the choice of the root node does not influence the result. If
only_clades is a character vector, then tree$tip.label must exist. If node names are included
in only_clades, then tree$node.label must also exist.

The asymptotic average time complexity of this function for a balanced binary tree is O(NC*NC*Nanc
+ Ntips), where NC is the number of tips/nodes considered (e.g., the length of only_clades) and
Nanc is the average number of ancestors per tip.

Value

A 2D numeric matrix of size NC x NC, where NC is the number of tips/nodes considered, and
with the entry in row r and column c listing the distance between the r-th and the c-th clade con-
sidered (e.g., between clades only_clades[r] and only_clades[c]). Note that if only_clades
was specified, then the rows and columns in the returned distance matrix correspond to the entries
in only_clades (i.e., in the same order). If only_clades was NULL, then the rows and columns in
the returned distance matrix correspond to tips (1,..,Ntips) and nodes (Ntips+1,..,Ntips+Nnodes)

Author(s)

Stilianos Louca

See Also

[get_all_distances_to_root](#), [get_pairwise_distances](#)

Examples

```
# generate a random tree
Ntips = 100
tree  = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# calculate distances between all internal nodes
only_clades = c((Ntips+1):(Ntips+tree$Nnode))
distances = get_all_pairwise_distances(tree, only_clades)

# reroot at some other node
tree = root_at_node(tree, new_root_node=20, update_indices=FALSE)
new_distances = get_all_pairwise_distances(tree, only_clades)

# verify that distances remained unchanged
plot(distances,new_distances,type='p')
```

---

get_clade_list   *Get a representation of a tree as a table listing tips/nodes.*

Description

Given a tree in standard "phylo" format, calculate an alternative representation of the tree structure as a list of tips/nodes with basic information on parents, children and incoming edge lengths. This function is analogous to the function read.tree.nodes in the R package phybase.

Usage

```
get_clade_list(tree, postorder=FALSE, missing_value=NA)
```

Arguments

tree	A tree of class "phylo". If postorder==TRUE, then the tree must be rooted.
postorder	Logical, specifying whether nodes should be ordered and indexed in postorder traversal, i.e. with the root node listed last. Note that regardless of the value of postorder, tips will always be listed first and indexed in the order in which they are listed in the input tree.
missing_value	Value to be used to denote missing information in the returned arrays, for example to denote the (non-existing) parent of the root node.

Details

This function is analogous to the function read.tree.nodes in the R package phybase v1.4, but becomes multiple orders of magnitude faster than the latter for large trees (i.e. with 1000-1000,000 tips). Specifically, calling get_clade_list with postorder=TRUE and missing_value=-9 on a bifurcating tree yields a similar behavior as calling read.tree.nodes with the argument "name" set to the tree's tip labels.

The input tree can include monofurcations, bifurcations and multifurcations. The asymptotic average time complexity of this function is O(Nedges), where Nedges is the number of edges in the tree.

Value

A named list with the following elements:

success	Logical, indicating whether model fitting succeeded. If FALSE, the returned list will include an additional "error" element (character) providing a description of the error; in that case all other return variables may be undefined.
Nsplits	The maximum number of children of any node in the tree. For strictly bifurcating trees this will be 2.
clades	2D integer matrix of size Nclades x (Nsplits+1), with every row representing a specific tip/node in the tree. If postorder==FALSE, then rows are in the same order as tips/nodes in the original tree, otherwise nodes (but not tips) will be re-ordered and re-indexed in postorder fashion, with the root being the last row. The first column lists the parent node index, the remaining columns list the child tip/node indices. For the root, the parent index will be set to missing_value; for the tips, the child indices will be set to missing_value. For nodes with fewer than Nsplits children, superfluous column entries will also be missing_value.
lengths	Numeric vector of size Nclades, listing the lengths of the incoming edges at each tip/node in clades. For the root, the value will be missing_value. If the tree's edge_length was NULL, then lengths will be NULL as well.
old2new_clade	Integer vector of size Nclades, mapping old tip/node indices to tip/node indices in the returned clades and lengths arrays. If postorder==FALSE, this will simply be c(1:Nclades).

Author(s)

Stilianos Louca

Examples

```
# generate a random bifurcating tree
tree = generate_random_tree(list(birth_rate_intercept=1),
                            max_tips=100)$tree

# get tree structure as clade list
# then convert into a similar format as would be
# returned by phybase::read.tree.nodes v1.4
results = get_clade_list(tree,postorder=TRUE,missing_value=-9)
nodematrix = cbind( results$clades,
                    results$lengths,
                    matrix(-9,nrow=nrow(results$clades),ncol=3))
phybaseformat = list(   nodes = nodematrix,
                        names = tree$tip.label,
                        root  = TRUE)
```

---

get_independent_contrasts
                    Phylogenetic independent contrasts for continuous traits.

Description

Calculate phylogenetic independent contrasts (PICs) for one or more continuous traits on a phylo-
genetic tree, as described by Felsenstein (1985). The trait states are assumed to be known for all
tips of the tree. PICs are commonly used to calculate correlations between multiple traits, while
accounting for shared evolutionary history at the tips. This function also returns an estimate for the
state of the root or, equivalently, the phylogenetically weighted mean of the tip states (Garland et
al., 1999).

Usage

```
get_independent_contrasts(tree,
                          tip_states,
                          scaled = TRUE,
                          only_bifurcations = FALSE,
                          check_input = TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	A numeric vector of size Ntips, or a 2D numeric matrix of size Ntips x Ntraits, specifying the numeric state of each trait at each tip in the tree.
scaled	Logical, specifying whether to divide (standardize) PICs by the square root of their expected variance, as recommended by Felsenstein (1985).
only_bifurcations	
	Logical, specifying whether to only calculate PICs for bifurcating nodes. If FALSE, then multifurcations are temporarily expanded to bifurcations, and an additional PIC is calculated for each created bifurcation. If TRUE, then multifurcations are not expanded and PICs will not be calculated for them.
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

If the tree is bifurcating, then one PIC is returned for each node. If multifurcations are present and
only_bifurcations==FALSE, these are internally expanded to bifurcations and an additional PIC
is returned for each such bifurcation. PICs are never returned for monofurcating nodes. Hence,
in general the number of returned PICs is the number of bifurcations in the tree, potentially after
multifurcations have been expanded to bifurcations (if only_bifurcations==FALSE).

If `tree$edge.length` is missing, each edge in the tree is assumed to have length 1. The tree may include multifurcations (i.e. nodes with more than 2 children) as well as monofurcations (i.e. nodes with only one child). Edges with length 0 will be adjusted internally to some tiny length (chosen to be much smaller than the smallest non-zero length).

Tips must be represented in `tip_states` in the same order as in `tree$tip.label`. The vector `tip_states` need not include item names; if it does, however, they are checked for consistency (if `check_input==TRUE`).

The function has asymptotic time complexity O(Nedges x Ntraits). It is more efficient to calculate PICs of multiple traits with the same function call, than to calculate PICs for each trait separately. For a single trait, this function is equivalent to the function `ape::pic`, with the difference that it can handle multifurcating trees.

Value

A list with the following elements:

PICs	A numeric vector (if `tip_states` is a vector) or a numeric matrix (if `tip_states` is a matrix), listing the phylogenetic independent contrasts for each trait and for each bifurcating node (potentially after multifurcations have been expanded). If a matrix, then `PICs[:,T]` will list the PICs for the T-th trait. Note that the order of elements in this vector (or rows, if `PICs` is a matrix) is not necesssarily the order of nodes in the tree, and that `PICs` may contain fewer or more elements (or rows) than there were nodes in the input tree.
distances	Numeric vector of the same size as `PICs`. The "evolutionary distances" (or time) corresponding to the PICs under a Brownian motion model of trait evolution. These roughly correspond to the cumulative edge lengths between sister nodes from which PICs were calculated; hence their units are the same as those of edge lengths. They do not take into account the actual trait values. See Felsenstein (1985) for details.
nodes	Integer vector of the same size as `PICs`, listing the node indices for which PICs are returned. If `only_bifurcations==FALSE`, then this vector may contain NAs, corresponding to temporary nodes created during expansion of multifurcations.
	If `only_bifurcations==TRUE`, then this vector will only list nodes that were bifurcating in the input tree. In that case, `PICs[1]` will correspond to the node with name `tree$node.label[nodes[1]]`, whereas `PICs[2]` will correspond to the node with name `tree$node.label[nodes[2]]`, and so on.
root_state	Numeric vector of size Ntraits, listing the globally estimated state for the root or, equivalently, the phylogenetically weighted mean of the tip states.
root_standard_error	Numeric vector of size Ntraits, listing the phylogenetically estimated standard errors of the root state under a Brownian motion model. The standard errors have the same units as the traits and depend both on the tree topology as well as the tip states. Calculated according to the procedure described by Garland et al. (1999, page 377).
root_CI95	Numeric vector of size Ntraits, listing the radius (half width) of the 95% confidence interval of the root state. Calculated according to the procedure described by Garland et al. (1999, page 377). Note that in contrast to the CI95 returned by

the ace function in the ape package (v. 0.5-64), root_CI95 has the same units
as the traits and depends both on the tree topology as well as the tip states.

Author(s)

Stilianos Louca

References

J. Felsenstein (1985). Phylogenies and the Comparative Method. The American Naturalist. 125:1-
15.

T. Garland Jr., P. E. Midford, A. R. Ives (1999). An introduction to phylogenetically based statistical
methods, with a new method for confidence intervals on ancestral values. American Zoologist.
39:374-388.

See Also

asr_independent_contrasts

Examples

```
# generate random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# simulate a continuous trait
tip_states = simulate_bm_model(tree, diffusivity=0.1, include_nodes=FALSE)$tip_states;

# calculate PICs
results = get_independent_contrasts(tree, tip_states, scaled=TRUE, only_bifurcations=TRUE)

# assign PICs to the bifurcating nodes in the input tree
PIC_per_node = rep(NA, tree$Nnode)
valids = which(!is.na(results$nodes))
PIC_per_node[results$nodes[valids]] = results$PICs[valids]
```

---

get_mrca_of_set   *Most recent common ancestor of a set of tips/nodes.*

Description

Given a rooted phylogenetic tree and a set of tips and/or nodes ("descendants"), calculate the most
recent common ancestor (MRCA) of those descendants.

Usage

```
get_mrca_of_set(tree, descendants)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
descendants	An integer vector or character vector, specifying the tips/nodes for which to find the MRCA. If an integer vector, it must list indices of tips (from 1 to Ntips) and/or nodes (from Ntips+1 to Ntips+Nnodes), where Ntips and Nnodes is the number of tips and nodes in the tree, respectively. If a character vector, it must list tip and/or node names. In this case tree must include tip.label, as well as node.label if nodes are included in descendants.

Details

The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). Duplicate entries in descendants are ignored.

Value

An integer in 1,..,(Ntips+Nnodes), representing the MRCA using the same index as in tree$edge. If the MRCA is a tip, then this index will be in 1,..,Ntips. If the MRCA is a node, then this index will be in (Ntips+1),..,(Ntips+Nnodes).

Author(s)

Stilianos Louca

See Also

[get_pairwise_mrcas](), [get_tips_for_mrcas]()

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# pick 3 random tips or nodes
descendants = sample.int(n=(Ntips+tree$Nnode), size=3, replace=FALSE)

# calculate MRCA of picked descendants
mrca = get_mrca_of_set(tree, descendants)
```

---

get_pairwise_distances

Get distances between pairs of tips or nodes.

Description

Calculate phylogenetic ("patristic") distances between tips or nodes in some list A and tips or nodes in a second list B of the same size.

Usage

```
get_pairwise_distances(tree, A, B, as_edge_counts=FALSE, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
A	An integer vector or character vector of size Npairs, specifying the first of the two members of each pair for which to calculate the distance. If an integer vector, it must list indices of tips (from 1 to Ntips) and/or nodes (from Ntips+1 to Ntips+Nnodes). If a character vector, it must list tip and/or node names.
B	An integer vector or character vector of size Npairs, specifying the second of the two members of each pair for which to calculate the distance. If an integer vector, it must list indices of tips (from 1 to Ntips) and/or nodes (from Ntips+1 to Ntips+Nnodes). If a character vector, it must list tip and/or node names.
check_input	Logical, whether to perform basic validations of the input data. If you know for certain that your input is valid, you can set this to FALSE to reduce computation time.
as_edge_counts	Logical, specifying whether distances should be calculated in terms of edge counts, rather than cumulative edge lengths. This is the same as if all edges had length 1.

Details

The "patristic distance" between two tips and/or nodes is the shortest cumulative branch length that must be traversed along the tree in order to reach one tip/node from the other. Given a list of tips and/or nodes A, and a 2nd list of tips and/or nodes B of the same size, this function will calculate patristic distance between each pair (A[i], B[i]), where i=1,2,..,Npairs.

If tree$edge.length is missing, then each edge is assumed to be of length 1; this is the same as setting as_edge_counts=TRUE. The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child). The input tree must be rooted at some node for technical reasons (see function root_at_node), but the choice of the root node does not influence the result. If A and/or B is a character vector, then tree$tip.label must exist. If node names are included in A and/or B, then tree$node.label must also exist.

The asymptotic average time complexity of this function for a balanced binary tree is O(Ntips+Npairs*log2(Ntips)).

Value

A numeric vector of size Npairs, with the i-th element being the patristic distance between the tips/nodes A[i] and B[i].

Author(s)

Stilianos Louca

See Also

[get_all_distances_to_root](), [get_all_pairwise_distances]()

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# pick 3 random pairs of tips or nodes
Npairs = 3
A = sample.int(n=(Ntips+tree$Nnode), size=Npairs, replace=FALSE)
B = sample.int(n=(Ntips+tree$Nnode), size=Npairs, replace=FALSE)

# calculate distances
distances = get_pairwise_distances(tree, A, B)

# reroot at some other node
tree = root_at_node(tree, new_root_node=20, update_indices=FALSE)
new_distances = get_pairwise_distances(tree, A, B)

# verify that distances remained unchanged
print(distances)
print(new_distances)
```

---

get_pairwise_mrcas   *Get most recent common ancestors of tip/node pairs.*

Description

Given a rooted phylogenetic tree and one or more pairs of tips and/or nodes, for each pair of tips/nodes find the most recent common ancestor (MRCA). If one clade is descendant of the other clade, the latter will be returned as MRCA.

Usage

```
get_pairwise_mrcas(tree, A, B, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
A	An integer vector or character vector of size Npairs, specifying the first of the two members of each pair of tips/nodes for which to find the MRCA. If an integer vector, it must list indices of tips (from 1 to Ntips) and/or nodes (from Ntips+1 to Ntips+Nnodes). If a character vector, it must list tip and/or node names.
B	An integer vector or character vector of size Npairs, specifying the second of the two members of each pair of tips/nodes for which to find the MRCA. If an integer vector, it must list indices of tips (from 1 to Ntips) and/or nodes (from Ntips+1 to Ntips+Nnodes). If a character vector, it must list tip and/or node names.
check_input	Logical, whether to perform basic validations of the input data. If you know for certain that your input is valid, you can set this to FALSE to reduce computation time.

Details

The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child). If tree$edge.length is missing, then each edge is assumed to be of length 1. Note that in some cases the MRCA of two tips may be a tip, namely when both tips are the same.

If A and/or B is a character vector, then tree$tip.label must exist. If node names are included in A and/or B, then tree$node.label must also exist.

The asymptotic average time complexity of this function is O(Nedges), where Nedges is the number of edges in the tree.

Value

An integer vector of size Npairs with values in 1,..,Ntips (tips) and/or in (Ntips+1),..,(Ntips+Nnodes) (nodes), with the i-th element being the index of the MRCA of tips/nodes A[i] and B[i].

Author(s)

Stilianos Louca

See Also

get_mrca_of_set, get_tips_for_mrcas

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# pick 3 random pairs of tips or nodes
Npairs = 3
```

```
A = sample.int(n=(Ntips+tree$Nnode), size=Npairs, replace=FALSE)
B = sample.int(n=(Ntips+tree$Nnode), size=Npairs, replace=FALSE)

# calculate MRCAs
MRCAs = get_pairwise_mrcas(tree, A, B)
```

---

get_random_diffusivity_matrix
Create a random diffusivity matrix for a Brownian motion model.

Description

Create a random diffusivity matrix for a Brownian motion model of multi-trait evolution. This may be useful for testing purposes. The diffusivity matrix is drawn from the Wishart distribution of symmetric, nonnegative-definite matrixes:

$$D = X^T \cdot X, \quad X[i,j] \sim N(0,V), \quad i = 1,..,n, j = 1,..,p,$$

where n is the degrees of freedom, p is the number of traits and V is a scalar scaling.

Usage

```
get_random_diffusivity_matrix(Ntraits, degrees=NULL, V=1)
```

Arguments

Ntraits	The number of traits modelled. Equal to the number of rows and the number of columns of the returned matrix.
degrees	Degrees of freedom for the Wishart distribution. Must be equal to or greater than Ntraits. Can also be NULL, which is the same as setting it equal to Ntraits.
V	Positive number. A scalar scaling for the Wishart distribution.

Value

A real-valued quadratic symmetric non-negative definite matrix of size Ntraits x Ntraits. Almost surely (in the probabilistic sense), this matrix will be positive definite.

Author(s)

Stilianos Louca

See Also

[get_random_mk_transition_matrix](#), [simulate_bm_model](#)

Examples

```
# generate a 5x5 diffusivity matrix
D = get_random_diffusivity_matrix(Ntraits=5)

# check that it is indeed positive definite
if(all(eigen(D)$values>0)){
  cat("Indeed positive definite\n");
}else{
  cat("Not positive definite\n");
}
```

---

get_random_mk_transition_matrix

Create a random transition matrix for an Mk model.

Description

Create a random transition matrix for a fixed-rates continuous-time Markov model of discrete trait evolution ("Mk model"). This may be useful for testing purposes.

Usage

```
get_random_mk_transition_matrix(Nstates, rate_model, min_rate=0, max_rate=1)
```

Arguments

Nstates	The number of distinct states represented in the transition matrix (number of rows & columns).
rate_model	Rate model that the transition matrix must satisfy. Can be "ER" (all rates equal), "SYM" (transition rate i–>j is equal to transition rate j–>i), "ARD" (all rates can be different) or "SUEDE" (only stepwise transitions i–>i+1 and i–>i-1 allowed, all 'up' transitions are equal, all 'down' transitions are equal).
min_rate	A non-negative number, specifying the minimum rate in off-diagonal entries of the transition matrix.
max_rate	A non-negative number, specifying the maximum rate in off-diagonal entries of the transition matrix. Must not be smaller than min_rate.

Value

A real-valued quadratic matrix of size Nstates x Nstates, representing a transition matrix for an Mk model. Each row will sum to 0. The [r,c]-th entry represents the transition rate r–>c. The number of unique off-diagonal rates will depend on the rate_model chosen.

Author(s)

Stilianos Louca

See Also

exponentiate_matrix, get_stationary_distribution

Examples

```
# generate a 5x5 Markov transition rate matrix
Q = get_random_mk_transition_matrix(Nstates=5, rate_model="ARD")
```

---

get_reds	*Calculate relative evolutionary divergences in a tree.*

Description

Calculate the relative evolutionary divergence (RED) of each node in a rooted phylogenetic tree. The RED of a node is a measure of its relative placement between the root and the node's descending tips (Parks et al. 2018). The root's RED is always 0, the RED of each tip is 1, and the RED of each node is between 0 and 1.

Usage

```
get_reds(tree)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

Details

The RED of a node measures its relative placement between the root and the node's descending tips (Parks et al. 2018). The root's RED is set to 0. Traversing from root to tips (preorder traversal), for each node the RED is set to $P + (a/(a+b)) \cdot (1 - P)$, where P is the RED of the node's parent, a is the edge length connecting the node to its parent, and b is the average distance from the node to its descending tips. The RED of a tip would always be 1.

The RED may be useful for defining taxonomic ranks based on a molecular phylogeny (e.g. see Parks et al. 2018). This function is similar to the PhyloRank v0.0.27 script published by Parks et al. (2018).

The time complexity of this function is O(Nedges). The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). If tree$edge.length is NULL, then all edges in the input tree are assumed to have length 1.

Value

A numeric vector of length Nnodes, listing the RED of each node in the tree. The REDs of tips are not included, since these would always be equal to 1.

Author(s)

Stilianos Louca

References

D. H. Parks, M. Chuvochina et al. (2018). A proposal for a standardized bacterial taxonomy based on genome phylogeny. bioRxiv 256800. DOI:10.1101/256800

Examples

```
# generate a random tree
params = list(birth_rate_intercept=1, death_rate_intercept=0.8)
tree = generate_random_tree(params, max_time=100, coalescent=FALSE)$tree

# calculate and print REDs
REDs = get_reds(tree)
print(REDs)
```

---

get_stationary_distribution

Stationary distribution of Markov transition matrix.

Description

Calculate the stationary probability distribution vector p for a transition matrix Q of a continuous-time Markov chain. That is, calculate $p \in [0,1]^n$ such that sum(p)==0 and $p^T Q = 0$.

Usage

```
get_stationary_distribution(Q)
```

Arguments

Q   A valid transition rate matrix of size Nstates x Nstates, i.e. a quadratic matrix in which every row sums up to zero.

Details

A stationary distribution of a discrete-state continuous-time Markov chain is a probability distribution across states that remains constant over time, i.e. $p^T Q = 0$. Note that in some cases (i.e. if Q is not irreducible), there may be multiple distinct stationary distributions. In that case, which one is returned by this function is unpredictable. Internally, p is estimated by stepwise minimization of the norm of $p^T Q$, starting with the vector p in which every entry equals 1/Nstates.

Value

A numeric vector of size Nstates and with non-negative entries, satisfying the conditions p%*%Q==0 and sum(p)==1.0.

Author(s)

Stilianos Louca

See Also

exponentiate_matrix

Examples

```
# generate a random 5x5 Markov transition matrix
Q = get_random_mk_transition_matrix(Nstates=5, rate_model="ARD")

# calculate stationary probability distribution
p = get_stationary_distribution(Q)
print(p)

# test correctness (p*Q should be 0, apart from rounding errors)
cat(sprintf("max(abs(p*Q)) = %g\n",max(abs(p %*% Q))))
```

---

get_subtrees_at_nodes   *Extract subtrees descending from specific nodes.*

Description

Given a tree and a list of focal nodes, extract the subtrees descending from those focal nodes, with the focal nodes becoming the roots of the extracted subtrees.

Usage

```
get_subtrees_at_nodes(tree, nodes)
```

Arguments

tree	A tree of class "phylo".
nodes	Character vector or integer vector specifying the names or indices, respectively, of the focal nodes at which to extract the subtrees. If an integer vector, entries must be between 1 and tree$Nnode. If a character vector, each entry must be a valid entry in tree$node.label.

Details

The input tree need not be rooted, however "descendance" from a focal node is inferred based on the direction of edges in tree$edge. The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child).

Value

A list with the following elements:

subtrees	List of the same length as nodes, with each element being a new tree of class "phylo", containing the subtrees descending from the focal nodes. Each subtree will be rooted at the corresponding focal node.
new2old_tip	List of the same length as nodes, with the n-th element being an integer vector with values in 1,..,Ntips, mapping tip indices of the n-th subtree to tip indices in the original tree. In particular, tree$tip.label[new2old_tip[[n]]] will be equal to subtrees[[n]]$tip.label.
new2old_node	List of the same length as nodes, with the n-th element being an integer vector with values in 1,..,Nnodes, mapping node indices of the n-th subtree to node indices in the original tree.
	For example, new2old_node[[2]][1] is the index that the 1st node of the 2nd subtree had within the original tree. In particular, tree$node.label[new2old_node[[n]]] will be equal to subtrees[[n]]$node.label (if node labels are available).
new2old_edge	List of the same length as nodes, with the n-th element being an integer vector with values in 1,..,Nedges, mapping edge indices of the n-th subtree to edge indices in the original tree. In particular, tree$edge.length[new2old_edge[[n]]] will be equal to subtrees[[n]]$edge.length (if edge lengths are available).

Author(s)

Stilianos Louca

See Also

get_subtree_at_node,

get_subtree_with_tips

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# extract subtrees descending from random nodes
nodes = sample.int(tree$Nnode,size=10)
subtrees = get_subtrees_at_nodes(tree, nodes)$subtrees

# print summaries of extracted subtrees
for(n in length(nodes)){
 cat(sprintf("Subtree at %d-th node has %d tips\n",nodes[n],length(subtrees[[n]]$tip.label)))
}
```

---

get_subtree_at_node   *Extract a subtree descending from a specific node.*

Description

Given a tree and a focal node, extract the subtree descending from the focal node and place the focal node as the root of the extracted subtree.

Usage

```
get_subtree_at_node(tree, node)
```

Arguments

tree	A tree of class "phylo".
node	Character or integer specifying the name or index, respectively, of the focal node at which to extract the subtree. If an integer, it must be between 1 and tree$Nnode. If a character, it must be a valid entry in tree$node.label.

Details

The input tree need not be rooted, however "descendance" from the focal node is inferred based on the direction of edges in tree$edge. The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child).

Value

A list with the following elements:

subtree	A new tree of class "phylo", containing the subtree descending from the focal node. This tree will be rooted, with the new root being the focal node.
new2old_tip	Integer vector of length Ntips_kept (=number of tips in the extracted subtree) with values in 1,..,Ntips, mapping tip indices of the subtree to tip indices in the original tree. In particular, tree$tip.label[new2old_tip] will be equal to subtree$tip.label.
new2old_node	Integer vector of length Nnodes_kept (=number of nodes in the extracted subtree) with values in 1,..,Nnodes, mapping node indices of the subtree to node indices in the original tree.
	For example, new2old_node[1] is the index that the first node of the subtree had within the original tree. In particular, tree$node.label[new2old_node] will be equal to subtree$node.label (if node labels are available).
new2old_edge	Integer vector of length Nedges_kept (=number of edges in the extracted subtree), with values in 1,..,Nedges, mapping edge indices of the subtree to edge indices in the original tree. In particular, tree$edge.length[new2old_edge] will be equal to subtree$edge.length (if edge lengths are available).

Author(s)

Stilianos Louca

See Also

[get_subtree_with_tips](#)

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# extract subtree descending from a random node
node = sample.int(tree$Nnode,size=1)
subtree = get_subtree_at_node(tree, node)$subtree

# print summary of subtree
cat(sprintf("Subtree at %d-th node has %d tips\n",node,length(subtree$tip.label)))
```

---

get_subtree_with_tips   *Extract a subtree spanning a specific subset of tips.*

Description

Given a rooted tree and a subset of tips, extract the subtree containing only those tips. The root of
the tree is kept.

Usage

```
get_subtree_with_tips(tree,
                      only_tips             = NULL,
                      omit_tips             = NULL,
                      collapse_monofurcations = TRUE,
                      force_keep_root       = FALSE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
only_tips	Either a character vector listing tip names to keep, or an integer vector listing tip indices to keep (between 1 and Ntips). Can also be NULL. Tips listed in only_tips not found in the tree will be silently ignored.
omit_tips	Either a character vector listing tip names to omit, or an integer vector listing tip indices to omit (between 1 and Ntips). Can also be NULL. Tips listed in omit_tips not found in the tree will be silently ignored.

collapse_monofurcations

A logical specifying whether nodes with a single outgoing edge remaining should be collapsed (removed). Incoming and outgoing edge of such nodes will be concatenated into a single edge, connecting the parent (or earlier) and child (or later) of the node. In that case, the returned tree will have edge lengths that reflect the concatenated edges.

force_keep_root

Logical, specifying whether to keep the root even if collapse_monofurcations==TRUE and the root of the subtree is left with a single child. If FALSE, and collapse_monofurcations==TRUE, the root may be removed and one of its descendants may become root.

Details

If both only_tips and omit_tips are NULL, then all tips are kept and the tree remains unchanged. If both only_tips and omit_tips are non-NULL, then only tips listed in only_tips and not listed in omit_tips will be kept. If only_tips and/or omit_tips is a character vector listing tip names, then tree$tip.label must exist.

If the input tree does not include edge.length, each edge in the input tree is assumed to have length 1. The root of the tree (which is always kept) is assumed to be the unique node with no incoming edge. The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child).

The asymptotic time complexity of this function is O(Nnodes+Ntips), where Ntips is the number of tips and Nnodes the number of nodes in the input tree.

When only_tips==NULL, omit_tips!=NULL, collapse_monofurcations==TRUE and force_keep_root==FALSE, this function is analogous to the function drop.tip in the ape package with option trim_internal=TRUE (v. 0.5-64).

Value

A list with the following elements:

subtree   A new tree of class "phylo", containing only the tips specified by tips_to_keep and the nodes & edges connecting those tips to the root. The returned tree will include edge.lengh as a member variable, listing the lengths of the remaining (possibly concatenated) edges.

root_shift   Numeric, indicating the phylogenetic distance between the old and the new root. Will always be non-negative.

new2old_tip   Integer vector of length Ntips_kept (=number of tips in the extracted subtree) with values in 1,..,Ntips, mapping tip indices of the subtree to tip indices in the original tree. In particular, tree$tip.label[new2old_tip] will be equal to subtree$tip.label.

new2old_node   Integer vector of length Nnodes_kept (=number of nodes in the extracted subtree) with values in 1,..,Nnodes, mapping node indices of the subtree to node indices in the original tree.

For example, new2old_node[1] is the index that the first node of the subtree had within the original tree. In particular, tree$node.label[new2old_node] will be equal to subtree$node.label (if node labels are available).

old2new_tip	Integer vector of length Ntips, with values in 1,..,Ntips_kept, mapping tip indices of the original tree to tip indices in the subtree (a value of 0 is used whenever a tip is absent in the subtree). This is essentially the inverse of the mapping new2old_tip.
old2new_node	Integer vector of length Nnodes, with values in 1,..,Nnodes_kept, mapping node indices of the original tree to node indices in the subtree (a value of 0 is used whenever a node is absent in the subtree). This is essentially the inverse of the mapping new2old_node.

Author(s)

Stilianos Louca

See Also

[get_subtree_at_node](#)

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# choose a random subset of tips
tip_subset = sample.int(Ntips, size=as.integer(Ntips/10), replace=FALSE)

# extract subtree spanning the chosen tip subset
subtree = get_subtree_with_tips(tree, only_tips=tip_subset)$subtree

# print summary of subtree
cat(sprintf("Subtree has %d tips and %d nodes\n",length(subtree$tip.label),subtree$Nnode))
```

---

get_tips_for_mrcas   *Find tips with specific most recent common ancestors.*

Description

Given a rooted phylogenetic tree and a list of nodes ("MRCA nodes"), for each MRCA node find a set of descending tips ("MRCA-defining tips") such that their most recent common ancestor (MRCA) is that node. This may be useful for cases where nodes need to be described as MRCAs of tip pairs for input to certain phylogenetics algorithms (e.g., for tree dating).

Usage

```
get_tips_for_mrcas(tree, mrca_nodes, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
mrca_nodes	Either an integer vector or a character vector, listing the nodes for each of which an MRCA-defining set of tips is to be found. If an integer vector, it should list node indices (i.e. from 1 to Nnodes). If a character vector, it should list node names; in that case tree$node.label must exist.
check_input	Logical, whether to perform basic validations of the input data. If you know for certain that your input is valid, you can set this to FALSE to reduce computation time.

Details

At most 2 MRCA-defining tips are assigned to each MRCA node. This function assumes that each of the mrca_nodes has at least two children or has a child that is a tip (otherwise the problem is not well-defined). The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child).

The asymptotic time complexity of this function is O(Ntips+Nnodes) + O(Nmrcas), where Ntips is the number of tips, Nnodes is the number of nodes in the tree and Nmrcas is equal to length(mrca_nodes).

Value

A list of the same size as mrca_nodes, whose n-th element is an integer vector of tip indices (i.e. with values in 1,..,Ntips) whose MRCA is the n-th node listed in mrca_nodes.

Author(s)

Stilianos Louca

See Also

[get_pairwise_mrcas](), [get_mrca_of_set]()

Examples

```
# generate a random tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# pick random nodes
focal_nodes = sample.int(n=tree$Nnode, size=3, replace=FALSE)

# get tips for mrcas
tips_per_focal_node = get_tips_for_mrcas(tree, focal_nodes);

# check correctness (i.e. calculate actual MRCAs of tips)
for(n in 1:length(focal_nodes)){
  mrca = get_mrca_of_set(tree, tips_per_focal_node[[n]])
  cat(sprintf("Focal node = %d, should match mrca of tips = %d\n",focal_nodes[n],mrca-Ntips))
}
```

---

get_trait_acf   *Phylogenetic autocorrelation function of a numeric trait.*

Description

Given a rooted phylogenetic tree and a numeric (typically continuous) trait with known value (state) on each tip, calculate the phylogenetic autocorrelation function (ACF) of the trait. The ACF is a function of phylogenetic distance x, where ACF(x) is the Pearson autocorrelation of the trait between two tips, provided that the tips have phylogenetic ("patristic") distance x. The function get_trait_acf also calculates the mean absolute difference and the mean relative difference of the trait between any two random tips at phylogenetic distance x (see details below).

Usage

```
get_trait_acf(tree, tip_states, Npairs=10000, Nbins=10)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with
                no incoming edge.

tip_states   A numeric vector of size Ntips, specifying the value of the trait at each tip in the
                tree. Note that tip_states[i] (where i is an integer index) must correspond to the
                i-th tip in the tree.

Npairs   Total number of random tip pairs to draw. A greater number of tip pairs will
                improve the accuracy of the estimated ACF within each distance bin. Tip pairs
                are drawn randomly with replacement. If Npairs<=0, then every tip pair is
                included exactly once.

Nbins   Number of distance bins to consider within the range of phylogenetic distances
                encountered between tip pairs in the tree. A greater number of bins will in-
                crease the resolution of the ACF as a function of phylogenetic distance, but will
                decrease the number of tip pairs falling within each bin (which reduces the ac-
                curacy of the estimated ACF).

Details

The phylogenetic autocorrelation function (ACF) of a trait can give insight into the evolutionary processes shaping its distribution across clades. An ACF that decays slowly with increasing phy- logenetic distance indicates a strong phylogenetic conservatism of the trait, whereas a rapidly de- caying ACF indicates weak phylogenetic conservatism. Similarly, if the mean absolute difference in trait value between two random tips increases with phylogenetic distance, this indicates a phylo- genetic autocorrelation of the trait (Zaneveld et al. 2014). Here, phylogenetic distance between tips refers to their patristic distance, i.e. the minimum cumulative edge length required to connect the two tips.

Since the phylogenetic distances between all possible tip pairs do not cover a continuoum (as there is only a finite number of tips), this function randomly draws tip pairs from the tree, maps them onto a finite set of equally-sized distance bins and then estimates the ACF for the centroid of each distance

bin based on tip pairs in that bin. In practice, as a next step one would usually plot the estimated ACF (returned vector `autocorrelations`) over the centroids of the distance bins (returned vector `distances`).

The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). If `tree$edge.length` is missing, then every edge is assumed to have length 1. The input tree must be rooted at some node for technical reasons (see function [root_at_node](), but the choice of the root node does not influence the result.

Value

A list with the following elements:

`distances`   Numeric vector of size Nbins, storing the centroid phylogenetic distance of each distance bin in increasing order. The first and last distance bin approximately span the full range of phylogenetic distances encountered between any two random tips in the tree.

`autocorrelations`
                Numeric vector of size Nbins, storing the estimated Pearson autocorrelation of the trait for each distance bin.

`mean_abs_differences`
                Numeric vector of size Nbins, storing the mean absolute difference of the trait between tip pairs in each distance bin.

`mean_rel_differences`
                Numeric vector of size Nbins, storing the mean relative difference of the trait between tip pairs in each distance bin. The relative difference between two values X and Y is 0 if $X == Y$, and equal to

$$\frac{|X - Y|}{0.5 \cdot (|X| + |Y|)}$$

otherwise.

`Npairs_per_distance`
                Integer vector of size Nbins, storing the number of random tip pairs associated with each distance bin.

Author(s)

Stilianos Louca

References

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

[get_trait_depth](), [geographic_acf]()

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=1000)$tree

# simulate continuous trait evolution on the tree
tip_states = simulate_bm_model(tree, diffusivity=1)$tip_states

# calculate autocorrelation function
ACF = get_trait_acf(tree, tip_states, Npairs=1e7, Nbins=10)

# plot ACF (autocorrelation vs phylogenetic distance)
plot(ACF$distances, ACF$autocorrelations, type="l", xlab="distance", ylab="ACF")
```

---

get_trait_depth   *Calculate depth of phylogenetic conservatism for a binary trait.*

Description

Given a rooted phylogenetic tree and presences/absences of a binary trait for each tip, calculate the mean phylogenetic depth at which the trait is conserved across clades, in terms of the consenTRAIT metric introduced by Martiny et al (2013). This is the mean depth of clades that are positive in the trait (i.e. in which a sufficient fraction of tips exhibits the trait).

Usage

```
get_trait_depth(tree,
                tip_states,
                min_fraction        = 0.9,
                count_singletons    = TRUE,
                singleton_resolution= 0,
                weighted            = FALSE,
                Npermutations       = 0)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	A numeric vector of size Ntips indicating absence (value <=0) or presence (value >0) of a particular trait at each tip of the tree. Note that tip_states[i] (where i is an integer index) must correspond to the i-th tip in the tree.
min_fraction	Minimum fraction of tips in a clade exhibiting the trait, for the clade to be considered "positive" in the trait. In the original paper by Martiny et al (2013), this was 0.9.
count_singletons	
	Logical, specifying whether to include singletons in the statistics (tips positive in the trait, but not part of a larger positive clade). The phylogenetic depth of singletons is taken to be half the length of their incoming edge, as proposed by Martiny et al (2013). If FALSE, singletons are ignored.

singleton_resolution

> Numeric, specifying the phylogenetic resolution at which to resolve singletons. Any clade found to be positive in a trait will be considered a singleton if the distance of the clade's root to all descending tips is below this threshold.

weighted   Whether to weight positive clades by their number of positive tips. If FALSE, each positive clades is weighted equally, as proposed by Martiny et al (2013).

Npermutations   Number of random permutations for estimating the statistical significance of the mean trait depth. If zero (default), the statistical significance is not calculated.

Details

This function calculates the "consenTRAIT" metric (or variants thereof) proposed by Martiny et al. (2013) for measuring the mean phylogenetic depth at which a binary trait (e.g. presence/absence of a particular metabolic function) is conserved across clades. A greater mean depth means that the trait tends to be conserved in deeper-rooting clades. In their original paper, Martiny et al. proposed to consider a trait as conserved in a clade (i.e. marking a clade as "positive" in the trait) if at least 90% of the clade's tips exhibit the trait (i.e. are "positive" in the trait). This fraction can be controlled using the min_fraction parameter. The depth of a clade is taken as the average distance of its tips to the clade's root.

The default parameters of this function reflect the original choices made by Martiny et al. (2013), however in some cases it may be sensible to adjust them. For example, if you suspect a high risk of false positives in the detection of a trait, it may be worth setting count_singletons to FALSE to avoid skewing the distribution of conservation depths towards shallower depths due to false positives.

The statistical significance of the calculated mean depth, i.e. the probability of encountering such a mean dept or higher by chance, can be estimated based on a null model in which each tip is randomly and independently re-assigned a presence or absence of the trait. In the null model, the probability that a tip exhibits the trait is set to the fraction of positive entries in tip_states.

The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child). If tree$edge.length is missing, then every edge is assumed to have length 1.

Value

A list with the following elements:

mean_depth   Mean phylogenetic depth of clades that are positive in the trait.

var_depth   Variance of phylogenetic depths of clades that are positive in the trait.

min_depth   Minimum phylogenetic depth of clades that are positive in the trait.

max_depth   Maximum phylogenetic depth of clades that are positive in the trait.

Npositives   Number of clades that are positive in the trait.

P   Statistical significance (P-values) of mean_depth, under a null model of random trait presences/absences (see details above). This is the probability that under the null model, the mean_depth would be at least as high as observed in the data.

mean_random_depth

> Mean random mean_depth, under a null model of random trait presences/absences (see details above).

positive_clades

> Integer vector, listing indices of tips and nodes (from 1 to Ntips+Nnodes) that were found to be positive in the trait and counted towards the statistic.

positives_per_clade

> Integer vector of size Ntips+Nnodes, listing the number of descending tips per clade (tip or node) that were positive in the trait.

mean_depth_per_clade

> Numeric vector of size Ntips+Nnodes, listing the mean phylogenetic depth of each clade (tip or node), i.e. the average distance to all its descending tips.

Author(s)

Stilianos Louca

References

A. C. Martiny, K. Treseder and G. Pusch (2013). Phylogenetic trait conservatism of functional traits in microorganisms. ISME Journal. 7:830-838.

See Also

get_trait_acf

Examples

```
## Not run:
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=1000)$tree

# simulate binary trait evolution on the tree
Q = get_random_mk_transition_matrix(Nstates=2, rate_model="ARD", max_rate=0.1)
tip_states = simulate_mk_model(tree, Q)$tip_states

# change states from 1/2 to 0/1 (presence/absence)
tip_states = tip_states - 1

# calculate phylogenetic conservatism of trait
results = get_trait_depth(tree, tip_states, count_singletons=FALSE, weighted=TRUE)
cat(sprintf("Mean depth = %g, std = %g\n",results$mean_depth,sqrt(results$var_depth)))

## End(Not run)
```

---

get_trait_stats_over_time

> *Calculate mean & standard deviation of a numeric trait on a dated tree over time.*

Description

Given a rooted and dated phylogenetic tree, and a scalar numeric trait with known value on each node and tip of the tree, calculate the mean and the variance of the trait's states across the tree at discrete time points. For example, if the trait represents "body size", then this function calculates the mean body size of extant clades over time.

Usage

```
get_trait_stats_over_time(tree, states, Ntimes=NULL, times=NULL, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo", where edge lengths represent time intervals (or similar).
states	Numeric vector, specifying the trait's state at each tip and each node of the tree (in the order in which tips & nodes are indexed). May include NA or NaN if values are missing for some tips/nodes.
Ntimes	Integer, number of equidistant time points for which to calculade clade counts. Can also be NULL, in which case times must be provided.
times	Integer vector, listing time points (in ascending order) for which to calculate clade counts. Can also be NULL, in which case Ntimes must be provided.
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

If tree$edge.length is missing, then every edge in the tree is assumed to be of length 1. The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child). The tree need not be ultrametric (e.g. may include extinct tips), although in general this function only makes sense if edge lengths correspond to time (or similar).

Either Ntimes or times must be non-NULL, but not both. states need not include names; if it does, then these are checked to be in the same order as in the tree (if check_input==TRUE).

Value

A list with the following elements:

Ntimes	Integer, indicating the number of returned time points. Equal to the provided Ntimes if applicable.
times	Numeric vector of size Ntimes, listing the considered time points in increasing order. If times was provided as an argument to the function, then this will be the same as provided.
clade_counts	Integer vector of size Ntimes, listing the number of tips or nodes considered at each time point.
means	Numeric vector of size Ntimes, listing the arithmetic mean of trait states at each time point.

stds   Numeric vector of size Ntimes, listing the population standard deviation of trait
                states at each time point.

Author(s)

Stilianos Louca

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1), max_tips=1000)$tree

# simulate a numeric trait under Brownian-motion
trait = simulate_bm_model(tree, diffusivity=1)
states = c(trait$tip_states,trait$node_states)

# calculate trait stats over time
results = get_trait_stats_over_time(tree, states, Ntimes=100)

# plot trait stats over time (mean +/- std)
M = results$means
S = results$stds
matplot(x=results$times,
        y=matrix(c(M-S,M+S),ncol=2,byrow=FALSE),
        main = "Simulated BM trait over time",
        lty = 1, col="black",
        type="l", xlab="time", ylab="mean +/- std")
```

---

get_transition_index_matrix
                    Create an index matrix for a Markov transition model.

Description

Create an index matrix encoding the parametric structure of the transition rates in a discrete-state
continuous-time Markov model (e.g., Mk model of trait evolution). Such an index matrix is required
by certain functions for mapping independent rate parameters to transition rates. For example, an
index matrix may encode the information that each rate i–>j is equal to its reversed counterpart j–>i.

Usage

```
get_transition_index_matrix(Nstates, rate_model)
```

Arguments

Nstates   Integer, the number of distinct states represented in the transition matrix (number
                of rows & columns).

rate_model	Rate model that the transition matrix must satisfy. Can be "ER" (all rates equal), "SYM" (transition rate i–>j is equal to transition rate j–>i), "ARD" (all rates can be different) or "SUEDE" (only stepwise transitions i–>i+1 and i–>i-1 allowed, all 'up' transitions are equal, all 'down' transitions are equal).

Details

The returned index matrix will include as many different positive integers as there are independent rate parameters in the requested rate model, plus potentially the value 0 (which has a special meaning, see below).

Value

A named list with the following elements:

index_matrix	Integer matrix of size Nstates x Nstates, with values between 0 and Nstates, assigning each entry in the transition matrix to an independent transition rate parameter. A value of 0 means that the corresponding rate is fixed to zero (if off-diagonal) or will be adjusted to ensure a valid Markov transition rate matrix (if on the diagonal).
Nrates	Integer, the number of independent rate parameters in the model.

Author(s)

Stilianos Louca

See Also

get_random_mk_transition_matrix

---

get_tree_span	*Get min and max distance of any tip to the root.*

Description

Given a rooted phylogenetic tree, calculate the minimum and maximum phylogenetic distance (cumulative branch length) of any tip from the root.

Usage

```
get_tree_span(tree, as_edge_count=FALSE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
as_edge_count	Logical, specifying whether distances should be counted in number of edges, rather than cumulative edge length. This is the same as if all edges had length 1.

Details

If `tree$edge.length` is missing, then every edge in the tree is assumed to be of length 1. The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child). The asymptotic average time complexity of this function is O(Nedges), where Nedges is the number of edges in the tree.

Value

A named list with the following elements:

min_distance	Minimum phylogenetic distance that any of the tips has to the root.
max_distance	Maximum phylogenetic distance that any of the tips has to the root.

Author(s)

Stilianos Louca

See Also

get_pairwise_distances

Examples

```
# generate a random tree
Ntips   = 1000
params  = list(birth_rate_intercept=1, death_rate_intercept=0.5)
tree    = generate_random_tree(params, max_tips=Ntips, coalescent=FALSE)$tree

# calculate min & max tip distances from root
tree_span = get_tree_span(tree)
cat(sprintf("Tip min dist = %g, max dist = %g\n",
            tree_span$min_distance,
            tree_span$max_distance))
```

---

get_tree_traversal_root_to_tips

Traverse tree from root to tips.

Description

Create data structures for traversing a tree from root to tips, and for efficient retrieval of a node's outgoing edges and children.

Usage

```
get_tree_traversal_root_to_tips(tree, include_tips)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
include_tips	Include tips in the tarversal queue. If FALSE, then only nodes are included in the queue.

Details

Many dynamic programming algorithms for phylogenetics involve traversing the tree in a certain direction (root to tips or tips to root), and efficient (O(1) complexity) access to a node's direct children can significantly speed up those algorithms. This function is meant to provide data structures that allow traversing the tree's nodes (and optionally tips) in such an order that each node is traversed prior to its descendants (root–>tips) or such that each node is traversed after its descendants (tips–>root). This function is mainly meant for use in other algorithms, and is probably of little relevance to the average user.

The tree may include multi-furcations as well as mono-furcations (i.e. nodes with only one child).

The asymptotic time and memory complexity of this function is O(Ntips), where Ntips is the number of tips in the tree.

Value

A list with the following elements:

queue	An integer vector of size Nnodes (if include_tips was FALSE) or of size Nnodes+Ntips (if include_tips was TRUE), listing indices of nodes (and optionally tips) in the order root–>tips described above. In particular, queue[1] will be the index of the tree's root (typically Ntips+1).
edges	An integer vector of size Nedges (=nrow(tree$edge)), listing indices of edges (corresponding to tree$edge) such that outgoing edges of the same node are listed in consequtive order.
node2first_edge	An integer vector of size Nnodes listing the location of the first outgoing edge of each node in edges. That is, edges[node2first_edge[n]] points to the first outgoing edge of node n in tree$edge.
node2last_edge	An integer vector of size Nnodes listing the location of the last outgoing edge of each node in edges. That is, edges[node2last_edge[n]] points to the last outgoing edge of node n in tree$edge. The total number of outgoing edges of a node is thus given by 1+node2last_edge[n]-node2first_edge[n].

Author(s)

Stilianos Louca

See Also

reorder_tree_edges

Examples

```
## Not run:
# generate a random tree
tree = generate_random_tree(list(birth_rate_factor=1), max_tips=100)$tree

# get tree traversal
traversal = get_tree_traversal_root_to_tips(tree, include_tips=TRUE)

## End(Not run)
```

---

hsp_binomial   *Hidden state prediction for a binary trait based on the binomial distri-*
                             bution.

Description

Estimate the state probabilities for a binary trait at ancestral nodes and tips with unknown (hidden) state, by fitting the probability parameter of a binomial distribution to empirical state frequencies. For each node, the states of its descending tips are assumed to be drawn randomly and independently according to some a prior unknown probability distribution. The probability P1 (probability of any random descending tip being in state 1) is estimated separately for each node based on the observed states in the descending tips via maximum likelihood.

This function can account for potential state-measurement errors, hidden states and reveal biases (i.e., tips in one particular state being more likely to be measured than in the other state). Only nodes with a number of non-hidden tips above a certain threshold are included in the ML-estimation phase. All other nodes and hidden tips are then assigned the probabilities estimated for the most closely related ancestral node with estimated probabilities. This function is a generalization of [hsp_empirical_probabilities](#) that can account for potential state-measurement errors and reveal biases.

Usage

```
hsp_binomial( tree,
              tip_states,
              reveal_probs  = NULL,
              state1_probs  = NULL,
              min_revealed  = 1,
              max_STE       = Inf,
              check_input   = TRUE)
```

Arguments

tree	A rooted tree of class "phylo".
tip_states	Integer vector of length Ntips, specifying the state of each tip in the tree (either 1 or 2). `tip_states` can include NA to indicate a hidden (non-measured) tip state.

reveal_probs   2D numeric matrix of size Ntips x 2, listing tip-specific reveal probabilities at each tip conditional on the tip's true state. Hence reveal_probs[n,s] is the probability that tip n would have a measured (non-hidden) state if its true state was s. May also be a vector of length 2 (same reveal_probs for all tips) or NULL (unbiased reveal probs).

state1_probs   2D numeric matrix of size Ntips x 2, listing the probability of measuring state 1 (potentially erroneously) at each tip conditional upon its true state and conditional upon its state having been measured (i.e., being non-hidden). For example, for an incompletely sequenced genome with completion level C_n and state 1 indicating presence and state 2 indicating absence of a gene, and assuming error-free detection of genes within the covered regions, one has state1_probs[n,1] = C_n and state1_probs[n,2]=0. state1_probs may also be a vector of length 2 (same probabilities for all tips) or NULL. If NULL, state measurements are assumed error-free, and hence this is the same as c(1,0).

min_revealed   Non-negative integer, specifying the minimum number of tips with non-hidden state that must descend from a node for estimating its P1 via maximum likelihood. For nodes with too few descending tips with non-hidden state, the probability P1 will not be estimated via maximum likelihood, and instead will be set to the P1 estimated for the nearest possible ancestral node. It is advised to set this threshold greater than zero (typical values are 2–10).

max_STE   Non-negative numeric, specifying the maximum acceptable estimated standard error (STE) for the estimated probability P1 for a node. If the STE for a node exceeds this threshold, the P1 for that node is set to the P1 of the nearest ancestor with STE below that threshold. Setting this to Inf disables this functionality. The STE is estimated based on the Observed Fisher Information Criterion (which, strictly speaking, only provides a lower bound for the STE).

check_input   Logical, specifying whether to perform some additional time-consuming checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

This function currently only supports binary traits, and states must be represented by integers 1 or 2. Any NA entries in tip_states are interpreted as hidden (non-revealed) states.

The algorithm proceeds in two phases ("ASR" phase and "HSP" phase). In the ASR phase the state probability P1 is estimated separately for every node and tip satisfying the thresholds min_revealed and max_STE, via maximum-likelihood. In the HSP phase, the P1 of nodes and tips not included in the ASR phase is set to the P1 of the nearest ancestral node with estimated P1, as described by Zaneveld and Thurber (2014).

This function yields estimates for the state probabilities P1 (note that P2=1-P1). In order to obtain point estimates for tip states one needs to interpret these probabilities in a meaningful way, for example by choosing as point estimate for each tip the state with highest probability P1 or P2; the closest that probability is to 1, the more reliable the point estimate will be.

The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). This function has asymptotic time complexity O(Nedges x Nstates). Tips must be represented in tip_states in the same order as in tree$tip.label. The

vector `tip_states` need not include names; if it does, however, they are checked for consistency (if `check_input==TRUE`).

Value

A list with the following elements:

success	Logical, indicating whether HSP was successful. If `FALSE`, an additional element `error` (character) will be returned describing the error, while all other return values may be `NULL`.
P1	Numeric vector of length Ntips+Nnodes, listing the estimated probability of being in state 1 for each tip and node. A value of P1[n]=0 or P1[n]=1 means that the n-th tip/node is in state 2 or state 1 with absolute certainty, respectively. Note that even tips with non-hidden state may have have a P1 that is neither 0 or 1, if state measurements are erroneous (i.e., if `state1_probs[n,]` differs from `(1,0)`).
STE	Numeric vector of length Ntips+Nnodes, listing the standard error of the estimated P1 at each tip and node, according to the Observed Fisher Information Criterion. Note that the latter strictly speaking only provides a lower bound on the standard error.
reveal_counts	Integer vector of length Ntips+Nnodes, listing the number of tips with non-hidden state descending from each tip and node.
inheritted	Logical vector of length Ntips+Nnodes, specifying for each tip or node whether its returned P1 was directly maximum-likelihood estimated duirng the ASR phase (`inheritted[n]==FALSE`) or set to the P1 estimated for an ancestral node during the HSP phase (`inheritted[n]==TRUE`).

Author(s)

Stilianos Louca

References

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

hsp_max_parsimony, hsp_mk_model, hsp_empirical_probabilities

Examples

```
## Not run:
# generate random tree
Ntips =50
tree = generate_random_tree(list(birth_rate_factor=1),max_tips=Ntips)$tree

# simulate a binary trait on the tips
```

```
Q = get_random_mk_transition_matrix(Nstates=2, rate_model="ER", min_rate=0.1, max_rate=0.5)
tip_states = simulate_mk_model(tree, Q)$tip_states

# print tip states
cat(sprintf("True tip states:\n"))
print(tip_states)

# hide some of the tip states
# include a reveal bias
reveal_probs = c(0.8, 0.3)
revealed = sapply(1:Ntips, FUN=function(n) rbinom(n=1,size=1,prob=reveal_probs[tip_states[n]]))
input_tip_states = tip_states
input_tip_states[!revealed] = NA

# predict state probabilities P1 and P2
hsp = hsp_binomial(tree, input_tip_states, reveal_probs=reveal_probs, max_STE=0.2)
probs = cbind(hsp$P1,1-hsp$P1)

# pick most likely state as a point estimate
# only accept point estimate if probability is sufficiently high
estimated_tip_states = max.col(probs[1:Ntips,])
estimated_tip_states[probs[cbind(1:Ntips,estimated_tip_states)]<0.8] = NA
cat(sprintf("ML-predicted tip states:\n"))
print(estimated_tip_states)

# calculate fraction of correct predictions
predicted = which((!revealed) & (!is.na(estimated_tip_states)))
if(length(predicted)>0){
  Ncorrect  = sum(tip_states[predicted]==estimated_tip_states[predicted])
  cat(sprintf("%.2g%% of predictions are correct\n",(100.0*Ncorrect)/length(predicted)))
}else{
  cat(sprintf("None of the tip states could be reliably predicted\n"))
}

## End(Not run)
```

---

hsp_empirical_probabilities

Hidden state prediction via empirical probabilities.

Description

Reconstruct ancestral discrete states of nodes and predict unknown (hidden) states of tips on a tree based on empirical state probabilities across tips. This is a very crude HSP method, and other more sophisticated methods should be preferred (e.g. hsp_mk_model).

Usage

```
hsp_empirical_probabilities(tree, tip_states,
                            Nstates=NULL, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	An integer vector of size Ntips, specifying the state of each tip in the tree as an integer from 1 to Nstates, where Nstates is the possible number of states (see below). tip_states can include NA to indicate an unknown tip state that is to be predicted.
Nstates	Either NULL, or an integer specifying the number of possible states of the trait. If NULL, then it will be computed based on the maximum non-NA value encountered in tip_states
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

For this function, the trait's states must be represented by integers within 1,..,Nstates, where Nstates is the total number of possible states. If the states are originally in some other format (e.g. characters or factors), you should map them to a set of integers 1,..,Nstates. You can easily map any set of discrete states to integers using the function [map_to_state_space](#).

Any NA entries in tip_states are interpreted as unknown states. Prior to ancestral state reconstruction, the tree is temporarily pruned, keeping only tips with known state. The function then calculates the empirical state probabilities for each node in the pruned tree, based on the states across tips descending from each node. The state probabilities of tips with unknown state are set to those of the most recent ancestor with reconstructed states, as described by Zaneveld and Thurber (2014).

The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). This function has asymptotic time complexity O(Nedges x Nstates).

Tips must be represented in tip_states in the same order as in tree$tip.label. The vector tip_states need not include names; if it does, however, they are checked for consistency (if check_input==TRUE).

This function is meant for reconstructing ancestral states in all nodes of a tree as well as predicting the states of tips with an a priori unknown state. If the state of all tips is known and only ancestral state reconstruction is needed, consider using functions such as [asr_empirical_probabilities](#) for improved efficiency.

Value

A list with the following elements:

success	Logical, indicating whether HSP was successful. If FALSE, some return values may be NULL.
likelihoods	A 2D numeric matrix, listing the probability of each tip and node being in each state. This matrix will have (Ntips+Nnodes) rows and Nstates columns, where Nstates was either explicitly provided as an argument or inferred based on the

number of unique values in `tip_states` (if `Nstates` was passed as NULL). In the latter case, the column names of this matrix will be the unique values found in `tip_states`. The rows in this matrix will be in the order in which tips and nodes are indexed in the tree, i.e. the rows 1,..,Ntips store the probabilities for tips, while rows (Ntips+1),..,(Ntips+Nnodes) store the probabilities for nodes. Each row in this matrix will sum up to 1. Note that the return value is named this way for compatibility with other HSP functions.

Author(s)

Stilianos Louca

References

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

[hsp_max_parsimony](#), [hsp_mk_model](#), [map_to_state_space](#)

Examples

```
## Not run:
# generate random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a discrete trait
Nstates = 5
Q = get_random_mk_transition_matrix(Nstates, rate_model="ER", max_rate=0.1)
tip_states = simulate_mk_model(tree, Q)$tip_states

# print states of first 20 tips
print(tip_states[1:20])

# set half of the tips to unknown state
tip_states[sample.int(Ntips,size=as.integer(Ntips/2),replace=FALSE)] = NA

# reconstruct all tip states via MPR
likelihoods = hsp_empirical_probabilities(tree, tip_states, Nstates)$likelihoods
estimated_tip_states = max.col(likelihoods[1:Ntips,])

# print estimated states of first 20 tips
print(estimated_tip_states[1:20])

## End(Not run)
```

---

hsp_independent_contrasts

Hidden state prediction via phylogenetic independent contrasts.

Description

Reconstruct ancestral states of a continuous (numeric) trait for nodes and predict unknown (hidden) states for tips on a tree using phylogenetic independent contrasts.

Usage

```
hsp_independent_contrasts(tree, tip_states, weighted=TRUE, check_input=TRUE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

tip_states   A numeric vector of size Ntips, specifying the state of each tip in the tree. `tip_states` can include NA to indicate an unknown tip state that is to be predicted.

weighted   Logical, specifying whether to weight transition costs by the inverted edge lengths during ancestral state reconstruction. This corresponds to the "weighted squared-change parsimony" reconstruction by Maddison (1991) for a Brownian motion model of trait evolution.

check_input   Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

Any NA entries in `tip_states` are interpreted as unknown (hidden) states to be estimated. Prior to ancestral state reconstruction, the tree is temporarily prunned, keeping only tips with known state. The function then uses a postorder traversal algorithm to calculate the intermediate "X" variables (a state estimate for each node) introduced by Felsenstein (1985) in his phylogenetic independent contrasts method. Note that these are only local estimates, i.e. for each node the estimate is only based on the tip states in the subtree descending from that node (see discussion in Garland and Ives, 2000). The states of tips with hidden state are set to those of the most recent ancestor with reconstructed state, as described by Zaneveld and Thurber (2014).

This function has asymptotic time complexity O(Nedges). If `tree$edge.length` is missing, each edge in the tree is assumed to have length 1. This is the same as setting `weighted=FALSE`. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child).

Tips must be represented in `tip_states` in the same order as in `tree$tip.label`. The vector `tip_states` need not include item names; if it does, however, they are checked for consistency (if `check_input==TRUE`).

This function is meant for reconstructing ancestral states in all nodes of a tree as well as predicting the states of tips with an a priory unknown state. If the state of all tips is known and only ancestral state reconstruction is needed, consider using the function `asr_independent_contrasts` for improved efficiency.

Value

A list with the following elements:

success
: Logical, indicating whether HSP was successful. If `FALSE`, some return values may be `NULL`.

states
: A numeric vector of size Ntips+Nnodes, listing the reconstructed state of each tip and node. The entries in this vector will be in the order in which tips and nodes are indexed in `tree$edge`.

total_sum_of_squared_changes
: The total sum of squared changes in tree, minimized by the (optionally weighted) squared-change parsimony algorithm. This is equation 7 in (Maddison, 1991). Note that for the root, phylogenetic independent contrasts is equivalent to Maddison's squared-change parsimony.

Author(s)

Stilianos Louca

References

J. Felsenstein (1985). Phylogenies and the comparative method. The American Naturalist. 125:1-15.

T. Jr. Garland and A. R. Ives (2000). Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. The American Naturalist. 155:346-364.

W. P. Maddison (1991). Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Systematic Zoology. 40:304-314.

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

`asr_squared_change_parsimony` `hsp_max_parsimony`, `hsp_mk_model`,

Examples

```
# generate random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a continuous trait
tip_states = simulate_ou_model(tree, stationary_mean=0, spread=1, decay_rate=0.001)$tip_states
```

```
# print tip states
print(as.vector(tip_states))

# set half of the tips to unknown state
tip_states[sample.int(Ntips,size=as.integer(Ntips/2),replace=FALSE)] = NA

# reconstruct all tip states via weighted PIC
estimated_states = hsp_independent_contrasts(tree, tip_states, weighted=TRUE)$states

# print estimated tip states
print(estimated_states[1:Ntips])
```

---

hsp_max_parsimony   *Hidden state prediction via maximum parsimony.*

Description

Reconstruct ancestral discrete states of nodes and predict unknown (hidden) states of tips on a tree using maximum parsimony. Transition costs can vary between transitions, and can optionally be weighted by edge length.

Usage

```
hsp_max_parsimony(tree, tip_states, Nstates=NULL,
                  transition_costs="all_equal",
                  edge_exponent=0.0, weight_by_scenarios=TRUE,
                  check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	An integer vector of size Ntips, specifying the state of each tip in the tree as an integer from 1 to Nstates, where Nstates is the possible number of states (see below). tip_states can include NA to indicate an unknown tip state that is to be predicted.
Nstates	Either NULL, or an integer specifying the number of possible states of the trait. If NULL, then it will be computed based on the maximum non-NA value encountered in tip_states

transition_costs

> Same as for the function [asr_max_parsimony](#).

edge_exponent   Same as for the function [asr_max_parsimony](#).

weight_by_scenarios

> Logical, indicating whether to weight each optimal state of a node by the number of optimal maximum-parsimony scenarios in which the node is in that state. If FALSE, then all possible states of a node are weighted equally (i.e. are assigned equal probabilities).

check_input   Logical, specifying whether to perform some basic checks on the validity of the
               input data. If you are certain that your input data are valid, you can set this to
               FALSE to reduce computation.

Details

For this function, the trait's states must be represented by integers within 1,..,Nstates, where Nstates
is the total number of possible states. If the states are originally in some other format (e.g. characters
or factors), you should map them to a set of integers 1,..,Nstates. The order of states (if relevant)
should be reflected in their integer representation. For example, if your original states are "small",
"medium" and "large" and transition_costs=="sequential", it is advised to represent these
states as integers 1,2,3. You can easily map any set of discrete states to integers using the function
[map_to_state_space](#).

Any NA entries in tip_states are interpreted as unknown states. Prior to ancestral state recon-
struction, the tree is temporarily pruned, keeping only tips with known state. The function then
applies Sankoff's (1975) dynamic programming algorithm for ancestral state reconstruction, which
determines the smallest number (or least costly if transition costs are uneven) of state changes along
edges needed to reproduce the known tip states. The state probabilities of tips with unknown state
are set to those of the most recent ancestor with reconstructed states, as described by Zaneveld and
Thurber (2014). This function has asymptotic time complexity O(Ntips+Nnodes x Nstates).

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. If edge_exponent
is 0, then edge lengths do not influence the result. If edge_exponent!=0, then all edges must have
non-zero length. The tree may include multi-furcations (i.e. nodes with more than 2 children) as
well as mono-furcations (i.e. nodes with only one child).

Tips must be represented in tip_states in the same order as in tree$tip.label. None of the
input vectors or matrixes need include row or column names; if they do, however, they are checked
for consistency (if check_input==TRUE).

This function is meant for reconstructing ancestral states in all nodes of a tree as well as predicting
the states of tips with an a priory unknown state. If the state of all tips is known and only ances-
tral state reconstruction is needed, consider using the function [asr_max_parsimony](#) for improved
efficiency.

Value

A list with the following elements:

success   Logical, indicating whether HSP was successful. If FALSE, some return values
               may be NULL.

likelihoods   A 2D numeric matrix, listing the probability of each tip and node being in each
               state. This matrix will have (Ntips+Nnodes) rows and Nstates columns, where
               Nstates was either explicitly provided as an argument or inferred based on the
               number of unique values in tip_states (if Nstates was passed as NULL). In
               the latter case, the column names of this matrix will be the unique values found
               in tip_states. The rows in this matrix will be in the order in which tips and
               nodes are indexed in the tree, i.e. the rows 1,..,Ntips store the probabilities for
               tips, while rows (Ntips+1),..,(Ntips+Nnodes) store the probabilities for nodes.
               Each row in this matrix will sum up to 1. Note that the return value is named
               this way for compatibility with other HSP functions.

Author(s)

Stilianos Louca

References

D. Sankoff (1975). Minimal mutation trees of sequences. SIAM Journal of Applied Mathematics. 28:35-42.

J. Felsenstein (2004). Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

asr_max_parsimony, asr_mk_model, hsp_mk_model, map_to_state_space

Examples

```
## Not run:
# generate random tree
Ntips = 10
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a discrete trait
Nstates = 5
Q = get_random_mk_transition_matrix(Nstates, rate_model="ER")
tip_states = simulate_mk_model(tree, Q)$tip_states

# print tip states
print(tip_states)

# set half of the tips to unknown state
tip_states[sample.int(Ntips,size=as.integer(Ntips/2),replace=FALSE)] = NA

# reconstruct all tip states via MPR
likelihoods = hsp_max_parsimony(tree, tip_states, Nstates)$likelihoods
estimated_tip_states = max.col(likelihoods[1:Ntips,])

# print estimated tip states
print(estimated_tip_states)

## End(Not run)
```

---

hsp_mk_model   *Hidden state prediction with Mk models and rerooting*

Description

Reconstruct ancestral states of a discrete trait and predict unknown (hidden) states of tips using a fixed-rates continuous-time Markov model (a.k.a. "Mk model"). This function can fit the model (i.e. estimate the transition matrix) using maximum likelihood, or use a specified transition matrix. The function can optionally calculate marginal ancestral state likelihoods for each node in the tree, using the rerooting method by Yang et al. (1995). A subset of the tips may have completely unknown states; in this case the fitted Markov model is used to predict their state likelihoods based on their most recent reconstructed ancestor, as described by Zaneveld and Thurber (2014). The function can account for biases in which tips have known state ("reveal bias").

Usage

```
hsp_mk_model( tree,
              tip_states,
              Nstates = NULL,
              reveal_fractions = NULL,
              tip_priors = NULL,
              rate_model = "ER",
              transition_matrix = NULL,
              include_likelihoods = TRUE,
              root_prior = "empirical",
              Ntrials = 1,
              optim_algorithm = "nlminb",
              optim_max_iterations = 200,
              optim_rel_tol = 1e-8,
              store_exponentials = TRUE,
              check_input = TRUE,
              Nthreads = 1)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

tip_states   An integer vector of size Ntips, specifying the state of each tip in the tree in terms of an integer from 1 to Nstates, where Nstates is the possible number of states (see below). Can also be NULL, in which case tip_priors must not be NULL (see below). tip_states can include NA to indicate an unknown (hidden) tip state that is to be predicted.

Nstates   Either NULL, or an integer specifying the number of possible states of the trait. If Nstates==NULL, then it will be computed based on the maximum non-NA value encountered in tip_states or based on the number of columns in tip_priors (whichever is non-NULL).

reveal_fractions

                  Either NULL, or a numeric vector of size Nstates, specifying the fraction of tips with revealed (i.e., non-hidden) state, depending on the tip state. That is, reveal_fractions[s] is the probability that a given tip at state s will have known (i.e., non-hidden) state, conditional upon being included in the tree. If the tree only contains a random subset of species (sampled independently of

each species' state), then reveal_fractions[s] is the probability of knowing the state of a species (regardless of whether it is included in the tree), if its state is s. This variable can be used to account for biases in which tips have known state, depending on their state. Only the relative ratios among reveal fractions matter, i.e. multiplying reveal_fractions with a constant factor has no effect.

tip_priors   A 2D numeric matrix of size Ntips x Nstates, where Nstates is the possible number of states for the character modelled. Can also be NULL. Each row of this matrix must be a probability vector, i.e. it must only contain non-negative entries and must sum up to 1. The [i,s]-th entry should be the prior probability of tip i being in state s. If you know for certain that tip i is in some state s, you can set the corresponding entry to 1 and all other entries in that row to 0. A row can include NA to indicate that neither the state nor the probability distribution of a state are known for that tip. If for all tips you either know the exact state or have no information at all, you can also use tip_states instead. If tip_priors==NULL, then tip_states must not be NULL (see above).

rate_model   Rate model to be used for fitting the transition rate matrix. Similar to the rate_model option in the function [asr_mk_model](). See the details of [asr_mk_model]() on the assumptions of each rate_model.

transition_matrix

Either a numeric quadratic matrix of size Nstates x Nstates containing fixed transition rates, or NULL. The [r,c]-th entry in this matrix should store the transition (probability) rate from the state r to state c. Each row in this matrix must have sum zero. If NULL, then the transition rates will be estimated using maximum likelihood, based on the rate_model specified.

include_likelihoods

Boolean, specifying whether to include the marginal state likelihoods for all tips and nodes, as returned variables. Setting this to TRUE can substantially increase computation time. If FALSE, the Mk model is merely fitted, but ancestral states and hidden tip states are not reconstructed.

root_prior   Prior probability distribution of the root's states. Similar to the root_prior option in the function [asr_mk_model]().

Ntrials   Number of trials (starting points) for fitting the transition matrix. Only relevant if transition_matrix=NULL. A higher number may reduce the risk of landing in a local non-global optimum of the likelihood function, but will increase computation time during fitting.

optim_algorithm

Either "optim" or "nlminb", specifying which optimization algorithm to use for maximum-likelihood estimation of the transition matrix. Only relevant if transition_matrix==NULL.

optim_max_iterations

Maximum number of iterations (per fitting trial) allowed for optimizing the likelihood function.

optim_rel_tol   Relative tolerance (stop criterion) for optimizing the likelihood function.

store_exponentials

Logical, specifying whether to pre-calculate and store exponentials of the transition matrix during calculation of ancestral likelihoods. This may reduce computation time because each exponential is only calculated once, but will use up

more memory since all exponentials are stored. Only relevant if `include_ancestral_likelihoods` is TRUE, otherwise exponentials are never stored.

check_input   Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Nthreads   Number of parallel threads to use for running multiple fitting trials simultaneously. This only makes sense if your computer has multiple cores/CPUs and Ntrials>1, and is only relevant if `transition_matrix==NULL`.

Details

For this function, the trait's states must be represented by integers within 1,..,Nstates, where Nstates is the total number of possible states. Note that Nstates can be chosen to be larger than the number of states observed in the tips of the present tree, to account for potential states not yet observed. If the trait's states are originally in some other format (e.g. characters or factors), you should map them to a set of integers 1,..,Nstates. The order of states (if applicable) should be reflected in their integer representation. For example, if your original states are "small", "medium" and "large" and `rate_model=="SUEDE"`, it is advised to represent these states as integers 1,2,3. You can easily map any set of discrete states to integers using the function [map_to_state_space](#).

This function allows the specification of the precise tip states (if these are known) using the vector `tip_states`. Alternatively, if some tip states are only known in terms of a probability distribution, you can pass these probability distributions using the matrix `tip_priors`. Note that exactly one of the two arguments, `tip_states` or `tip_priors`, must be non-NULL. In either case, the presence of NA in `tip_states` or in a row of `tip_priors` is interpreted as an absence of information about the tip's state (i.e. the tip has "hidden state").

Tips must be represented in `tip_states` or `tip_priors` in the same order as in `tree$tip.label`. None of the input vectors or matrixes need include row or column names; if they do, however, they are checked for consistency (if `check_input==TRUE`).

This method assumes that the tree is either complete (i.e. includes all species), or that the tree's tips represent a random subset of species that have been sampled independent of their state. The function does not require that tip state knowledge is independent of tip state, provided that the associated biases are known (provided via `reveal_fractions`). The rerooting method by Yang et al (2015) is used to reconstruct the marginal ancestral state likelihoods for each node by treating the node as a root and calculating its conditional scaled likelihoods. The state likelihoods of tips with hidden states are calculated from those of the most recent ancestor with previously calculated state likelihoods, using the exponentiated transition matrix along the connecting edges (essentially using the rerooting method). Attention: The state likelihoods for tips with known states or with provided priors are not modified, i.e. they are as provided in the input. In other words, for those tips the returned state likelihoods should not be considered as posteriors in a Bayesian sense.

If `tree$edge.length` is missing, each edge in the tree is assumed to have length 1. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child).

Value

A list with the following elements:

success	Logical, indicating whether HSP was successful. If FALSE, some return values may be NULL.
Nstates	Integer, specifying the number of modeled trait states.
transition_matrix	
	A numeric quadratic matrix of size Nstates x Nstates, containing the transition rates of the Markov model. The [r,c]-th entry is the transition rate from state r to state c. Will be the same as the input transition_matrix, if the latter was not NULL.
loglikelihood	Log-likelihood of the Markov model. If transition_matrix was NULL in the input, then this will be the log-likelihood maximized during fitting.
likelihoods	A 2D numeric matrix, listing the probability of each tip and node being in each state. Only included if include_likelihoods was TRUE. This matrix will have (Ntips+Nnodes) rows and Nstates columns, where Nstates was either explicitly provided as an argument, or inferred from tip_states or tip_priors (whichever was non-NULL). The rows in this matrix will be in the order in which tips and nodes are indexed in the tree, i.e. rows 1,..,Ntips store the probabilities for tips, while rows (Ntips+1),...,(Ntips+Nnodes) store the probabilities for nodes. For example, likelihoods[1,3] will store the probability that tip 1 is in state 3. Each row in this matrix will sum up to 1. Note that for tips with known state or fully provided prior, the likelihoods will be unchanged, i.e. these are not the posteriors in a Bayesian sense.

Author(s)

Stilianos Louca

References

Z. Yang, S. Kumar and M. Nei (1995). A new method for inference of ancestral nucleotide and amino acid sequences. Genetics. 141:1641-1650.

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

[hsp_max_parsimony](#), [hsp_squared_change_parsimony](#), [asr_mk_model](#), [map_to_state_space](#)

Examples

```
## Not run:
# generate random tree
Ntips = 1000
tree  = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a discrete trait
Nstates = 5
Q = get_random_mk_transition_matrix(Nstates, rate_model="ER", max_rate=0.01)
tip_states = simulate_mk_model(tree, Q)$tip_states
```

```
cat(sprintf("Simulated ER transition rate=%g\n",Q[1,2]))

# print states for first 20 tips
print(tip_states[1:20])

# set half of the tips to unknown state
# chose tips randomly, regardless of their state (no biases)
tip_states[sample.int(Ntips,size=as.integer(Ntips/2),replace=FALSE)] = NA

# reconstruct all tip states via Mk model max-likelihood
results = hsp_mk_model(tree, tip_states, Nstates, rate_model="ER", Ntrials=2, Nthreads=2)
estimated_tip_states = max.col(results$likelihoods[1:Ntips,])

# print Mk model fitting summary
cat(sprintf("Mk model: log-likelihood=%g\n",results$loglikelihood))
cat(sprintf("Universal (ER) transition rate=%g\n",results$transition_matrix[1,2]))

# print estimated states for first 20 tips
print(estimated_tip_states[1:20])

## End(Not run)
```

---

hsp_nearest_neighbor   *Hidden state prediction based on nearest neighbor.*

Description

Predict unknown (hidden) character states of tips on a tree using nearest neighbor matching.

Usage

```
hsp_nearest_neighbor(tree, tip_states, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo".
tip_states	A vector of length Ntips, specifying the state of each tip in the tree. Tip states can be any valid data type (e.g., characters, integers, continuous numbers, and so on). NA values denote unknown (hidden) tip states to be predicted.
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

For each tip with unknown state, this function seeks the closest tip with known state, in terms of patristic distance. The state of the closest tip is then used as a prediction of the unknown state. In the case of multiple equal matches, the precise outcome is unpredictable (this is unlikely to occur

if edge lengths are continuous numbers, but may happen frequently if e.g. edge lengths are all of unit length). This algorithm is arguably one of the crudest methods for predicting character states, so use at your own discretion.

Any NA entries in `tip_states` are interpreted as unknown states. If `tree$edge.length` is missing, each edge in the tree is assumed to have length 1. The tree may include multifurcations (i.e. nodes with more than 2 children) as well as monofurcations (i.e. nodes with only one child). Tips must be represented in `tip_states` in the same order as in `tree$tip.label`. `tip_states` need not include names; if names are included, however, they are checked for consistency with the tree's tip labels (if `check_input==TRUE`).

Value

A list with the following elements:

success	Logical, indicating whether HSP was successful. If `FALSE`, some return values may be `NULL`.
states	Vector of length Ntips, listing the known and predicted state for each tip.
nearest_neighbors	
	Integer vector of length Ntips, listing for each tip the index of the nearest tip with known state. Hence, `nearest_neighbors[n]` specifies the tip from which the unknown state of tip n was inferred. If tip n had known state, `nearest_neighbors[n]` will be n.
nearest_distances	
	Numeric vector of length Ntips, listing for each tip the patristic distance to the nearest tip with known state. For tips with known state, distances will be zero.

Author(s)

Stilianos Louca

References

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

[hsp_max_parsimony](), [hsp_mk_model](),

Examples

```
## Not run:
# generate random tree
Ntips = 20
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a binary trait
Q = get_random_mk_transition_matrix(2, rate_model="ER")
tip_states = simulate_mk_model(tree, Q)$tip_states
```

```
# print tip states
print(tip_states)

# set half of the tips to unknown state
tip_states[sample.int(Ntips,size=as.integer(Ntips/2),replace=FALSE)] = NA

# reconstruct all tip states via nearest neighbor
predicted_states = hsp_nearest_neighbor(tree, tip_states)$states

# print predicted tip states
print(predicted_states)

## End(Not run)
```

hsp_squared_change_parsimony

Hidden state prediction via squared-change parsimony.

Description

Reconstruct ancestral states of a continuous (numeric) trait for nodes and predict unknown (hidden) states for tips on a tree using squared-change (or weighted squared-change) parsimony (Maddison 1991).

Usage

```
hsp_squared_change_parsimony(tree, tip_states, weighted=TRUE, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
tip_states	A numeric vector of size Ntips, specifying the state of each tip in the tree. tip_states can include NA to indicate an unknown tip state that is to be predicted.
weighted	Logical, specifying whether to weight transition costs by the inverted edge lengths during ancestral state reconstruction. This corresponds to the "weighted squared-change parsimony" reconstruction by Maddison (1991) for a Brownian motion model of trait evolution.
check_input	Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

Any NA entries in `tip_states` are interpreted as unknown (hidden) states to be estimated. Prior to ancestral state reconstruction, the tree is temporarily pruned, keeping only tips with known state. The function then uses Maddison's squared-change parsimony algorithm to reconstruct the globally parsimonious state at each node (Maddison 1991). The states of tips with hidden state are set to those of the most recent ancestor with reconstructed state, as described by Zaneveld and Thurber (2014). This function has asymptotic time complexity O(Nedges). If `tree$edge.length` is missing, each edge in the tree is assumed to have length 1. This is the same as setting `weighted=FALSE`. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child).

Tips must be represented in `tip_states` in the same order as in `tree$tip.label`. The vector `tip_states` need not include item names; if it does, however, they are checked for consistency (if `check_input==TRUE`).

This function is meant for reconstructing ancestral states in all nodes of a tree as well as predicting the states of tips with an a priori unknown state. If the state of all tips is known and only ancestral state reconstruction is needed, consider using the function `asr_squared_change_parsimony` for improved efficiency.

Value

A list with the following elements:

states   A numeric vector of size Ntips+Nnodes, listing the reconstructed state of each tip and node. The entries in this vector will be in the order in which tips and nodes are indexed in `tree$edge`.

total_sum_of_squared_changes

The total sum of squared changes, minimized by the (optionally weighted) squared-change parsimony algorithm. This is equation 7 in (Maddison, 1991).

Author(s)

Stilianos Louca

References

W. P. Maddison (1991). Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. Systematic Zoology. 40:304-314.

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

`asr_squared_change_parsimony` `hsp_max_parsimony`, `hsp_mk_model`, `map_to_state_space`

Examples

```
# generate random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a continuous trait
tip_states = simulate_ou_model(tree, stationary_mean=0, spread=1, decay_rate=0.001)$tip_states

# print tip states
print(tip_states)

# set half of the tips to unknown state
tip_states[sample.int(Ntips,size=as.integer(Ntips/2),replace=FALSE)] = NA

# reconstruct all tip states via weighted SCP
estimated_states = hsp_squared_change_parsimony(tree, tip_states, weighted=TRUE)$states

# print estimated tip states
print(estimated_states[1:Ntips])
```

---

hsp_subtree_averaging   *Hidden state prediction via subtree averaging.*

Description

Reconstruct ancestral states of a continuous (numeric) trait for nodes and predict unknown (hidden) states for tips on a tree using subtree averaging.

Usage

```
hsp_subtree_averaging(tree, tip_states, check_input=TRUE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

tip_states   A numeric vector of size Ntips, specifying the state of each tip in the tree. tip_states can include NA to indicate an unknown tip state that is to be predicted.

check_input   Logical, specifying whether to perform some basic checks on the validity of the input data. If you are certain that your input data are valid, you can set this to FALSE to reduce computation.

Details

Any NA entries in `tip_states` are interpreted as unknown (hidden) states to be estimated. For each node the reconstructed state is set to the arithmetic average state of all tips with known state and descending from that node. For each tip with hidden state and each node whose descending tips all have hidden states, the state is set to the state of the closest ancestral node with known or reconstructed state, while traversing from root to tips (Zaneveld and Thurber 2014). Note that reconstructed node states are only local estimates, i.e. for each node the estimate is only based on the tip states in the subtree descending from that node.

Tips must be represented in `tip_states` in the same order as in `tree$tip.label`. The vector `tip_states` need not include item names; if it does, however, they are checked for consistency (if `check_input==TRUE`). This function has asymptotic time complexity O(Nedges).

This function is meant for reconstructing ancestral states in all nodes of a tree as well as predicting the states of tips with an a priori unknown state. If the state of all tips is known and only ancestral state reconstruction is needed, consider using the function [asr_subtree_averaging](#) for improved efficiency.

Value

A list with the following elements:

success	Logical, indicating whether HSP was successful.
states	A numeric vector of size Ntips+Nnodes, listing the reconstructed state of each tip and node. The entries in this vector will be in the order in which tips and nodes are indexed in `tree$edge`.

Author(s)

Stilianos Louca

References

J. R. Zaneveld and R. L. V. Thurber (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 5:431.

See Also

[asr_subtree_averaging](#), [hsp_squared_change_parsimony](#)

Examples

```
# generate random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# simulate a continuous trait
tip_states = simulate_ou_model(tree, stationary_mean=0, spread=1, decay_rate=0.001)$tip_states

# print tip states
```

```
print(as.vector(tip_states))

# set half of the tips to unknown state
tip_states[sample.int(Ntips,size=as.integer(Ntips/2),replace=FALSE)] = NA

# reconstruct all tip states via subtree averaging
estimated_states = hsp_subtree_averaging(tree, tip_states)$states

# print estimated tip states
print(estimated_states[1:Ntips])
```

---

is_monophyletic	*Determine if a set of tips is monophyletic.*

Description

Given a rooted phylogenetic tree and a set of focal tips, this function determines whether the tips form a monophyletic group.

Usage

```
is_monophyletic(tree, focal_tips, check_input=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
focal_tips	Either an integer vector or a character vector, listing the tips to be checked for monophyly. If an integer vector, it should list tip indices (i.e. from 1 to Ntips). If a character vector, it should list tip names; in that case tree$tip.label must exist.
check_input	Logical, whether to perform basic validations of the input data. If you know for certain that your input is valid, you can set this to FALSE to reduce computation time.

Details

This function first finds the most recent common ancestor (MRCA) of the focal tips, and then checks if all tips descending from that MRCA fall within the focal tip set.

Value

A logical, indicating whether the focal tips form a monophyletic set.

Author(s)

Stilianos Louca

See Also

[get_mrca_of_set](#)

Examples

```
# generate random tree
Ntips = 100
tree  = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# pick a random subset of focal tips
focal_tips = which(sample.int(2,size=Ntips,replace=TRUE)==1)

# check if focal tips form a monophyletic group
is_monophyletic(tree, focal_tips)
```

---

join_rooted_trees   *Join two rooted trees.*

Description

Given two rooted phylogenetic trees, place one tree (tree2) onto an edge of the other tree (tree1), so that tree2 becomes a monophyletic group of the final joined tree. As a special case, this function can join two trees at their roots, i.e. so that both are disjoint monophyletic clades of the final tree, splitting at the new root.

Usage

```
join_rooted_trees(  tree1,
                    tree2,
                    target_edge1,
                    target_edge_length1,
                    root_edge_length2)
```

Arguments

tree1	A rooted tree of class "phylo".
tree2	A rooted tree of class "phylo". This tree will become a monophyletic subclade of the final joined tree.
target_edge1	Integer, edge index in tree1 onto which tree2 is to be joined. If <=0, then this refers to the hypothetical edge leading into the root of tree1, in which case both trees will become disjoint monophyletic subclades of the final joined tree.
target_edge_length1	
	Numeric, length of the edge segment in tree1 from the joining-point to the next child node, i.e. how far from the child of target_edge1 should the joining occur. If target_edge1<=0, then target_edge_length1 is the distance of the root of tree1 from the final joined tree's root.
root_edge_length2	
	Numeric, length of the edge leading into the root of tree2, i.e. the distance from the joining point to the root of tree2.

Details

The input trees may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). If any of the input trees does not have edge lengths (i.e., edge.length is NULL), then its edge lengths are assumed to all be 1.

The tips of the two input trees will become the tips of the final joined tree. The nodes of the two input trees will become nodes of the final joined tree, however one additional node will be added at the joining point. Tip labels and node labels (if available) of the joined tree are inheritted from the two input trees.

Value

A list with the following elements:

tree   A new rooted tree of class "phylo", representing the joined tree.

clade1_to_clade

                 Integer vector of length Ntips1+Nnodes1, mapping tip/node indices of the input tree1 to tip/node indices in the final joined tree.

clade2_to_clade

                 Integer vector of length Ntips2+Nnodes2, mapping tip/node indices of the input tree2 to tip/node indices in the final joined tree.

Author(s)

Stilianos Louca

See Also

[split_tree_at_height](#)

Examples

```
# generate two random trees, include tip & node names
tree1 = generate_random_tree(list(birth_rate_intercept=1),
                             max_tips=10,
                             tip_basename="tip1.",
                             node_basename="node1.")$tree
tree2 = generate_random_tree(list(birth_rate_intercept=1),
                             max_tips=5,
                             tip_basename="tip2.",
                             node_basename="node2.")$tree

# join trees at their roots
# each subtree's root should have distance 1 from the new root
joined_tree = join_rooted_trees(tree1,
                                tree2,
                                target_edge1=0,
                                target_edge_length1=1,
                                root_edge_length2=1)$tree
```

---

loglikelihood_hbd   *Galculate the log-likelihood of a homogenous birth-death model.*

Description

Given a rooted ultrametric timetree, and a homogenous birth-death (HBD) model, i.e., with specia-
tion rate λ, extinction rate μ and sampling fraction ρ, calculate the likelihood of the tree under the
model. The speciation and extinction rates may be time-dependent. "Homogenous" refers to the
assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction
rates (in the literature this is sometimes referred to simply as "birth-death model"). Alternatively to
λ and μ, the likelihood may also be calculated based on the pulled diversification rate (PDR; Louca
et al. 2018) and the product $\rho(0) \cdot \lambda(0)$, or based on the pulled speciation rate (PSR). In either case,
the time-profiles of λ, μ, the PDR or the PSR are specified as piecewise polynomially functions
(splines), defined on a discrete grid of ages.

Usage

```
loglikelihood_hbd(tree,
                  oldest_age        = NULL,
                  age0              = 0,
                  rho0              = NULL,
                  rholambda0        = NULL,
                  age_grid          = NULL,
                  lambda            = NULL,
                  mu                = NULL,
                  PDR               = NULL,
                  PSR               = NULL,
                  splines_degree    = 1,
                  condition         = "auto",
                  max_model_runtime = -1,
                  relative_dt       = 1e-3)
```

Arguments

tree   A rooted ultrametric tree of class "phylo".

oldest_age Strictly positive numeric, specifying the oldest time before present ("age") to
            consider when calculating the likelihood. If this is equal to or greater than the
            root age, then oldest_age is taken as the stem age, and the classical formula by
            Morlon et al. (2011) is used. If oldest_age is less than the root age, the tree is
            split into multiple subtrees at that age by treating every edge crossing that age as
            the stem of a subtree, and each subtree is considered an independent realization
            of the HBD model stemming at that age. This can be useful for avoiding points
            in the tree close to the root, where estimation uncertainty is generally higher. If
            oldest_age==NULL, it is automatically set to the root age.

age0   Non-negative numeric, specifying the youngest age (time before present) to con-
            sider for fitting, and with respect to which rho and rholambda0 are defined. If

age0>0, then rho refers to the sampling fraction at age age0, and rholambda0 to the product between rho and the speciation rate at age age0. See below for more details.

rho0   Numeric between 0 (exclusive) and 1 (inclusive), specifying the sampling fraction of the tree at age0, i.e. the fraction of lineages extant at age0 that are included in the tree. Note that if $rho0 < 1$, lineages extant at age0 are assumed to have been sampled randomly at equal probabilities. Can also be NULL, in which case rholambda0 and PDR (see below) must be provided.

rholambda0   Strictly positive numeric, specifying the product of the sampling fraction and the speciation rateat age0, units 1/time. Can be NULL, in which case rarefaction, lambda and mu must be provided.

age_grid   Numeric vector, listing discrete ages (time before present) on which either λ and μ, or the PDR, are specified. Listed ages must be strictly increasing, and must cover at least the full considered age interval (from age0 to oldest_age). Can also be NULL or a vector of size 1, in which case the speciation rate, extinction rate and PDR are assumed to be time-independent.

lambda   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing speciation rates (in units 1/time) at the ages listed in age_grid. Speciation rates should be non-negative, and are assumed to vary polynomially between grid points (see argument splines_degree). If NULL, then either PDR and rholambda0, or PSR alone, must be provided.

mu   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing extinction rates (in units 1/time)at the ages listed in age_grid. Extinction rates should be non-negative, and are assumed to vary polynomially between grid points (see argument splines_degree). If NULL, then PDR and rholambda0, or PSR alone, must be provided.

PDR   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing pulled diversification rates (in units 1/time) at the ages listed in age_grid. PDRs can be negative or positive, and are assumed to vary polynomially between grid points (see argument splines_degree). If NULL, then either lambda and mu, or PSR alone, must be provided.

PSR   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing pulled speciation rates (in units 1/time) at the ages listed in age_grid. PSRs should be non-negative, and are assumed to vary polynomially between grid points (see argument splines_degree). If NULL, then either lambda and mu, or PDR and rholambda0, must be provided.

splines_degree   Integer, either 0,1,2 or 3, specifying the polynomial degree of the provided lambda, mu, PDR and PSR (whichever applicable) between grid points in age_grid. For example, if splines_degree==1, then the provided lambda, mu, PDR and PSR are interpreted as piecewise-linear curves; if splines_degree==2 they are interpreted as quadratic splines; if splines_degree==3 they are interpreted as cubic splines. The splines_degree influences the analytical properties of the curve, e.g. splines_degree==1 guarantees a continuous curve, splines_degree==2 guarantees a continuous curve and continuous derivative, and so on.

condition   Character, either "crown", "stem", "auto" or "none" (the last one is only available if lambda and mu are given), specifying on what to condition the likelihood.

If "crown", the likelihood is conditioned on the survival of the two daughter lineages branching off at the root. If "stem", the likelihood is conditioned on the survival of the stem lineage. Note that "crown" really only makes sense when `oldest_age` is equal to the root age, while "stem" is recommended if `oldest_age` differs from the root age. "none" is usually not recommended and is only available when `lambda` and `mu` are provided. If "auto", the condition is chosen according to the recommendations mentioned earlier.

`max_model_runtime`

Numeric, maximum allowed runtime (in seconds) for evaluating the likelihood. If the likelihood calculation takes longer than this (appoximate) threshold, it halts and returns with an error. If negative (default), this option is ignored.

`relative_dt`   Strictly positive numeric (unitless), specifying the maximum relative time step allowed for integration over time. Smaller values increase integration accuracy but increase computation time. Typical values are 0.0001-0.001. The default is usually sufficient.

Details

If `age0>0`, the input tree is essentially trimmed at `age0` (omitting anything younger than `age0`), and the is likelihood calculated for the trimmed tree while shifting time appropriately. In that case, `rho0` is interpreted as the sampling fraction at `age0`, i.e. the fraction of lineages extant at `age0` that are repreented in the tree. Similarly, `rholambda0` is the product of the sampling fraction and λ at `age0`.

This function supports three alternative parameterizations of HBD models, either using the speciation and extinction rates and sampling fraction (λ, μ and $\rho(\tau_o)$ (for some arbitrary age τ_o), or using the pulled diversification rate (PDR) and the product $\rho(\tau_o) \cdot \lambda(\tau_o$ (sampling fraction times speciation rate at τ_o), or using the pulled speciation rate (PSR). The latter two options should be interpreted as a parameterization of congruence classes, i.e. sets of models that have the same likelihood, rather than specific models, since multiple combinations of λ, μ and $\rho(\tau_o)$ can have identical PDRs, $\rho(\tau_o) \cdot \lambda(\tau_o)$ and PSRs (Louca and Pennell, in review).

For large trees the asymptotic time complexity of this function is O(Nips). The tree may include monofurcations as well as multifurcations, and the likelihood formula accounts for those (i.e., as if monofurcations were omitted and multifurcations were expanded into bifurcations).

Value

A named list with the following elements:

`success`   Logical, indicating whether the calculation was successful. If `FALSE`, then the returned list includes an additional 'error' element (character) containing a description of the error; all other return variables may be undefined.

`loglikelihood` Numeric. If `success==TRUE`, this will be the natural logarithm of the likelihood of the tree under the given model.

Author(s)

Stilianos Louca

References

H. Morlon, T. L. Parsons, J. B. Plotkin (2011). Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences. 108:16327-16332.

S. Louca et al. (2018). Bacterial diversification through geological time. Nature Ecology & Evolution. 2:1458-1467.

S. Louca and M. W. Pennell (in review as of 2019)

See Also

simulate_deterministic_hbd

fit_hbd_model_parametric

fit_hbd_model_on_grid

fit_hbd_pdr_on_grid

fit_hbd_pdr_parametric

Examples

```
# generate a random tree with constant rates
Ntips  = 100
params = list(birth_rate_factor=1, death_rate_factor=0.2, rarefaction=0.5)
tree   = generate_random_tree(params, max_tips=Ntips, coalescent=TRUE)$tree

# get the loglikelihood for an HBD model with the same parameters that generated the tree
# in particular, assuming time-independent speciation & extinction rates
LL = loglikelihood_hbd( tree,
                        rho0     = params$rarefaction,
                        age_grid = NULL, # assume time-independent rates
                        lambda   = params$birth_rate_factor,
                        mu       = params$death_rate_factor)
if(LL$success){
  cat(sprintf("Loglikelihood for constant-rates model = %g\n",LL$loglikelihood))
}

# get the likelihood for a model with exponentially decreasing (in forward time) lambda & mu
beta     = 0.01 # exponential decay rate of lambda over time
age_grid = seq(from=0, to=100, by=0.1) # choose a sufficiently fine age grid
lambda   = 1*exp(beta*age_grid) # define lambda on the age grid
mu       = 0.2*lambda # assume similarly shaped but smaller mu
LL = loglikelihood_hbd( tree,
                        rho0     = params$rarefaction,
                        age_grid = age_grid,
                        lambda   = lambda,
                        mu       = mu)
if(LL$success){
  cat(sprintf("Loglikelihood for exponential-rates model = %g\n",LL$loglikelihood))
}
```

---

map_to_state_space   *Map states of a discrete trait to integers.*

Description

Given a list of states (e.g., for each tip in a tree), map the unique states to integers 1,..,Nstates, where Nstates is the number of possible states. This function can be used to translate states that are originally represented by characters or factors, into integer states as required by ancestral state reconstruction and hidden state prediction functions in this package.

Usage

```
map_to_state_space(raw_states, fill_gaps=FALSE,
                   sort_order="natural", include_state_values=FALSE)
```

Arguments

raw_states	A vector of values (states), each of which can be converted to a different character. This list can include the same value multiple times, for example if values represent the trait's states for tips in a tree.
fill_gaps	Logical. If TRUE, then states are converted to integers using as.integer(as.character()), and then all missing intermediate integer values are included as additional possible states. For example, if raw_states contained the values 2,4,6, then 3 and 5 are assumed to also be possible states.
sort_order	Character, specifying the order in which raw_states should be mapped to ascending integers. Either "natural" or "alphabetical". If "natural", numerical parts of characters are sorted numerically, e.g. as in "3"<"a2"<"a12"<"b1".
include_state_values	
	Logical, specifying whether to also return a numerical version of the unique states. For example, the states "3","a2","4.5" will be mapped to the numeric values 3, NA, 4.5.

Details

Several ancestral state reconstruction and hidden state prediction algorithms in the castor package (e.g., asr_max_parsimony) require that the focal trait's states are represented by integer indices within 1,..,Nstates. These indices are then associated, afor example, with column and row indices in the transition cost matrix (in the case of maximum parsimony reconstruction) or with column indices in the returned matrix containing marginal ancestral state probabilities (e.g., in [asr_mk_model](#)). The function map_to_state_space can be used to conveniently convert a set of discrete states into integers, for use with the aforementioned algorithms.

Value

A list with the following elements:

Nstates	Integer. Number of possible states for the trait, based on the unique values encountered in `raw_states` (after conversion to characters). This may be larger than the number of unique values in `raw_states`, if `fill_gaps` was set to TRUE.
state_names	Character vector of size Nstates, storing the original name (character version) of each state. For example, if `raw_states` was c("b1","3","a12","a2","b1","a2") and `sort_order=="natural"`, then `Nstates` will be 4 and `state_names` will be c("3","a2","a12","b1").
state_values	Optional, only included if `include_state_values==TRUE`. A numeric vector of size `Nstates`, providing the numerical value for each unique state.
mapped_states	Integer vector of size equal to `length(raw_states)`, listing the integer representation of each value in `raw_states`.
name2index	An integer vector of size Nstates, with `names(name2index)` set to `state_names`. This vector can be used to map any new list of states (in character format) to their integer representation. In particular, `name2index[as.character(raw_states)]` is equal to `mapped_states`.

Author(s)

Stilianos Louca

Examples

```
# generate a sequence of random states
unique_states = c("b","c","a")
raw_states = unique_states[sample.int(3,size=10,replace=TRUE)]

# map to integer state space
mapping = map_to_state_space(raw_states)

cat(sprintf("Checking that original unique states is the same as the one inferred:\n"))
print(unique_states)
print(mapping$state_names)

cat(sprintf("Checking reversibility of mapping:\n"))
print(raw_states)
print(mapping$state_names[mapping$mapped_states])
```

---

merge_nodes_to_multifurcations

Merge specific nodes into multifurcations.

Description

Given a rooted tree, merge one or more nodes "upwards" into their parent nodes, thus effectively generating multifurcations. Multiple generations of nodes (i.e., successive branching points) can be merged into a single "absorbing ancestor".

Usage

```
merge_nodes_to_multifurcations( tree,
                                 nodes_to_merge,
                                 merge_with_parents  = FALSE,
                                 keep_ancestral_ages = FALSE)
```

Arguments

tree   A rooted tree of class "phylo".

nodes_to_merge   Integer vector or character vector, listing nodes in the tree that should be merged with their parents (if `merge_with_parents=TRUE`) or with their children (if `merge_with_parents=FALSE`). If an integer vector, it must contain values in 1,..,Nnodes. If a character vector, it must list node labels, and the tree itself must also include node labels.

merge_with_parents

> Logical, specifying whether the nodes listed in `nodes_to_merge` should be merged with their parents. If `FALSE`, the specified nodes will be merged with their children (whenever these are not tips).

keep_ancestral_ages

> Logical, specifying whether the generated multifurcations should have the same age as the absorbing ancestor. If `FALSE`, then the age of a multifurcation will be the average of the absorbing ancestor's age and the ages of its merged child nodes (but constrained from below by the ages of non-merged descendants to avoid negative edge lengths). If `TRUE`, then the ages of multifurcations will be biased towards the root, since their age will be that of the absorbing ancestor.

Details

All tips in the input tree are kept and retain their original indices, however the returned tree will include fewer nodes and edges. Edge and node indices may change. When a node is merged into its parent, the incoming edge is lost, and the parent's age remains unchanged.

Nodes are merged in order from root to tips. Hence, if a node B is merged into ("absorbed by") its parent node A, and child node C is merged into node B, then effectively C ends up merged into node A (node A is the "absorbing ancestor").

If `tree$edge.length` is missing, then all edges in the input tree are assumed to have length 1.

Value

A list with the following elements:

tree   A new tree of class "phylo". The number of nodes in this tree, Nnodes_new, will generally be lower than of the input tree.

new2old_node   Integer vector of length Nnodes_new, mapping node indices in the new tree to node indices in the old tree. Note that nodes merged with their parents are not represented in this list.

old2new_node   Integer vector of length Nnodes, mapping node indices in the old tree to node indices in the new tree. Nodes merged with their parents (and thus missing from the new tree) will have value 0.

Nnodes_removed	Integer. Number of nodes removed from the tree, due to being merged into their parents.
Nedges_removed	Integer. Number of edges removed from the tree.

Author(s)

Stilianos Louca

See Also

[multifurcations_to_bifurcations](), [collapse_monofurcations]()

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1), max_tips=Ntips)$tree

# merge a few nodes with their parents,
# thus obtaining a multifurcating tree
nodes_to_merge = c(1,3,4)
new_tree = merge_nodes_to_multifurcations(tree, nodes_to_merge)$tree

# print summary of old and new tree
cat(sprintf("Old tree has %d nodes\n",tree$Nnode))
cat(sprintf("New tree has %d nodes\n",new_tree$Nnode))
```

---

merge_short_edges	*Eliminate short edges in a tree by merging nodes into multifurcations.*

Description

Given a rooted phylogenetic tree and an edge length threshold, merge nodes/tips into multifurcations when their incoming edges are shorter than the threshold.

Usage

```
merge_short_edges(tree,
                  edge_length_epsilon = 0,
                  force_keep_tips     = TRUE,
                  new_tip_prefix      = "ex.node.tip.")
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

edge_length_epsilon

        Non-negative numeric, specifying the maximum edge length for an edge to be considered "short" and thus to be eliminated. Typically 0 or some small positive number.

force_keep_tips

        Logical. If TRUE, then tips are always kept, even if their incoming edges are shorter than edge_length_epsilon. If FALSE, then tips with short incoming edges are removed from the tree; in that case some nodes may become tips.

new_tip_prefix   Character or NULL, specifying the prefix to use for new tip labels stemming from nodes. Only relevant if force_keep_tips==FALSE. If NULL, then labels of tips stemming from nodes will be the node labels from the original tree (in this case the original tree should include node labels).

Details

The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). Whenever a short edge is eliminated, the edges originating from its child are elongated according to the short edge's length. The corresponding grand-children become children of the short edge's parent. Short edges are eliminated in a depth-first-search manner, i.e. traversing from the root to the tips.

Note that existing monofurcations are retained. If force_keep_tips==FALSE, then new monofurcations may also be introduced due to tips being removed.

This function is conceptually similar to the function ape::di2multi.

Value

A list with the following elements:

tree   A new rooted tree of class "phylo", containing the (potentially multifurcating) tree.

new2old_clade   Integer vector of length equal to the number of tips+nodes in the new tree, with values in 1,..,Ntips+Nnodes, mapping tip/node indices of the new tree to tip/node indices in the original tree.

new2old_edge   Integer vector of length equal to the number of edges in the new tree, with values in 1,..,Nedges, mapping edge indices of the new tree to edge indices in the original tree.

Nedges_removed Integer. Number of edges that have been eliminated.

Author(s)

Stilianos Louca

See Also

[multifurcations_to_bifurcations](#)

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_factor=1),max_tips=Ntips)$tree

# set some edge lengths to zero
tree$edge.length[sample.int(n=Ntips, size=10, replace=FALSE)] = 0

# print number of edges
cat(sprintf("Original tree has %d edges\n",nrow(tree$edge)))

# eliminate any edges of length zero
merged = merge_short_edges(tree, edge_length_epsilon=0)$tree

# print number of edges
cat(sprintf("New tree has %d edges\n",nrow(merged$edge)))
```

---

model_adequacy_hbd   *Check if a birth-death model adequately explains a timetree.*

Description

Given a rooted ultrametric timetree and a homogenous birth-death model, check if the model adequately explains various aspects of the tree, such as the branch length and node age distributions and other test statistics. The function uses bootstrapping to simulate multiple hypothetical trees according to the model and then compares the distribution of those trees to the original tree. This function may be used to quantify the "goodness of fit" of a birth-death model to a timetree.

Usage

```
model_adequacy_hbd( tree,
                    models,
                    splines_degree  = 1,
                    extrapolate     = FALSE,
                    Nbootstraps     = 1000,
                    Nthreads        = 1)
```

Arguments

tree   A rooted ultrametric timetree of class "phylo".

models   Either a single HBD model or a list of HBD models, specifying the pool of models from which to randomly draw bootstraps. Every model should itself be a named list with some or all of the following elements:

                ages: Numeric vector, specifying discrete ages (times before present) in ascending order, on which the pulled speciation rate will be specified. Age increases from tips to root; the youngest tip in the input tree has age 0. The age grid must cover the present-day (age 0) and the root

- **PSR:** Numeric vector of size NG, listing the pulled speciation rate (PSR) of the HBD model at the corresponding ages. Between grid points, the PSR is assumed to either be constant (if splines_degree=0), or linearly (if splines_degree=1) or quadratically (if splines_degree=2) or cubically (if splines_degree=3). To calculate the PSR of an HBD model based on the speciation and extinction rate, see simulate_deterministic_hbd.

splines_degree   Integer, one of 0, 1, 2 or 3, specifying the polynomial degree of the PSR between age-grid points. For example, splines_degree=0 means piecewise constant, splines_degree=1 means piecewise linear and so on.

extrapolate   Logical, specifying whether to extrapolate the model variables λ, μ, ψ and κ (as constants) beyond the provided age grid all the way to stem_age and end_age if needed.

Nbootstraps   Integer, the number of bootstraps (simulations) to perform for calculating statistical significances. A larger number will increase the accuracy of estimated statistical significances.

Nthreads   Integer, number of parallel threads to use for bootstrapping. Note that on Windows machines this option is ignored.

Details

In addition to model selection, the adequacy of any chosen model should also be assessed in absolute terms, i.e. not just relative to other competing models (after all, all considered models might be bad). This function essentially determines how probable it is for hypothetical trees generated by a candidate model to resemble the tree at hand, in terms of various test statistics (such as the historically popular "gamma" statistic, or the Colless tree imbalance). In particular, the function uses a Kolmogorov-Smirnov test to check whether the probability distributions of edge lengths and node ages in the tree resemble those expected under the model. All statistical significances are calculated using bootstrapping, i.e. by simulating trees from the provided model with the same number of tips and the same root age as the original tree.

Note that even if an HBD model appears to adequately explain a given timetree, this does not mean that the model even approximately resembles the true diversification history (i.e., the true speciation and extinction rates) that generated the tree (Louca and Pennell 2020). Hence, it is generally more appropriate to say that a given model "congruence class" (or PSR) rather than a specific model (or speciation/extinction rate) explains the tree.

Value

A named list with the following elements:

success   Logical, indicating whether the model evaluation was successful. If FALSE, then an additional return variable, error, will contain a description of the error; in that case all other return variables may be undefined. Note that success does not say whether the model explains the tree, but rather whether the computation was performed without errors.

Nbootstraps   Integer, the number of bootstraps used.

tree_gamma   Numeric, gamma statistic (Pybus and Harvey 2000) of the original tree.

bootstrap_mean_gamma

        Numeric, mean gamma statistic across all bootstrap trees.

Pgamma   Numeric, two-sided statistical significance of the tree's gamma statistic under the provided null model, i.e. the probability that `abs(bootstrap_mean_gamma-tree_gamma)` would be as large as observed.

tree_Colless   Numeric, Colless imbalance statistic (Shao and Sokal, 1990) of the original tree.

bootstrap_mean_Colless

        Numeric, mean Colless statistic across all bootstrap trees.

PColless   Numeric, two-sided statistical significance of the tree's Colless statistic under the provided null model, i.e. the probability that `abs(bootstrap_mean_Colless-tree_Colless)` would be as large as observed.

tree_Sackin   Numeric, Sackin statistic (Sackin, 1972) of the original tree.

bootstrap_mean_Sackin

        Numeric, median Sackin statistic across all bootstrap trees.

PSackin   Numeric, two-sided statistical significance of the tree's Sackin statistic under the provided null model, i.e. the probability that `abs(bootstrap_mean_Sackin-tree_Sackin)` would be as large as observed.

tree_edgeKS   Numeric, Kolmogorov-Smirnov (KS) statistic of the original tree's edge lengths, i.e. the estimated maximum difference between the tree's and the model's (estimated) cumulative distribution function of edge lengths.

bootstrap_mean_edgeKS

        Numeric, mean KS statistic of the bootstrap trees' edge lengths.

PedgeKS   Numeric, the one-sided statistical significance of the tree's edge-length KS statistic, i.e. the probability that the KS statistic of any tree generated by the model would be larger than the original tree's KS statistic. A low value means that the probability distribution of edge lengths in the original tree differs strongly from that expected based on the model.

tree_nodeKS   Numeric, Kolmogorov-Smirnov (KS) statistic of the original tree's node ages (divergence times), i.e. the estimated maximum difference between the tree's and the model's (estimated) cumulative distribution function of node ages.

bootstrap_mean_nodeKS

        Numeric, mean KS statistic of the bootstrap trees' node ages.

PnodeKS   Numeric, the one-sided statistical significance of the tree's node-age KS statistic, i.e. the probability that the KS statistic of any tree generated by the model would be larger than the original tree's KS statistic. A low value means that the probability distribution of node ages in the original tree differs strongly from that expected based on the model.

statistical_tests

        Data frame, listing the above statistical test results in a more compact format (one test statistic per row).

LTT_ages   Numeric vector, listing ages (time before present) on which the tree's LTT will be defined.

tree_LTT   Numeric vector of the same length as `LTT_ages`, listing the number of lineages in the tree at every age in `LTT_ages`.

bootstrap_LTT_CI

> Named list containing the elements means, medians, CI50lower, CI50upper, CI95lower and CI95upper. Each of these elements is a numeric vector of length equal to LTT_ages, listing the mean or a specific percentile of LTTs of bootstrap trees at every age in LTT_ages. For example, bootstrap_LTT_CI$CI95lower[10] and bootstrap_LTT_CI$CI95upper[10] define the lower and upper bound, respectively, of the 95% confidence interval of LTTs generated by the model at age LTT_ages[10].

fraction_LTT_in_CI95

> Numeric, fraction of the tree's LTT contained within the equal-tailed 95%-confidence interval of the distribution of LTT values predicted by the model. For example, a value of 0.5 means that at half of the time points between the present-day and the root, the tree's LTT is contained with the 95%-CI of predicted LTTs.

Author(s)

Stilianos Louca

References

S. Louca and M. W. Pennell (2020). Extant timetrees are consistent with a myriad of diversification histories. Nature. 580:502-505.

O. G. Pybus and P. H. Harvey (2000). Testing macro-evolutionary models using incomplete molecular phylogenies. Proceedings of the Royal Society of London. Series B: Biological Sciences. 267:2267-2272.

M. J. Sackin (1972). "Good" and "Bad" Phenograms. Systematic Biology. 21:225-226.

K.T. Shao, R. R. Sokal (1990). Tree Balance. Systematic Biology. 39:266-276.

See Also

[simulate_deterministic_hbd](#),

Examples

```
# generate a tree
tree = castor::generate_tree_hbd_reverse(Ntips  = 50,
                                         lambda = 1,
                                         mu     = 0.5,
                                         rho    = 1)$trees[[1]]
root_age = castor::get_tree_span(tree)$max_distance

# define & simulate a somewhat different BD model
model = simulate_deterministic_hbd(LTT0        = 50,
                                   oldest_age  = root_age,
                                   lambda      = 1.5,
                                   mu          = 0.5,
                                   rho0        = 1)

# compare the tree to the model
```

```
adequacy = model_adequacy_hbd(tree,
                              models      = model,
                              Nbootstraps = 100,
                              Nthreads    = 2)
if(!adequacy$success){
    cat(sprintf("Adequacy test failed: %s\n",adequacy$error))
}else{
    print(adequacy$statistical_tests)
}
```

---

model_adequacy_hbds   *Check if a birth-death-sampling model adequately explains a timetree.*

Description

Given a rooted timetree and a homogenous birth-death-sampling model (e.g., as used in molecular epidemiology), check if the model adequately explains various aspects of the tree, such as the branch length and node age distributions and other test statistics. The function uses bootstrapping to simulate multiple hypothetical trees according to the model and then compares the distribution of those trees to the original tree. This function may be used to quantify the "goodness of fit" of a birth-death-sampling model to a timetree. For background on the HBDS model see the documentation for `generate_tree_hbds`.

Usage

```
model_adequacy_hbds(tree,
                    models,
                    splines_degree    = 1,
                    extrapolate       = FALSE,
                    Nbootstraps       = 1000,
                    max_sim_attempts  = 1000,
                    Nthreads          = 1,
                    max_extant_tips   = NULL,
                    max_model_runtime = NULL)
```

Arguments

tree   A rooted timetree of class "phylo".

models   Either a single HBDS model or a list of HBDS models, specifying the pool of models from which to randomly draw bootstraps. Every model should itself be a named list with some or all of the following elements:

                stem_age: Numeric, the age (time before present) at which the HBDS process started. If NULL, this is automatically set to the input tree's root age.

                - end_age : Numeric, the age (time before present) at which the HBDS process halted. This will typically be 0 (i.e., at the tree's youngest tip), however it may also be negative if the process actually halted after the youngest tip was sampled.

- ages: Numeric vector, specifying discrete ages (times before present) in ascending order, on which all model variables (e.g., λ, μ and ψ) will be specified. Age increases from tips to root; the youngest tip in the input tree has age 0. The age grid must cover stem_age and end_age.
- lambda: Numeric vector of the same length as ages, listing the speciation rate (λ) of the HBDS model at the corresponding ages. Between grid points, the speciation rate is assumed to either be constant (if splines_degree=0), or linearly (if splines_degree=1) or quadratically (if splines_degree=2) or cubically (if splines_degree=3).
- mu: Numeric vector of the same length as ages, listing the extinction rate (μ) of the HBDS model at the corresponding ages. Between grid points, the extinction rate is assumed to either be constant (if splines_degree=0), or linearly (if splines_degree=1) or quadratically (if splines_degree=2) or cubically (if splines_degree=3). Note that in epidemiological models μ usually corresponds to the recovery rate plus the death rate of infected hosts. If mu is not included, it is assumed to be zero.
- psi: Optional numeric vector of the same length as ages, listing the Poissonian sampling rate (μ) of the HBDS model at the corresponding ages. Between grid points, the sampling rate is assumed to either be constant (if splines_degree=0), or linearly (if splines_degree=1) or quadratically (if splines_degree=2) or cubically (if splines_degree=3). If psi is not included, it is assumed to be zero.
- kappa: Optional numeric vector of the same length as ages, listing the retention probability upon sampling (κ) of the HBDS model at the corresponding ages. Between grid points, the retention probability is assumed to either be constant (if splines_degree=0), or linearly (if splines_degree=1) or quadratically (if splines_degree=2) or cubically (if splines_degree=3). Note that since kappa are actual probabilities, they must all be between 0 and 1. If kappa is not included, it is assumed to be zero.
- CSA_ages: Numeric vector listing the ages (time before present) of concentrated sampling attempts, in ascending order. If empty or NULL, no concentrated sampling attempts are included, i.e. all sampling is assumed to be done according to the Poissonian rate ψ.
- CSA_probs: Optional numeric vector, of the same length as CSA_ages, specifying the sampling probabilities for each concentrated sampling attempt listed in CSA_ages. Hence, a lineage extant at age CSA_ages[k] has probability CSA_probs[k] of being sampled. Note that since CSA_probs are actual probabilities, they must all be between 0 and 1. CSA_probs must be provided if and only if CSA_ages is provided.
- CSA_kappas: Optional numeric vector, of the same length as CSA_ages, specifying the retention probability upon sampling for each concentrated sampling attempt listed in CSA_ages. Note that since CSA_kappas are actual probabilities, they must all be between 0 and 1. CSA_kappas must be provided if and only if CSA_ages is provided.

If you are assessing the adequacy of a single model with specific parameters, then models can be a single model. If you want to assess the adequacy of a distribution of models, such as sampled from the posterior distribution during a Bayesian analysis, models should list those posterior models.

splines_degree Integer, one of 0, 1, 2 or 3, specifying the polynomial degree of the model parameters λ, μ, ψ and κ between age-grid points. For example, `splines_degree=0` means piecewise constant, `splines_degree=1` means piecewise linear and so on.

extrapolate Logical, specifying whether to extrapolate the model variables λ, μ, ψ and κ (as constants) beyond the provided age grid all the way to `stem_age` and `end_age` if needed.

Nbootstraps Integer, the number of bootstraps (simulations) to perform for calculating statistical significances. A larger number will increase the accuracy of estimated statistical significances.

max_sim_attempts
        Integer, maximum number of simulation attempts per bootstrap, before giving up. Multiple attempts may be needed if the HBDS model has a high probability of leading to extinction early on.

Nthreads Integer, number of parallel threads to use for bootstrapping. Note that on Windows machines this option is ignored.

max_extant_tips
        Integer, optional maximum number of extant tips per simulation. A simulation is aborted (and that bootstrap iteration skipped) if the number of extant tips exceeds this threshold. Use this to avoid occasional explosions of runtimes, for example due to very large generated trees.

max_model_runtime
        Numeric, optional maximum computation time (in seconds) to allow for each HBDS model simulation (per bootstrap). Use this to avoid occasional explosions of runtimes, for example due to very large generated trees. Aborted simulations will be omitted from the bootstrap statistics. If NULL or <=0, this option is ignored.

Details

In addition to model selection, the adequacy of any chosen model should also be assessed in absolute terms, i.e. not just relative to other competing models (after all, all considered models might be bad). This function essentially determines how probable it is for hypothetical trees generated by a candidate model (or a distribution of candidate models) to resemble the tree at hand, in terms of various test statistics. In particular, the function uses a Kolmogorov-Smirnov test to check whether the probability distributions of edge lengths and node ages in the tree resemble those expected under the provided models. All statistical significances are calculated using bootstrapping, i.e. by simulating trees from the provided models. For every bootstrap, a model is randomly chosen from the provided `models` list.

Note that even if an HBDS model appears to adequately explain a given timetree, this does not mean that the model even approximately resembles the true diversification history (i.e., the true speciation, extinction and sampling rates) that generated the tree (Louca and Pennell 2020). Hence, it is generally more appropriate to say that a given model "congruence class" rather than a specific model explains the tree.

Note that here "age" refers to time before present, i.e. age increases from tips to roots and the youngest tip in the input tree has age 0. In some situations the process that generated the tree (or which is being compared to the tree) might have halted after the last tip was sampled, in which case

end_age should be negative. Similarly, the process may have started prior to the tree's root (e.g., sampled tips coalesce at a later time than when the monitoring started), in which case stem_age should be greater than the root's age.

For convenience, it is possible to specify a model without providing an explicit age grid (i.e., omitting ages); in such a model λ, μ, ψ and κ are assumed to be time-independent, and hence lambda, mu, psi and kappa must be provided as single numerics (or not provided at all).

Value

A named list with the following elements:

success	Logical, indicating whether the model evaluation was successful. If FALSE, then an additional return variable, error, will contain a description of the error; in that case all other return variables may be undefined. Note that success does not say whether the model explains the tree, but rather whether the computation was performed without errors.
Nbootstraps	Integer, the number of bootstraps used.
tree_Ntips	Integer, the number of tips in the original tree.
bootstrap_mean_Ntips	
	Numeric, mean number of tips in the bootstrap trees.
PNtips	Numeric, two-sided statistical significance of the tree's number of tips under the provided null model, i.e. the probability that abs(bootstrap_mean_Ntips-tree_Ntips) would be as large as observed.
tree_Colless	Numeric, Colless imbalance statistic (Shao and Sokal, 1990) of the original tree.
bootstrap_mean_Colless	
	Numeric, mean Colless statistic across all bootstrap trees.
PColless	Numeric, two-sided statistical significance of the tree's Colless statistic under the provided null model, i.e. the probability that abs(bootstrap_mean_Colless-tree_Colless) would be as large as observed.
tree_Sackin	Numeric, Sackin statistic (Sackin, 1972) of the original tree.
bootstrap_mean_Sackin	
	Numeric, median Sackin statistic across all bootstrap trees.
PSackin	Numeric, two-sided statistical significance of the tree's Sackin statistic under the provided null model, i.e. the probability that abs(bootstrap_mean_Sackin-tree_Sackin) would be as large as observed.
tree_edgeKS	Numeric, Kolmogorov-Smirnov (KS) statistic of the original tree's edge lengths, i.e. the estimated maximum difference between the tree's and the model's (estimated) cumulative distribution function of edge lengths.
bootstrap_mean_edgeKS	
	Numeric, mean KS statistic of the bootstrap trees' edge lengths.
PedgeKS	Numeric, the one-sided statistical significance of the tree's edge-length KS statistic, i.e. the probability that the KS statistic of any tree generated by the model would be larger than the original tree's KS statistic. A low value means that the probability distribution of edge lengths in the original tree differs strongly from that expected based on the model.

tree_tipKS   Numeric, Kolmogorov-Smirnov (KS) statistic of the original tree's tip ages (sampling times before present), i.e. the estimated maximum difference between the tree's and the model's (estimated) cumulative distribution function of tip ages.

bootstrap_mean_tipKS
                    Numeric, mean KS statistic of the bootstrap trees' tip ages.

PtipKS   Numeric, the one-sided statistical significance of the tree's tip-age KS statistic, i.e. the probability that the KS statistic of any tree generated by the model would be larger than the original tree's KS statistic. A low value means that the probability distribution of tip ages in the original tree differs strongly from that expected based on the model.

tree_nodeKS   Numeric, Kolmogorov-Smirnov (KS) statistic of the original tree's node ages (divergence times before present), i.e. the estimated maximum difference between the tree's and the model's (estimated) cumulative distribution function of node ages.

bootstrap_mean_nodeKS
                    Numeric, mean KS statistic of the bootstrap trees' node ages.

PnodeKS   Numeric, the one-sided statistical significance of the tree's node-age KS statistic, i.e. the probability that the KS statistic of any tree generated by the model would be larger than the original tree's KS statistic. A low value means that the probability distribution of node ages in the original tree differs strongly from that expected based on the model.

statistical_tests
                    Data frame, listing the above statistical test results in a more compact format (one test statistic per row).

LTT_ages   Numeric vector, listing ages (time before present) on which the tree's LTT will be defined.

tree_LTT   Numeric vector of the same length as LTT_ages, listing the number of lineages in the tree at every age in LTT_ages.

bootstrap_LTT_CI
                    Named list containing the elements means, medians, CI50lower, CI50upper, CI95lower and CI95upper. Each of these elements is a numeric vector of length equal to LTT_ages, listing the mean or a specific percentile of LTTs of bootstrap trees at every age in LTT_ages. For example, bootstrap_LTT_CI$CI95lower[10] and bootstrap_LTT_CI$CI95upper[10] define the lower and upper bound, respectively, of the 95% confidence interval of LTTs generated by the model at age LTT_ages[10].

fraction_LTT_in_CI95
                    Numeric, fraction of the tree's LTT contained within the equal-tailed 95%-confidence interval of the distribution of LTT values predicted by the model. For example, a value of 0.5 means that at half of the time points between the present-day and the root, the tree's LTT is contained with the 95%-CI of predicted LTTs.

Author(s)

Stilianos Louca

References

S. Louca and M. W. Pennell (2020). Extant timetrees are consistent with a myriad of diversification histories. Nature. 580:502-505.

O. G. Pybus and P. H. Harvey (2000). Testing macro-evolutionary models using incomplete molecular phylogenies. Proceedings of the Royal Society of London. Series B: Biological Sciences. 267:2267-2272.

M. J. Sackin (1972). "Good" and "Bad" Phenograms. Systematic Biology. 21:225-226.

K.T. Shao, R. R. Sokal (1990). Tree Balance. Systematic Biology. 39:266-276.

See Also

simulate_deterministic_hbds, generate_tree_hbds

Examples

```
## Not run:
# generate a tree based on a simple HBDS process
max_time = 10
gen = castor::generate_tree_hbds(max_time         = max_time,
                                 lambda           = 1,
                                 mu               = 0.1,
                                 psi              = 0.1,
                                 no_full_extinction = TRUE)
if(!gen$success) stop(sprintf("Could not generate tree: %s",gen$error))
tree     = gen$tree
root_age = castor::get_tree_span(tree)$max_distance

# determine age of the stem, i.e. when the HBDS process started
stem_age = gen$root_time + root_age

# determine age at which the HBDS simulation was halted.
# This might be slightly negative, e.g. if the process
# halted after the last sampled tip
end_age = root_age - (gen$final_time-gen$root_time)

# compare the tree to a slightly different model
model = list(stem_age    = stem_age,
             end_age     = end_age,
             lambda      = 1.2,
             mu          = 0.1,
             psi         = 0.2)
adequacy = model_adequacy_hbds( tree,
                                models = model,
                                Nbootstraps = 100)
if(!adequacy$success){
    cat(sprintf("Adequacy test failed: %s\n",adequacy$error))
}else{
    print(adequacy$statistical_tests)
}

## End(Not run)
```

---

multifurcations_to_bifurcations

Expand multifurcations to bifurcations.

Description

Eliminate multifurcations from a phylogenetic tree, by replacing each multifurcation with multiple bifurcations.

Usage

```
multifurcations_to_bifurcations(tree, dummy_edge_length=0,
                                new_node_basename="node.",
                                new_node_start_index=NULL)
```

Arguments

tree   A tree of class "phylo".

dummy_edge_length

          Non-negative numeric. Length to be used for new (dummy) edges when breaking multifurcations into bifurcations. Typically this will be 0, but can also be a positive number if zero edge lengths are not desired in the returned tree.

new_node_basename

          Character. Name prefix to be used for added nodes (e.g. "node." or "new.node."). Only relevant if the input tree included node labels.

new_node_start_index

          Integer. First index for naming added nodes. Can also be NULL, in which case this is set to Nnodes+1, where Nnodes is the number of nodes in the input tree.

Details

For each multifurcating node (i.e. with more than 2 children), all children but one will be placed on new bifurcating nodes, connected to the original node through one or more dummy edges.

The input tree need not be rooted, however descendance from each node is inferred based on the direction of edges in tree$edge. The input tree may include multifurcations (i.e. nodes with more than 2 children) as well as monofurcations (i.e. nodes with only one child). Monofurcations are kept in the returned tree.

All tips and nodes in the input tree retain their original indices, however the returned tree may include additional nodes and edges. Edge indices may change.

If tree$edge.length is missing, then all edges in the input tree are assumed to have length 1. The returned tree will include edge.length, with all new edges having length equal to dummy_edge_length.

Value

A list with the following elements:

tree	A new tree of class "phylo", containing only bifurcations (and monofurcations, if these existed in the input tree).
old2new_edge	Integer vector of length Nedges, mapping edge indices in the old tree to edge indices in the new tree.
Nnodes_added	Integer. Number of nodes added to the new tree.

Author(s)

Stilianos Louca

See Also

collapse_monofurcations

Examples

```
# generate a random multifurcating tree
Ntips = 1000
tree = generate_random_tree(list(birth_rate_intercept=1), Ntips, Nsplits=5)$tree

# expand multifurcations to bifurcations
new_tree = multifurcations_to_bifurcations(tree)$tree

# print summary of old and new tree
cat(sprintf("Old tree has %d nodes\n",tree$Nnode))
cat(sprintf("New tree has %d nodes\n",new_tree$Nnode))
```

---

pick_random_tips   *Pick random subsets of tips on a tree.*

Description

Given a rooted phylogenetic tree, this function picks random subsets of tips by traversing the tree from root to tips, choosing a random child at each node until reaching a tip. Multiple random independent subsets can be generated if needed.

Usage

```
pick_random_tips( tree,
                  size              = 1,
                  Nsubsets          = 1,
                  with_replacement  = TRUE,
                  drop_dims         = TRUE)
```

Arguments

`tree`	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
`size`	Integer. The size of each random subset of tips.
`Nsubsets`	Integer. Number of independent subsets to pick.
`with_replacement`	
	Logical. If `TRUE`, each tip can be picked multiple times within a subset (i.e. are "replaced" in the urn). If `FALSE`, tips are picked without replacement in each subset. In that case, `size` must not be greater than the number of tips in the tree.
`drop_dims`	Logical, specifying whether to return a vector (instead of a matrix) if `Nsubsets==1`.

Details

If `with_replacement==TRUE`, then each child of a node is equally probable to be traversed and each tip can be included multiple times in a subset. If `with_replacement==FALSE`, then only children with at least one descending tip not included in the subset remain available for traversal; each available child of a node has equal probability to be traversed. In any case, it is always possible for separate subsets to include the same tips.

This random sampling algorithm differs from a uniform sampling of tips at equal probabilities; instead, this algorithm ensures that sister clades have equal probabilities to be picked (if `with_replacement==TRUE` or if `size«Ntips`).

The time required by this function per random subset decreases with the number of subsets requested.

Value

A 2D integer matrix of size Nsubsets x size, with each row containing indices of randomly picked tips (i.e. in 1,..,Ntips) within a specific subset. If `drop_dims==TRUE` and `Nsubsets==1`, then a vector is returned instead of a matrix.

Author(s)

Stilianos Louca

Examples

```
# generate random tree
Ntips = 1000
tree  = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# pick random tip subsets
Nsubsets = 100
size     = 50
subsets = pick_random_tips(tree, size, Nsubsets, with_replacement=FALSE)

# count the number of times each tip was picked in a subset ("popularity")
popularities = table(subsets)
```

```
# plot histogram of tip popularities
hist(popularities,breaks=20,xlab="popularity",ylab="# tips",main="tip popularities")
```

---

read_tree   *Load a tree from a string or file in Newick (parenthetic) format.*

Description

Load a phylogenetic tree from a file or a string, in Newick (parenthetic) format. Any valid Newick format is acceptable. Extended variants including edge labels and edge numbers are also supported.

Usage

```
read_tree( string  = "",
           file    = "",
           edge_order              = "cladewise",
           include_edge_lengths    = TRUE,
           look_for_edge_labels    = FALSE,
           look_for_edge_numbers   = FALSE,
           include_node_labels     = TRUE,
           underscores_as_blanks   = FALSE,
           check_label_uniqueness  = FALSE,
           interpret_quotes        = FALSE,
           trim_white              = TRUE)
```

Arguments

string	A character containing a single tree in Newick format. Can be used alternatively to `file`.
file	Character, a path to an input text file containing a single tree in Newick format. Can be used alternatively to `string`.
edge_order	Character, one of "cladewise" or "pruningwise", specifying the order in which edges should be listed in the returned tree. This does not influence the topology of the tree or the tip/node labeling, it only affects the way edges are numbered internally.

include_edge_lengths

Logical, specifying whether edge lengths (if available) should be included in the returned tree.

look_for_edge_labels

Logical, specifying whether edge labels may be present in the input tree. If edge labels are found, they are included in the returned tree as a character vector `edge.label`. Edge labels are sought inside square brackets, which are not part of the standard Newick format used by some tree creation software (Matsen 2012). If `look_for_edge_labels==FALSE`, square brackets are read verbatim just like any other character.

look_for_edge_numbers

Logical, specifying whether edge numbers (non-negative integers) may be present in the input tree. If edge numbers are found, they are included in the returned tree as an integer vector edge.number. Edge numbers are sought inside curly braces, which are not part of the standard Newick format but used by some tree creation software (Matsen 2012). If look_for_edge_numbers==FALSE, curly braces are read verbatim just like any other character.

include_node_labels

Logical, specifying whether node labels (if available) should be included in the returned tree.

underscores_as_blanks

Logical, specifying whether underscores ("_") in tip and node labels should be replaced by spaces (" "). This is common behavior in other tree parsers. In any case, tip, node and edge labels (if available) are also allowed to contain explicit whitespace (except for newline characters).

check_label_uniqueness

Logical, specifying whether to check if all tip labels are unique.

interpret_quotes

Logical, specifying whether to interpret quotes as delimiters of tip/node/edge labels. If FALSE, then quotes are read verbatim just like any other character.

trim_white   Logical, specifying whether to trim flanking whitespace from tip, node and edge labels.

Details

This function is comparable to (but typically much faster than) the ape function read.tree. The function supports trees with monofurcations and multifurcations, trees with or without tip/node labels, and trees with or without edge lengths. The time complexity is linear in the number of edges in the tree.

Either file or string must be specified, but not both. The tree may be arbitrarily split across multiple lines, but no other non-whitespace text is permitted in string or in the input file. Flanking whitespace (space, tab, newlines) is ignored.

Value

A single rooted phylogenetic tree in "phylo" format.

Author(s)

Stilianos Louca

References

Frederick A. Matsen et al. (2012). A format for phylogenetic placements. PLOS One. 7:e31009

See Also

write_tree

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=100)$tree

# obtain a string representation of the tree in Newick format
Newick_string = write_tree(tree)

# re-parse tree from string
parsed_tree = read_tree(Newick_string)
```

reconstruct_past_diversification
                              Reconstruct past diversification dynamics from a diversity time series.

Description

Given a time series of past diversities (coalescent or not), this function estimates instantaneous birth (speciation) and death (extinction) rates that would lead to the observed diversity time series. The function is based on a deterministic model (or the continuum limit of a stochastic cladogenic model), in which instantaneous birth and death rates lead to a predictable growth of a tree (one new species per birth event). The reconstruction is non-parametric, i.e. does not rely on fitting a parameterized model. The reconstruction is only accurate in the deterministic limit, i.e. for high diversities where the stochastic nature of the cladogenic process diminishes. Of particular importance is the case where the time series is coalescent, i.e. represents the diversity (lineages-through-time) that would be represented in a coalescent tree with extinctions.

Note: This function is included for legacy reasons mainly. In most cases users should instead use the functions `fit_hbd_model_on_grid` and `fit_hbd_model_parametric` to fit birth-death models, or the functions `fit_hbd_pdr_on_grid`, `fit_hbd_pdr_parametric` and `fit_hbd_psr_on_grid` to fit BD model congruence classes (aka. "pulled variables") to a tree.

Usage

```
reconstruct_past_diversification( times,
                                  diversities,
                                  birth_rates_pc           = NULL,
                                  rarefaction              = NULL,
                                  discovery_fractions      = NULL,
                                  discovery_fraction_slopes = NULL,
                                  max_age                  = NULL,
                                  coalescent               = FALSE,
                                  smoothing_span           = 0,
                                  smoothing_order          = 1)
```

Arguments

times   Numeric vector, listing the times at which diversities are given. Values must be
                 in ascending order.

diversities	Numeric vector of the same size as `times`, listing diversities (coalescent or not) at each time point.
birth_rates_pc	Numeric vector of the same size as `times`, listing known or assumed per-capita birth rates (speciation rates). Can also be of size 1, in which case the same per-capita birth rate is assumed throughout. Alternatively if `coalescent==TRUE`, then this vector can also be empty, in which case a constant per-capita birth rate is assumed and estimated from the slope of the coalescent diversities at the last time point. The last alternative is not available when `coalescent==FALSE`.
rarefaction	Numeric between 0 and 1. Optional rarefaction fraction assumed for the diversities at the very end. Set to 1 to assume no rarefaction was performed.

discovery_fractions

Numeric array of size Ntimes, listing the fractions of extant lineages represented in the tree over time. Hence, `discovery_fraction[t]` is the probability that a lineage at time `times[t]` with extant representatives will be represented in the tree. Can be used as an alternative to `rarefaction`, for example if discovery of extant species is non-random or phylogenetically biased. Experimental, so leave this `NULL` if you don't know what it means.

discovery_fraction_slopes

Numeric array of size Ntimes, listing the 1st derivative of `discovery_fractions` (w.r.t. time) over time. If `NULL`, this will be estimated from `discovery_fractions` via basic finite differences if needed. Experimental, so leave this `NULL` if you don't know what it means.

max_age	Numeric. Optional maximum distance from the end time to be considered. If `NULL` or <=0 or `Inf`, all provided time points are considered.
coalescent	Logical, indicating whether the provided diversities are from a coalescent tree (only including clades with extant representatives) or total diversities (extant species at each time point).
smoothing_span	Non-negative integer. Optional sliding window size (number of time points) for smoothening the diversities time series via Savitzky-Golay-filter. If <=2, no smoothing is done. Smoothening the time series can reduce the effects of noise on the reconstructed diversity dynamics.

smoothing_order

Integer between 1 and 4. Polynomial order of the Savitzky-Golay smoothing filter to be applied. Only relevant if `smoothing_span>2`. A value of 1 or 2 is typically recommended.

Details

This function can be used to fit a birth-death model to a coalescent diversity time series $N_c(\tau)$ at various ages τ, also known as "lineages-through-time" curve. The reconstruction of the total diversity $N(\tau)$ is based on the following formulas:

$$E(\tau) = 1 + \frac{\nu(\tau)}{\beta(\tau)},$$

$$N(\tau) = \frac{N_c}{1 - E(\tau)},$$

$$\nu(\tau) = \frac{1}{N_c(\tau)} \frac{dN_c(\tau)}{d\tau}$$

where $E(\tau)$ is the probability that a clade of size 1 at age τ went extinct by the end of the time series and β is the per-capita birth rate. If the per-capita birth rate is not explicitly provided for each time point (see argument `birth_rate_pc`), the function assumes that the per-capita birth rate (speciation rate) is constant at all times. If `birth_rates_pc==NULL` and `coalescent==TRUE`, the constant speciation rate is estimated as

$$\beta = -\frac{\nu(0)}{\rho},$$

where ρ is the fraction of species kept after rarefaction (see argument `rarefaction`).

Assuming a constant speciation rate may or may not result in accurate estimates of past total diversities and other quantities. If a time-varying speciation rate is suspected but not known, additional information on past diversification dynamics may be obtained using modified ("pulled") quantities that partly resemble the classical extinction rate, diversification rate and total diversity. Such quantities are the "pulled diversification rate":

$$\eta(\tau) = \delta(\tau) - \beta(\tau) + \frac{1}{\beta(\tau)} \frac{d\beta}{d\tau},$$

the "pulled extinction rate":

$$\delta_p(\tau) = \delta(\tau) + (\beta_o - \beta(\tau)) - \frac{1}{\beta(\tau)} \frac{d\beta}{d\tau},$$

and the "pulled total diversity":

$$N_p(\tau) = N(\tau) \cdot \frac{\beta_o}{\beta(\tau)},$$

where β_o is the provided or estimated (if not provided) speciation rate at the last time point. The advantage of these quantities is that they can be estimated from the coalescent diversities (lineages-through-time) without any assumptions on how β and δ varied over time. The disadvantage is that they differ from their "non-pulled" quantities ($\beta - \delta$, δ and N), in cases where β varied over time.

Value

A named list with the following elements:

success   Logical, specifying whether the reconstruction was successful. If `FALSE`, the remaining elements may not be defined.

Ntimes   Integer. Number of time points for which reconstruction is returned.

total_diversities

Numeric vector of the same size as `times`, listing the total diversity at each time point (number of extant lineages at each time point). If `coalescent==FALSE`, then these are the same as the `diversities` passed to the function.

coalescent_diversities

Numeric vector of the same size as `times`, listing the coalescent diversities at each time point (number of species with at least one extant descendant at the last time point). If `coalescent==TRUE`, then these are the same as the `diversities` passed to the function.

birth_rates   Numeric vector of the same size as `times`, listing the estimated birth rates (speciation events per time unit).

death_rates   Numeric vector of the same size as `times`, listing the estimated death rates (extinction events per time unit).

Psurvival   Numeric vector of the same size as `times`, listing the estimated fraction of lineages at each time point that eventually survive. `Psurvival[i]` is the probability that a clade of size 1 at time `times[i]` will be extant by the end of the time series. May be `NULL` in some cases.

Pdiscovery   Numeric vector of the same size as `times`, listing the estimated fraction of lineages at each time point that are eventually discovered, provided that they survive. `Pdiscovery[i]` is the probability that a clade of size 1 at time `times[i]` that is extant by the end of the time series, will be discovered. May be `NULL` in some cases.

Prepresentation

Numeric vector of the same size as `times`, listing the estimated fraction of lineages at each time point that eventually survive and are discovered. `Prepresentation[i]` is the probability that a clade of size 1 at time `times[i]` will be extant by the end of the time series and visible in the coalescent tree after rarefaction. Note that Prepresentation = Psurvival * Pdiscovery. May be `NULL` in some cases.

total_births   Numeric, giving the estimated total number of birth events that occurred between times `T-max_age` and `T`, where `T` is the last time point of the time series.

total_deaths   Numeric, giving the estimated total number of death events that occurred between times `T-max_age` and `T`, where `T` is the last time point of the time series.

last_birth_rate_pc

The provided or estimated (if not provided) speciation rate at the last time point. This corresponds to the birth rate divided by the estimated true diversity (prior to rarefaction) at the last time point.

last_death_rate_pc

The estimated extinction rate at the last time point. This corresponds to the death rate divided by the estimated true diversity (prior to rarefaction) at the last time point.

pulled_diversification_rates

Numeric vector of the same size as `times`, listing the estimated pulled diversification rates.

pulled_extinction_rates

Numeric vector of the same size as `times`, listing the estimated pulled extinction rates.

pulled_total_diversities

Numeric vector of the same size as `times`, listing the estimated pulled total diversities.

Author(s)

Stilianos Louca

References

Louca et al (2018). Bacterial diversification through geological time. Nature Ecology & Evolution. 2:1458-1467.

See Also

generate_random_tree, fit_tree_model, count_lineages_through_time, fit_hbd_model_parametric, fit_hbd_model_on_grid

Examples

```
####################################################
# EXAMPLE 1

# Generate a coalescent tree
params = list(birth_rate_intercept  = 0,
              birth_rate_factor     = 1,
              birth_rate_exponent   = 1,
              death_rate_intercept  = 0,
              death_rate_factor     = 0.05,
              death_rate_exponent   = 1.3,
              rarefaction           = 1)
simulation = generate_random_tree(params,max_time_eq=1,coalescent=TRUE)
tree = simulation$tree
time_span = simulation$final_time - simulation$root_time
cat(sprintf("Generated tree has %d tips, spans %g time units\n",length(tree$tip.label),time_span))

# Calculate diversity time series from the tree
counter = count_lineages_through_time(tree, times=seq(0,0.99*time_span,length.out=100))

# print coalescent diversities
print(counter$lineages)

# reconstruct diversification dynamics based on diversity time series
results = reconstruct_past_diversification( counter$times,
                                            counter$lineages,
                                            coalescent      = TRUE,
                                            smoothing_span  = 3,
                                            smoothing_order = 1)

# print reconstructed total diversities
print(results$total_diversities)

# plot coalescent and reconstructed true diversities
matplot(x      = counter$times,
        y      = matrix(c(counter$lineages,results$total_diversities), ncol=2, byrow=FALSE),
          type = "b",
          xlab = "time",
          ylab = "# clades",
          lty  = c(1,2), pch = c(1,0), col = c("red","blue"))
legend( "topleft",
          legend  = c("coalescent (simulated)","true (reconstructed)"),
```

```
              col     = c("red","blue"), lty = c(1,2), pch = c(1,0));



    #######################################################
    # EXAMPLE 2

    # Generate a non-coalescent tree
    params = list(birth_rate_intercept  = 0,
                  birth_rate_factor     = 1,
                  birth_rate_exponent   = 1,
                  death_rate_intercept  = 0,
                  death_rate_factor     = 0.05,
                  death_rate_exponent   = 1.3,
                  rarefaction           = 1)
    simulation = generate_random_tree(params,max_time_eq=1,coalescent=FALSE)
    tree = simulation$tree
    time_span = simulation$final_time - simulation$root_time
    cat(sprintf("Generated tree has %d tips, spans %g time units\n",length(tree$tip.label),time_span))

    # Calculate diversity time series from the tree
    counter = count_lineages_through_time(tree, times=seq(0,0.99*time_span,length.out=100))

    # print true diversities
    print(counter$lineages)

    # reconstruct diversification dynamics based on diversity time series
    results = reconstruct_past_diversification( counter$times,
                                                counter$lineages,
                                                birth_rates_pc  = params$birth_rate_factor,
                                                coalescent      = FALSE,
                                                smoothing_span  = 3,
                                                smoothing_order = 1)

    # print coalescent diversities
    print(results$coalescent_diversities)

    # plot coalescent and reconstructed true diversities
    matplot(x     = counter$times,
          y     = matrix(c(results$coalescent_diversities,counter$lineages), ncol=2, byrow=FALSE),
            type  = "b",
            xlab  = "time",
            ylab  = "# clades",
            lty   = c(1,2), pch = c(1,0), col = c("red","blue"))
    legend( "topleft",
            legend = c("coalescent (reconstructed)","true (simulated)"),
            col     = c("red","blue"), lty = c(1,2), pch = c(1,0));
```

---

reorder_tree_edges   *Reorder tree edges in preorder or postorder.*

Description

Given a rooted tree, this function reorders the rows in `tree$edge` so that they are listed in preorder (root–>tips) or postorder (tips–>root) traversal.

Usage

```
reorder_tree_edges(tree, root_to_tips=TRUE,
                    depth_first_search=TRUE,
                    index_only=FALSE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with
                     no incoming edge.

root_to_tips   Logical, specifying whether to sort edges in preorder traversal (root–>tips),
                     rather than in postorder traversal (tips–>roots).

depth_first_search
                     Logical, specifying whether the traversal (or the reversed traversal, if `root_to_tips`
                     is `FALSE`) should be in depth-first-search format rather than breadth-first-search
                     format.

index_only   Whether the function should only return a vector listing the reordered row in-
                     dices of the edge matrix, rather than a modified tree.

Details

This function does not change the tree structure, nor does it affect tip/node indices and names. It
merely changes the order in which edges are listed in the matrix `tree$edge`, so that edges are listed
in preorder or postorder traversal. Preorder traversal guarantees that each edge is listed before any
of its descending edges. Likewise, postorder guarantees that each edge is listed after any of its
descending edges.

With options `root_to_tips=TRUE` and `depth_first_search=TRUE`, this function is analogous to
the function `reorder` in the ape package with option `order="cladewise"`.

The tree can include multifurcations (nodes with more than 2 children) as well as monofurcations
(nodes with 1 child). This function has asymptotic time complexity O(Nedges).

Value

If `index_only==FALSE`, a tree object of class "phylo", with the rows in edge reordered such that
they are listed in direction root–>tips (if `root_to_tips==TRUE`) or tips–>root. The vector `tree$edge.length`
will also be updated in correspondence. Tip and node indices and names remain unchanged.

If `index_only=TRUE`, an integer vector (X) of size Nedges, listing the reordered row indices of
`tree$edge`, i.e. such that `tree$edge[X,]` would be the reordered edge matrix.

Author(s)

Stilianos Louca

See Also

get_tree_traversal_root_to_tips

Examples

```
## Not run:
# generate a random tree
tree = generate_random_tree(list(birth_rate_factor=1), max_tips=100)$tree

# get new tree with reordered edges
postorder_tree = reorder_tree_edges(tree, root_to_tips=FALSE)

## End(Not run)
```

---

root_at_midpoint   *Root or re-root a tree at the midpoint node.*

Description

Given a tree (rooted or unrooted), this function changes the direction of edges (tree$edge) such that the midpoint node becomes the new root (i.e. has no incoming edges and all other tips and nodes descend from it). The number of tips and the number of nodes remain unchanged. The midpoint node is the node whose maximum distance to any tip is smallest.

Usage

```
root_at_midpoint( tree,
                  update_indices  = TRUE,
                  as_edge_counts  = FALSE,
                  is_rooted       = FALSE)
```

Arguments

tree   A tree object of class "phylo". Can be unrooted or rooted (but see option
                is_rooted).

update_indices Logical, specifying whether to update the node indices such that the new root is
                the first node in the list (as is common convention). This will modify tree$node.label
                (if it exists) and also the node indices listed in tree$edge.

as_edge_counts Logical, specifying whether phylogenetic distances should be measured as cu-
                mulative edge counts. This is the same if all edges had length 1.

is_rooted   Logical, specifying whether the input tree can be assumed to be rooted. If you
                are not certain that the tree is rooted, set this to FALSE.

Details

The input tree may include an arbitrary number of incoming and outgoing edges per node (but only one edge per tip), and the direction of these edges can be arbitrary. Of course, the undirected graph defined by all edges must still be a valid tree. Only set is_rooted=TRUE if you are sure that the input tree is rooted.

If update_indices==FALSE, then node indices remain unchanged. If update_indices==TRUE (default), then node indices are modified such that the new root is the first node (i.e. with index Ntips+1 in edge and with index 1 in node.label), as is common convention. Setting update_indices=FALSE reduces the computation required for rerooting. Tip indices always remain unchanged.

The asymptotic time complexity of this function is O(Nedges).

Value

A tree object of class "phylo", with the edge element modified such that the maximum distance of the root to any tip is minimized. The elements tip.label, edge.length and root.edge (if they exist) are the same as for the input tree. If update_indices==FALSE, then the element node.label will also remain the same.

Author(s)

Stilianos Louca

See Also

[root_via_outgroup](), [root_at_node](), [root_in_edge]()

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# reroot the tree at its midpoint node
tree = root_at_midpoint(tree)
```

---

root_at_node   *Root or re-root a tree at a specific node.*

Description

Given a tree (rooted or unrooted) and a specific node, this function changes the direction of edges (tree$edge) such that the designated node becomes the root (i.e. has no incoming edges and all other tips and nodes descend from it). The number of tips and the number of nodes remain unchanged.

Usage

```
root_at_node(tree, new_root_node, update_indices=TRUE)
```

Arguments

`tree`	A tree object of class "phylo". Can be unrooted or rooted.
`new_root_node`	Character or integer specifying the name or index, respectively, of the node to be turned into root. If an integer, it must be between 1 and `tree$Nnode`. If a character, it must be a valid entry in `tree$node.label`.
`update_indices`	Logical, specifying whether to update the node indices such that the new root is the first node in the list (as is common convention). This will modify `tree$node.label` (if it exists) and also the node indices listed in `tree$edge`.

Details

The input tree may include an arbitrary number of incoming and outgoing edges per node (but only one edge per tip), and the direction of these edges can be arbitrary. Of course, the undirected graph defined by all edges must still be a valid tree. The asymptotic time complexity of this function is O(Nedges).

If `update_indices==FALSE`, then node indices remain unchanged. If `update_indices==TRUE` (default), then node indices are modified such that the new root is the first node (i.e. with index Ntips+1 in edge and with index 1 in `node.label`). This is common convention, but it may be undesirable if, for example, you are looping through all nodes in the tree and are only temporarily designating them as root. Setting `update_indices=FALSE` also reduces the computation required for rerooting. Tip indices always remain unchanged.

Value

A tree object of class "phylo", with the edge element modified such that the node `new_root_node` is root. The elements `tip.label`, `edge.length` and `root.edge` (if they exist) are the same as for the input tree. If `update_indices==FALSE`, then the element `node.label` will also remain the same.

Author(s)

Stilianos Louca

See Also

[root_via_outgroup](#), [root_at_midpoint](#), [root_in_edge](#)

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# reroot the tree at the 20-th node
new_root_node = 20
tree = root_at_node(tree, new_root_node, update_indices=FALSE)

# find new root index and compare with expectation
cat(sprintf("New root is %d, expected at %d\n",find_root(tree),new_root_node+Ntips))
```

---

root_in_edge   *Root or re-root a tree in the middle of an edge.*

Description

Given a tree (rooted or unrooted), this function places the new root in the middle of a specified edge, effectively adding one more node, one more edge and changing the direction of edges as required.

Usage

```
root_in_edge( tree,
              root_edge,
              new_root_name = NULL,
              collapse_monofurcations = TRUE)
```

Arguments

tree	A tree object of class "phylo". Can be unrooted or rooted.
root_edge	Integer, index of the edge into which the new root is to be placed. Must be between 1 and Nedges.
new_root_name	Character, optional, specifying the node name to use for the new root. Only used if tree$node.label is not NULL.
collapse_monofurcations	
	Logical, specifying whether monofurcations in the rerooted tree (e.g. stemming from the old root) should be collapsed by connecting incoming edges with outgoing edges.

Details

The input tree may include an arbitrary number of incoming and outgoing edges per node (but only one edge per tip), and the direction of these edges can be arbitrary. Of course, the undirected graph defined by all edges must still be a valid tree.

The number of tips in the rerooted tree remains unchanged, the number of nodes is increased by 1. Node indices may be modified. Tip indices always remain unchanged.

The asymptotic time complexity of this function is O(Nedges).

Value

A tree object of class "phylo", representing the (re-)rooted phylogenetic tree. The element tip.label is the same as for the input tree, but all other elements may have changed.

Author(s)

Stilianos Louca

See Also

[root_via_outgroup](#), [root_at_node](#), [root_at_midpoint](#)

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# reroot the tree inside some arbitrary edge
focal_edge = 120
tree = root_in_edge(tree, focal_edge)
```

---

root_via_outgroup   *Root or re-root a tree based on an outgroup tip.*

Description

Given a tree (rooted or unrooted) and a specific tip ("outgroup"), this function changes the direction of edges (`tree$edge`) such that the outgroup's parent node becomes the root. The number of tips and the number of nodes remain unchanged.

Usage

```
root_via_outgroup(tree, outgroup, update_indices=TRUE)
```

Arguments

tree	A tree object of class "phylo". Can be unrooted or rooted.
outgroup	Character or integer specifying the name or index, respectively, of the outgroup tip. If an integer, it must be between 1 and Ntips. If a character, it must be a valid entry in `tree$tip.label`.
update_indices	Logical, specifying whether to update the node indices such that the new root is the first node in the list (as is common convention). This will modify `tree$node.label` (if it exists) and also the node indices listed in `tree$edge`.

Details

The input tree may include an arbitrary number of incoming and outgoing edges per node (but only one edge per tip), and the direction of these edges can be arbitrary. Of course, the undirected graph defined by all edges must still be a valid tree. The asymptotic time complexity of this function is O(Nedges).

If `update_indices==FALSE`, then node indices remain unchanged. If `update_indices==TRUE` (default), then node indices are modified such that the new root is the first node (i.e. with index Ntips+1 in edge and with index 1 in `node.label`). This is common convention, but it may be undesirable in some cases. Setting `update_indices=FALSE` also reduces the computation required for rerooting. Tip indices always remain unchanged.

Value

A tree object of class "phylo", with the edge element modified such that the outgroup tip's parent node is root. The elements `tip.label`, `edge.length` and `root.edge` (if they exist) are the same as for the input tree. If `update_indices==FALSE`, then the element `node.label` will also remain the same.

Author(s)

Stilianos Louca

See Also

[root_at_node](), [root_at_midpoint](), [root_in_edge]()

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=Ntips)$tree

# reroot the tree using the 1st tip as outgroup
outgroup = 1
tree = root_via_outgroup(tree, outgroup, update_indices=FALSE)

# find new root index
cat(sprintf("New root is %d\n",find_root(tree)))
```

---

shift_clade_times   *Shift the time of specific nodes & tips.*

Description

Given a rooted tree, shift the times (distance from root) of specific tips & nodes.

Usage

```
shift_clade_times(  tree,
                    clades_to_shift,
                    time_shifts,
                    shift_descendants      = FALSE,
                    negative_edge_lengths  = "error")
```

Arguments

tree   A rooted tree of class "phylo".

clades_to_shift

Integer or character vector, listing the tips and/or nodes whose time is to be shifted. If an integer vector, values must correspond to indices and must be in the range 1,..,Ntips+Nnodes. If a character vector, values must correspond to tip and/or node labels in the tree; if node labels are listed, the tree must contain node labels (attribute node.label).

time_shifts   Numeric vector of the same length as clades_to_shift, specifying the time shifts to apply to every tip/node listed in clades_to_shift. Values can be negative (shift towards the root) or positive (shift away from the root).

shift_descendants

Logical, specifying whether to shift the entire descending subclade when shifting a node. If FALSE, the descending tips & nodes retain their original time (unless negative edges are created, see option negative_edge_lengths).

negative_edge_lengths

Character, specifying whether and how to fix negative edge lengths resulting from excessive shifting. Must be either "error", "allow" (allow and don't fix negative edge lengths), "move_all_descendants" (move all descendants forward as needed, to make the edge non-negative), "move_all_ancestors" (move all ancestors backward as needed, to make the edge non-negative), "move_child" (only move children to younger ages as needed, traversing the tree root->tips) or "move_parent" (only move parents to older ages as needed, traversing the tree tips->root). Note that "move_child" could result in tips moving, if an ancestral node is shifted too much towards younger ages. Similarly, "move_parent" could result in the root moving towards an older age if some descendant was shifted too far towards the root.

Details

The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). The input tree does not need to be ultrametric, but edge lengths are interpreted as time. If edge lengths are missing from the tree, it is assumed that each edge has length 1.

All tips, nodes and edges are kept and indexed as in the input tree; the only thing that changes are the edgen lengths.

Note that excessive shifting can result in negative edge lengths, which can be corrected in a variety of alternative ways (see option negative_edge_lengths). However, to avoid changing the overall span of the tree (root age and tip times) in an effort to fix negative edge lengths, you should generally not shift a clade beyond the boundaries of the tree (i.e., resulting in a negative time or a time beyond its descending tips).

Value

A list with the following elements:

success   Logical, specifying whether the operation was successful. If FALSE, an additional variable error is returned, briefly specifying the error, but all other return variables may be undefined.

tree   A new rooted tree of class "phylo", representing the tree with shifted clade times.

Author(s)

Stilianos Louca

See Also

[get_all_distances_to_root](), [trim_tree_at_height](), [get_tree_span]()

Examples

```
# generate a random tree, include node names
tree = generate_random_tree(list(birth_rate_intercept=1),
                            max_tips=20,
                            node_basename="node.")$tree

# shift a few nodes backward in time,
# changing as few of the remaining node timings as possible
clades_to_shift = c("node.2",   "node.5",   "node.6")
time_shifts     = c(-0.5,       -0.2,       -0.3)
new_tree        = shift_clade_times(tree,
                                    clades_to_shift,
                                    time_shifts,
                                    shift_descendants=FALSE,
                                    negative_edge_lengths="move_parent")$tree
```

---

simulate_bm_model   *Simulate a Brownian motion model for multivariate trait co-evolution.*

Description

Given a rooted phylogenetic tree and a Brownian motion (BM) model for the co-evolution of one or more continuous (numeric) unbounded traits, simulate random outcomes of the model on all nodes and/or tips of the tree. The function traverses nodes from root to tips and randomly assigns a multivariate state to each node or tip based on its parent's previously assigned state and the specified model parameters. The generated states have joint distributions consistent with the multivariate BM model. Optionally, multiple independent simulations can be performed using the same model.

Usage

```
simulate_bm_model(tree, diffusivity=NULL, sigma=NULL,
                  include_tips=TRUE, include_nodes=TRUE,
                  root_states=NULL, Nsimulations=1, drop_dims=TRUE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with
                no incoming edge.

diffusivity   Either NULL, or a single number, or a 2D quadratic positive definite symmetric matrix of size Ntraits x Ntraits. Diffusivity matrix ($"D"$) of the multivariate Brownian motion model (in units trait^2/edge_length). The convention is that if the root's state is fixed, then the covariance matrix of a node's state at distance L from the root will be $2LD$ (see mathematical details below).

sigma   Either NULL, or a single number, or a 2D matrix of size Ntraits x Ndegrees, where Ndegrees refers to the degrees of freedom of the model. Noise-amplitude coefficients of the multivariate Brownian motion model (in units trait/sqrt(edge_length)). This can be used as an alternative way to specify the Brownian motion model instead of through the diffusivity D. Note that $sigma \cdot \sigma^T = 2D$ (see mathematical details below).

include_tips   Include random states for the tips. If FALSE, no states will be returned for tips.

include_nodes   Include random states for the nodes. If FALSE, no states will be returned for nodes.

root_states   Numeric matrix of size NR x Ntraits (where NR can be arbitrary), specifying the state of the root for each simulation. If NR is smaller than Nsimulations, values in root_states are recycled in rotation. If root_states is NULL or empty, then the root state is set to 0 for all traits in all simulations.

Nsimulations   Number of random independent simulations to perform. For each node and/or tip, there will be Nsimulations random states generated.

drop_dims   Logical, specifying whether singleton dimensions should be dropped from tip_states and node_states, if Nsimulations==1 and/or Ntraits==1. If drop_dims==FALSE, then tip_states and tip_nodes will always be 3D matrices.

Details

The BM model for Ntraits co-evolving traits is defined by the stochastic differential equation

$$dX = \sigma \cdot dW$$

where W is a multidimensional Wiener process with Ndegrees independent components and σ is a matrix of size Ntraits x Ndegrees. Alternatively, the same model can be defined as a Fokker-Planck equation for the probability density ρ:

$$\frac{\partial \rho}{\partial t} = \sum_{i,j} D_{ij} \frac{\partial^2 \rho}{\partial x_i \partial x_j}.$$

The matrix D is referred to as the diffusivity matrix (or diffusion tensor), and $2D = \sigma \cdot \sigma^T$. Either diffusivity ($D$) or sigma ($\sigma$) may be used to specify the BM model, but not both.

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). The asymptotic time complexity of this function is O(Nedges*Nsimulations*Ntraits).

Value

A list with the following elements:

tip_states	Either NULL (if include_tips==FALSE), or a 3D numeric matrix of size Nsimulations x Ntips x Ntraits. The [r,c,i]-th entry of this matrix will be the state of trait i at tip c generated by the r-th simulation. If drop_dims==TRUE and Nsimulations==1 and Ntraits==1, then tip_states will be a vector.
node_states	Either NULL (if include_nodes==FALSE), or a 3D numeric matrix of size Nsimulations x Nnodes x Ntraits. The [r,c,i]-th entry of this matrix will be the state of trait i at node c generated by the r-th simulation. If drop_dims==TRUE and Nsimulations==1 and Ntraits==1, then node_states will be a vector.

Author(s)

Stilianos Louca

See Also

[simulate_ou_model](), [simulate_rou_model](), [simulate_mk_model](), [fit_bm_model]()

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=10000)$tree

# Example 1: Scalar case
# - - - - - - - - - - - - - - -
# simulate scalar continuous trait evolution on the tree
tip_states = simulate_bm_model(tree, diffusivity=1)$tip_states

# plot histogram of simulated tip states
hist(tip_states, breaks=20, xlab="state", main="Trait probability distribution", prob=TRUE)

# Example 2: Multivariate case
# - - - - - - - - - - - - - - - -
# simulate co-evolution of 2 traits with 3 degrees of freedom
Ntraits  = 2
Ndegrees = 3
sigma    = matrix(stats::rnorm(n=Ntraits*Ndegrees, mean=0, sd=1), ncol=Ndegrees)
tip_states = simulate_bm_model(tree, sigma=sigma, drop_dims=TRUE)$tip_states

# generate scatterplot of traits across tips
plot(tip_states[,1],tip_states[,2],xlab="trait 1",ylab="trait 2",cex=0.5)
```

---

simulate_deterministic_hbd

Simulate a deterministic homogenous birth-death model.

Description

Given a homogenous birth-death (HBD) model, i.e., with speciation rate λ, extinction rate μ and sampling fraction ρ, calculate various deterministic features of the model backwards in time, such as the total diversity over time. The speciation and extinction rates may be time-dependent. "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction rates (in the literature this is sometimes referred to simply as "birth-death model"; Morlon et al. 2011). "Deterministic" refers to the fact that all calculated properties are completely determined by the model's parameters (i.e. non-random), as if an infinitely large tree was generated (aka. "continuum limit").

Alternatively to λ, one may provide the pulled diversification rate (PDR; Louca et al. 2018) and the speciation rate at some fixed age, $\lambda(\tau_o)$. Similarly, alternatively to μ, one may provide the ratio of extinction over speciation rate, μ/λ. In either case, the time-profiles of λ, μ, μ/λ or the PDR are specified as piecewise polynomial functions (splines), defined on a discrete grid of ages.

Usage

```
simulate_deterministic_hbd(LTT0,
                           oldest_age,
                           age0            = 0,
                           rho0            = 1,
                           age_grid        = NULL,
                           lambda          = NULL,
                           mu              = NULL,
                           mu_over_lambda  = NULL,
                           PDR             = NULL,
                           lambda0         = NULL,
                           splines_degree  = 1,
                           relative_dt     = 1e-3,
                           allow_unreal    = FALSE)
```

Arguments

LTT0	The assumed number of sampled extant lineages at age0, defining the necessary initial condition for the simulation. If the HBD model is supposed to describe a specific timetree, then LTT0 should correspond to the number of lineages in the tree ("lineages through time") at age age0.
oldest_age	Strictly positive numeric, specifying the oldest time before present ("age") to include in the simulation.
age0	Non-negative numeric, specifying the age at which LTT0, lambda0 and rho are given. Typically this will be 0, i.e., corresponding to the present.
rho0	Numeric between 0 (exclusive) and 1 (inclusive), specifying the sampling fraction of the tree at age0, i.e. the fraction of lineages extant at age0 that are included in the tree (aka. "rarefaction"). Note that if rho0<1, lineages extant at age0 are assumed to have been sampled randomly at equal probabilities. Can also be NULL, in which case rho0=1 is assumed.
age_grid	Numeric vector, listing discrete ages (time before present) on which either λ and μ, or the PDR and μ, are specified. Listed ages must be strictly increasing, and

must cover at least the full considered age interval (from age0 to oldest_age). Can also be NULL or a vector of size 1, in which case the speciation rate, extinction rate and PDR are assumed to be time-independent.

lambda   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing speciation rates (λ, in units 1/time) at the ages listed in age_grid. Speciation rates should be non-negative, and are assumed to vary polynomially between grid points (see argument splines_degree). If NULL, then PDR and lambda0 must be provided.

mu   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing extinction rates (μ, in units 1/time) at the ages listed in age_grid. Extinction rates should be non-negative, and are assumed to vary polynomially between grid points (see argument splines_degree). Either mu or mu_over_lambda must be provided, but not both.

mu_over_lambda Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing the ratio of extinction rates over speciation rates (μ/λ) at the ages listed in age_grid. These ratios should be non-negative, and are assumed to vary polynomially between grid points (see argument splines_degree). Either mu or mu_over_lambda must be provided, but not both.

PDR   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL), listing pulled diversification rates (in units 1/time) at the ages listed in age_grid. PDRs can be negative or positive, and are assumed to vary polynomially between grid points (see argument splines_degree). If NULL, then lambda must be provided.

lambda0   Non-negative numeric, specifying the speciation rate (in units 1/time) at age0. Either lambda0 or lambda must be provided, but not both.

splines_degree Integer, either 0,1,2 or 3, specifying the polynomial degree of the provided lambda, mu and PDR between grid points in age_grid. For example, if splines_degree==1, then the provided lambda, mu and PDR are interpreted as piecewise-linear curves; if splines_degree==2 they are interpreted as quadratic splines; if splines_degree==3 they are interpreted as cubic splines. The splines_degree influences the analytical properties of the curve, e.g. splines_degree==1 guarantees a continuous curve, splines_degree==2 guarantees a continuous curve and continuous derivative, and so on.

relative_dt   Strictly positive numeric (unitless), specifying the maximum relative time step allowed for integration over time. Smaller values increase integration accuracy but increase computation time. Typical values are 0.0001-0.001. The default is usually sufficient.

allow_unreal   Logical, specifying whether HBD models with unrealistic parameters (e.g., negative μ) should be supported. This may be desired for example when examining model congruence classes with negative μ.

Details

This function supports the following alternative parameterizations of HBD models:

- Using the speciation rate λ and extinction rate μ.

- Using the speciation rate λ and the ratio μ/λ.

- Using the pulled diversification rate (PDR), the extinction rate and the speciation rate given at some fixed age0 (i.e. lambda0).

- Using the PDR, the ratio μ/λ and the speciation rate at some fixed age0.

The PDR is defined as $PDR = \lambda - \mu + \lambda^{-1} d\lambda/d\tau$, where τ is age (time before present). To avoid ambiguities, only one of the above parameterizations is accepted at a time. The sampling fraction at age0 (i.e., rho0) should always be provided; setting it to NULL is equivalent to setting it to 1.

Note that in the literature the sampling fraction usually refers to the fraction of lineages extant at present-day that have been sampled (included in the tree); this present-day sampling fraction is then used to parameterize birth-death cladogenic models. The sampling fraction can however be generalized to past times, by defining it as the probability that a lineage extant at any given time is included in the tree. The simulation function presented here allows for specifying this generalized sampling fraction at any age of choice, not just present-day.

The simulated LTT refers to a hypothetical tree sampled at age age_grid[1], i.e. LTT(t) will be the number of lineages extant at age t that survived until age age_grid[1] and have been sampled, given that the fraction of sampled extant lineages at age0 is rho0. Similarly, the returned Pextinct(t) (see below) is the probability that a lineage extant at age t would not survive until age_grid[1]. The same convention is used for the returned variables Pmissing, shadow_diversity, PER, PSR, SER and PND.

Value

A named list with the following elements:

success	Logical, indicating whether the calculation was successful. If FALSE, then the returned list includes an additional 'error' element (character) providing a description of the error; all other return variables may be undefined.
ages	Numerical vector of size NG, listing discrete ages (time before present) on which all returned time-curves are specified. Listed ages will be in ascending order, will cover exactly the range age_grid[1] - oldest_age, may be irregularly spaced, and may be finer than the original provided age_grid. Note that ages[1] corresponds to the latest time point (closer to the tips), while ages[NG] corresponds to the oldest time point (oldest_age).
total_diversity	
	Numerical vector of size NG, listing the predicted (deterministic) total diversity (number of extant species, denoted N) at the ages given in ages[].
shadow_diversity	
	Numerical vector of size NG, listing the predicted (deterministic) "shadow diversity" at the ages given in ages[]. The shadow diversity is defined as $N_s = N \cdot \rho(\tau_o)\lambda(\tau_o)/\lambda$, where τ_o is age0.
Pmissing	Numeric vector of size NG, listing the probability that a lineage, extant at a given age, will be absent from the tree either due to extinction or due to incomplete sampling.
Pextinct	Numeric vector of size NG, listing the probability that a lineage, extant at a given age, will be fully extinct at present. Note that always Pextinct<=Pmissing.

LTT	Numeric vector of size NG, listing the number of lineages represented in the tree at any given age, also known as "lineages-through-time" (LTT) curve. Note that LTT at age0 will be equal to LTT, and that values in LTT will be non-increasing with age.
lambda	Numeric vector of size NG, listing the speciation rate (in units 1/time) at the ages given in ages[].
mu	Numeric vector of size NG, listing the extinction rate (in units 1/time) at the ages given in ages[].
diversification_rate	
	Numeric vector of size NG, listing the net diversification rate ($\lambda - \mu$) at the ages given in ages[].
PDR	Numeric vector of size NG, listing the pulled diversification rate (PDR, in units 1/time) at the ages given in ages[].
PND	Numeric vector of size NG, listing the pulled normalized diversity (PND, in units 1/time) at the ages given in ages[]. The PND is defined as $PND = (N/N(\tau_o)) \cdot \lambda(\tau_o)/\lambda$.
SER	Numeric vector of size NG, listing the "shadow extinction rate" (SER, in units 1/time) at the ages given in ages[]. The SER is defined as $SER = \rho(\tau_o)\lambda(\tau_o) - PDR$.
PER	Numeric vector of size NG, listing the "pulled extinction rate" (PER, in units 1/time) at the ages given in ages[]. The PER is defined as $SER = \lambda(\tau_o) - PDR$ (Louca et al. 2018).
PSR	Numeric vector of size NG, listing the "pulled speciation rate" (PSR, in units 1/time) at the ages given in ages[]. The PSR is defined as $PSR = \lambda \cdot (1 - Pmissing)$.
rholambda0	Non-negative numeric, specifying the product of the sampling fraction and the speciation rate at age0, $\rho \cdot \lambda(\tau_o)$.

Author(s)

Stilianos Louca

References

H. Morlon, T. L. Parsons, J. B. Plotkin (2011). Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences. 108:16327-16332.

S. Louca et al. (2018). Bacterial diversification through geological time. Nature Ecology & Evolution. 2:1458-1467.

See Also

loglikelihood_hbd

Examples

```
# define an HBD model with exponentially decreasing speciation/extinction rates
Ntips     = 1000
beta      = 0.01 # exponential decay rate of lambda over time
oldest_age= 10
age_grid  = seq(from=0,to=oldest_age,by=0.1) # choose a sufficiently fine age grid
lambda    = 1*exp(beta*age_grid) # define lambda on the age grid
mu        = 0.2*lambda # assume similarly shaped but smaller mu

# simulate deterministic HBD model
simulation = simulate_deterministic_hbd(LTT0        = Ntips,
                                        oldest_age = oldest_age,
                                        rho0       = 0.5,
                                        age_grid   = age_grid,
                                        lambda     = lambda,
                                        mu         = mu)

# plot deterministic LTT
plot( x = simulation$ages, y = simulation$LTT, type='l',
      main='dLTT', xlab='age', ylab='lineages')
```

---

simulate_deterministic_hbds

Simulate a deterministic homogenous birth-death-sampling model.

Description

Given a homogenous birth-death-sampling (HBDS) model, i.e., with speciation rate λ, extinction rate μ, continuous (Poissonian) sampling rate ψ and retention probability κ, calculate various deterministic features of the model backwards in time, such as the total population size and the LTT over time. Continuously sampled lineages are kept in the pool of extant lineages at probability κ. The variables λ, μ, ψ and κ may depend on time. In addition, the model can include concentrated sampling attempts at a finite set of discrete time points $t_1, .., t_m$. "Homogenous" refers to the assumption that, at any given moment in time, all lineages exhibit the same speciation/extinction/sampling rates and retention proabbility. Every HBDS model is thus defined based on the values that λ, μ, ψ and κ take over time, as well as the sampling probabilities $\psi_1, .., \psi_m$ and retention probabilities $\kappa_1, .., \kappa_m$ during the concentrated sampling attempts. Special cases of this model are sometimes known as "birth-death-skyline plots" in the literature (Stadler 2013). In epidemiology, these models are often used to describe the phylogenies of viral strains sampled over the course of the epidemic. A "concentrated sampling attempt" is a brief but intensified sampling period that lasted much less than the typical timescales of speciation/extinction. "Deterministic" refers to the fact that all calculated properties are completely determined by the model's parameters (i.e. non-random), as if an infinitely large tree was generated (aka. "continuum limit"). The time-profiles of λ, μ, ψ and κ are specified as piecewise polynomial curves (splines), defined on a discrete grid of ages.

Usage

```
simulate_deterministic_hbds(age_grid                      = NULL,
```

```
lambda                         = NULL,
mu                             = NULL,
psi                            = NULL,
kappa                          = NULL,
splines_degree                 = 1,
CSA_ages                       = NULL,
CSA_probs                      = NULL,
CSA_kappas                     = NULL,
requested_ages                 = NULL,
age0                           = 0,
N0                             = NULL,
LTT0                           = NULL,
ODE_relative_dt                = 0.001,
ODE_relative_dy                = 1e-4)
```

Arguments

age_grid   Numeric vector, listing discrete ages (time before present) on which either λ and
                  μ, or the PDR and μ, are specified. Listed ages must be strictly increasing, and
                  must cover at least the full considered age interval (from age0 to oldest_age).
                  Can also be NULL or a vector of size 1, in which case the speciation rate, extinc-
                  tion rate and PDR are assumed to be time-independent.

lambda   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL),
                  listing speciation rates (λ, in units 1/time) at the ages listed in age_grid. Spe-
                  ciation rates should be non-negative, and are assumed to vary polynomially be-
                  tween grid points (see argument splines_degree).

mu   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL),
                  listing extinction rates (μ, in units 1/time) at the ages listed in age_grid. Ex-
                  tinction rates should be non-negative, and are assumed to vary polynomially
                  between grid points (see argument splines_degree).

psi   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL),
                  listing the continuous (Poissonian) sampling rate at the ages listed in age_grid.
                  Sampling rates should be non-negative, and are assumed to vary polynomially
                  between grid points (see argument splines_degree).

kappa   Numeric vector, of the same size as age_grid (or size 1 if age_grid==NULL),
                  listing the retention probabilities following Poissonian sampling events, at the
                  ages listed in age_grid. The listed values must be true probabilities, i.e. be-
                  tween 0 and 1, and are assumed to vary polynomially between grid points (see
                  argument splines_degree). The retention probability is the probability that a
                  continuously sampled lineage remains in the pool of extant lineages. Note that
                  many epidemiological models assume kappa to be zero.

splines_degree   Integer, either 0,1,2 or 3, specifying the polynomial degree of the provided
                  lambda, mu, psi and kappa between grid points in age_grid. For example,
                  if splines_degree==1, then the provided lambda, mu, psi and kappa are inter-
                  preted as piecewise-linear curves; if splines_degree==2 they are interpreted as
                  quadratic splines; if splines_degree==3 they are interpreted as cubic splines.
                  The splines_degree influences the analytical properties of the curve, e.g. splines_degree==1

guarantees a continuous curve, `splines_degree==2` guarantees a continuous curve and continuous derivative, and so on.

CSA_ages   Optional numeric vector, listing the ages of concentrated sampling attempts, in ascending order. Concentrated sampling is performed in addition to any continuous (Poissonian) sampling specified by `psi`.

CSA_probs   Optional numeric vector of the same size as `CSA_ages`, listing sampling probabilities at each concentrated sampling attempt. Note that in contrast to the sampling rates `psi`, the `CSA_probs` are interpreted as probabilities and must thus be between 0 and 1. `CSA_probs` must be provided if and only if `CSA_ages` is provided.

CSA_kappas   Optional numeric vector of the same size as `CSA_ages`, listing retention probabilities at each concentrated sampling event, i.e. the probability at which a sampled lineage is kept in the pool of extant lineages. Note that the `CSA_kappas` are probabilities and must thus be between 0 and 1. `CSA_kappas` must be provided if and only if `CSA_ages` is provided.

requested_ages Optional numeric vector, listing ages (in ascending order) at which the various model variables are requested. If `NULL`, it will be set to `age_grid`.

age0   Non-negative numeric, specifying the age at which `LTT0` and `pop_size0` are specified. Typically this will be 0, i.e., corresponding to the present.

N0   Positive numeric, specifying the number of extant species (sampled or not) at `age0`. Used to determine the "scaling factor" for the returned population sizes and LTT. Either `pop_size0` or `LTT0` must be provided, but not both.

LTT0   Positive numeric, specifying the number of lineages present in the tree at `age0`. Used to determine the "scaling factor" for the returned population sizes and LTT. Either `pop_size0` or `LTT0` must be provided, but not both.

ODE_relative_dt

                Positive unitless number, specifying the default relative time step for internally used ordinary differential equation solvers. Typical values are 0.01-0.001.

ODE_relative_dy

                Positive unitless number, specifying the relative difference between subsequent simulated and interpolated values, in internally used ODE solvers. Typical values are 1e-2 to 1e-5. A smaller `ODE_relative_dy` increases interpolation accuracy, but also increases memory requirements and adds runtime (scaling with the tree's age span, not with Ntips).

Details

The simulated LTT refers to a hypothetical tree sampled at age 0, i.e. LTT(t) will be the number of lineages extant at age t that survived and were sampled until by the present day. Note that if a concentrated sampling attempt occurs at age τ, then LTT(τ) is the number of lineages in the tree right before the occurrence of the sampling attempt, i.e., in the limit where τ is approached from above.

Note that in this function age always refers to time before present, i.e., present day age is 0, and age increases towards the root.

Value

A named list with the following elements:

success	Logical, indicating whether the calculation was successful. If FALSE, then the returned list includes an additional 'error' element (character) providing a description of the error; all other return variables may be undefined.
ages	Numerical vector of size NG, listing discrete ages (time before present) on which all returned time-curves are specified. Will be equal to requested_ages, if the latter was provided.

total_diversity

Numerical vector of size NG, listing the predicted (deterministic) total diversity (number of extant species, denoted N) at the ages given in ages[].

LTT	Numeric vector of size NG, listing the number of lineages represented in the tree at any given age, also known as "lineages-through-time" (LTT) curve. Note that LTT at age0 will be equal to LTT0 (if the latter was provided).
nLTT	Numeric vector of size NG, listing values of the normalized LTT at ages ages[]. The nLTT is calculated by dividing the LTT by its area-under-the-curve (AUC). The AUC is calculated by integrating the LTT over the time spanned by ages and using the trapezoid rule. Hence, the exact value of the AUC and of the nLTT depends on the span and resolution of ages[]. If you want the AUC to accurately refer to the entire area under the curve (i.e. across the full real axis), you should specify a sufficiently wide and sufficiently fine age grid (e.g., via requested_ages).
Pmissing	Numeric vector of size NG, listing the probability that a lineage, extant at a given age, will not be represented in the tree.
lambda	Numeric vector of size NG, listing the speciation rates at the ages given in ages[].
mu	Numeric vector of size NG, listing the extinctions rates at the ages given in ages[].
psi	Numeric vector of size NG, listing the Poissonian sampling rates at the ages given in ages[].
kappa	Numeric vector of size NG, listing the retention probabilities (for continuously sampled lineages) at the ages given in ages[].
PDR	Numeric vector of size NG, listing the pulled diversification rate (PDR, in units 1/time) at the ages given in ages[].
IPRP	Numeric vector of size NG, listing the age-integrated pulled diversification rate at the ages given in ages[], i.e. $IPDR(t) = \int_0^t PDR(s)ds$.
PSR	Numeric vector of size NG, listing the "pulled speciation rate" (PSR, in units 1/time) at the ages given in ages[]. The PSR is defined as $PSR = \lambda \cdot (1 - Pmissing)$.
PRP	Numeric vector of size NG, listing the "pulled retention probability" (PRP) at the ages given in ages[]. The PRP is defined as $PRP = \kappa \cdot (1 - Pmissing)$.

diversification_rate

Numeric vector of size NG, listing the net diversification rate (in units 1/time) at the ages given in ages[].

branching_density

> Numeric vector of size NG, listing the deterministic branching density (PSR * nLTT, in units nodes/time) at the ages given in `ages[]`.

sampling_density

> Numeric vector of size NG, listing the deterministic sampling density ($\psi \cdot N/AUC$, in units tips/time, where AUC is the area-under-the-curve calculated for the LTT) at the ages given in `ages[]`.

lambda_psi Numeric vector of size NG, listing the product of the speciation rate and Poissonian sampling rate (in units 1/time^2) at the ages given in `ages[]`.

kappa_psi Numeric vector of size NG, listing the product of the continuous sampling rate and the continuous retention probability (in units 1/time) at the ages given in `ages[]`.

Reff Numeric vector of size NG, listing the effective reproduction ratio ($R_e = \lambda/(\mu + \psi(1 - \kappa))$) at the ages given in `ages[]`.

removal_rate Numeric vector of size NG, listing the total removal rate ($\mu + \psi$), also known as "become uninfectious rate", at the ages given in `ages[]`.

sampling_proportion

> Numeric vector of size NG, listing the instantaneous sampling proportion ($\psi/(\mu + \psi)$) at the ages given in `ages[]`.

CSA_pulled_probs

> Numeric vector of size NG, listing the pulled concentrated sampling probabilities, $\tilde{\rho}_k = \rho_k/(1 - E)$.

CSA_psis Numeric vector of size NG, listing the continuous (Poissonian) sampling rates during the concentrated sampling attempts, $\psi(t_1), .., \psi(t_m)$.

CSA_PSRs Numeric vector of size NG, listing the pulled speciation rates during the concentrated sampling attempts.

Author(s)

Stilianos Louca

References

T. Stadler, D. Kuehnert, S. Bonhoeffer, A. J. Drummond (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). PNAS. 110:228-233.

See Also

[generate_tree_hbds](), [fit_hbds_model_parametric](), [simulate_deterministic_hbd]()

Examples

```
# define an HBDS model with exponentially decreasing speciation/extinction rates
# and constant Poissonian sampling rate psi
oldest_age= 10
age_grid = seq(from=0,to=oldest_age,by=0.1) # choose a sufficiently fine age grid
lambda   = 1*exp(0.01*age_grid) # define lambda on the age grid
mu       = 0.2*lambda # assume similarly shaped but smaller mu
```

```
# simulate deterministic HBD model
# scale LTT such that it is 100 at age 1
simulation = simulate_deterministic_hbds(age_grid   = age_grid,
                                         lambda      = lambda,
                                         mu          = mu,
                                         psi         = 0.1,
                                         age0        = 1,
                                         LTT0        = 100)
# plot deterministic LTT
plot( x = simulation$ages, y = simulation$LTT, type='l',
      main='dLTT', xlab='age', ylab='lineages', xlim=c(oldest_age,0))
```

---

simulate_diversification_model

Simulate a deterministic uniform speciation/extinction model.

Description

Simulate a speciation/extinction cladogenic model for diversity over time, in the derministic limit. Speciation (birth) and extinction (death) rates can each be constant or power-law functions of the number of extant species. For example,

$$B = I + F \cdot N^E,$$

where B is the birth rate, I is the intercept, F is the power-law factor, N is the current number of extant species and E is the power-law exponent. Optionally, the model can account for incomplete taxon sampling (rarefaction of tips) and for the effects of collapsing a tree at a non-zero resolution (i.e. clustering closely related tips into a single tip).

Usage

```
simulate_diversification_model( times,
                                parameters          = list(),
                                added_rates_times   = NULL,
                                added_birth_rates_pc = NULL,
                                added_death_rates_pc = NULL,
                                added_periodic      = FALSE,
                                start_time          = NULL,
                                final_time          = NULL,
                                start_diversity     = 1,
                                final_diversity     = NULL,
                                reverse             = FALSE,
                                include_coalescent  = FALSE,
                                include_event_rates = FALSE,
                                include_Nevents     = FALSE,
                                max_runtime         = NULL)
```

Arguments

times	Numeric vector, listing the times for which to calculate diversities, as predicted by the model. Values must be in ascending order.
parameters	A named list specifying the birth-death model parameters, with one or more of the following entries:

- birth_rate_intercept: Non-negative number. The intercept of the Poissonian rate at which new species (tips) are added. In units 1/time.
- birth_rate_factor: Non-negative number. The power-law factor of the Poissonian rate at which new species (tips) are added. In units 1/time.
- birth_rate_exponent: Numeric. The power-law exponent of the Poissonian rate at which new species (tips) are added. Unitless.
- death_rate_intercept: Non-negative number. The intercept of the Poissonian rate at which extant species (tips) go extinct. In units 1/time.
- death_rate_factor: Non-negative number. The power-law factor of the Poissonian rate at which extant species (tips) go extinct. In units 1/time.
- death_rate_exponent: Numeric. The power-law exponent of the Poissonian rate at which extant species (tips) go extinct. Unitless.
- resolution: Non-negative number. Time resolution at which the final tree is assumed to be collapsed. Units are time units. E.g. if this is 10, then all nodes of age 10 or less, are assumed to be collapsed into (represented by) a single tip. This can be used to model OTU trees, obtained after clustering strains by some similarity (=age) threshold. Set to 0 to disable collapsing. If left unspecified, this is set to 0.
- rarefaction: Numeric between 0 and 1, specifying the fraction of tips kept in the final tree after random subsampling. Rarefaction is assumed to occur after collapsing at the specified resolution (if applicable). This can be used to model incomplete taxon sampling. If left unspecified, this is set to 1.

added_rates_times	
	Numeric vector, listing time points (in ascending order) for a custom per-capita birth and/or death rates time series (see added_birth_rates_pc and added_death_rates_pc below). Can also be NULL, in which case the custom time series are ignored.
added_birth_rates_pc	
	Numeric vector of the same size as added_rates_times, listing per-capita birth rates to be added to the power law part. Added rates are interpolated linearly between time points in added_rates_times. Can also be NULL, in which case this option is ignored and birth rates are purely described by the power law.
added_death_rates_pc	
	Numeric vector of the same size as added_rates_times, listing per-capita death rates to be added to the power law part. Added rates are interpolated linearly between time points in added_rates_times. Can also be NULL, in which case this option is ignored and death rates are purely described by the power law.
added_periodic	Logical, indicating whether added_birth_rates_pc and added_death_rates_pc should be extended periodically if needed (i.e. if not defined for the entire simulation time). If FALSE, added birth & death rates are extended with zeros.

start_time　　　　　Numeric. Start time of the tree (<=times[1]). Can also be NULL, in which case it is set to the first value in times.

final_time　　　　　Numeric. Final (ending) time of the tree (>=max(times)). Can also be NULL, in which case it is set to the last value in times.

start_diversity

　　　　　　　　　Numeric. Total diversity at start_time. Only relevant if reverse==FALSE.

final_diversity

　　　　　　　　　Numeric. Total diversity at final_time, i.e. the final diversity of the tree (total extant species at age 0). Only relevant if reverse==TRUE.

reverse　　　　　　Logical. If TRUE, then the tree model is simulated in backward time direction. In that case, final_diversity is interpreted as the known diversity at the last time point, and all diversities at previous time points are calculated based on the model. If FALSE, then the model is simulated in forward-time, with initial diversity given by start_diversity.

include_coalescent

　　　　　　　　　Logical, specifying whether the diversity corresponding to a coalescent tree (i.e. the tree spanning only extant tips) should also be calculated. If coalescent==TRUE and the death rate is non-zero, then the coalescent diversities will generally be lower than the total diversities.

include_event_rates

　　　　　　　　　Logical. If TRUE, then the birth (speciation) and death (extinction) rates (for each time point) are included as returned values. This comes at a moderate computational overhead.

include_Nevents

　　　　　　　　　Logical. If TRUE, then the cumulative birth (speciation) and death (extinction) events (for each time point) are included as returned values. This comes at a moderate computational overhead.

max_runtime　　　　Numeric. Maximum runtime (in seconds) allowed for the simulation. If this time is surpassed, the simulation aborts.

Details

The simulation is deterministic, meaning that diversification is modeled using ordinary differential equations, not as a stochastic process. The simulation essentially computes the deterministic diversity over time, not an actual tree. For stochastic cladogenic simulations yielding a random tree, see [generate_random_tree](#) and [simulate_dsse](#).

In the special case where per-capita birth and death rates are constant (i.e. $I = 0$ and $E = 1$ for birth and death rates), this function uses an explicit analytical solution to the underlying differential equations, and is thus much faster than in the general case.

If rarefaction<1 and resolution>0, collapsing of closely related tips (at the resolution specified) is assumed to take place prior to rarefaction (i.e., subsampling applies to the already collapsed tips).

Value

A named list with the following elements:

success　　　　　　Logical, indicating whether the simulation was successful. If the simulation aborted due to runtime constraints (option max_runtime), success will be FALSE.

total_diversities

     Numeric vector of the same size as times, listing the total diversity (extant at each the time) for each time point in times.

coalescent_diversities

     Numeric vector of the same size as times, listing the coalescent diversity (i.e. as seen in the coalescent tree spanning only extant species) for each time point in times. Only included if include_coalescent==TRUE.

birth_rates   Numeric vector of the same size as times, listing the speciation (birth) rate at each time point. Only included if include_event_rates==TRUE.

death_rates   Numeric vector of the same size as times, listing the extinction (death) rate at each time point. Only included if include_event_rates==TRUE.

Nbirths   Numeric vector of the same size as times, listing the cumulative number of speciation (birth) events up to each time point. Only included if include_Nevents==TRUE.

Ndeaths   Numeric vector of the same size as times, listing the cumulative number of extinction (death) events up to each time point. Only included if include_Nevents==TRUE.

Author(s)

Stilianos Louca

See Also

[generate_random_tree](#), [count_lineages_through_time](#)

Examples

```
# Generate a tree
max_time = 100
parameters = list(birth_rate_intercept  = 10,
                  birth_rate_factor     = 0,
                  birth_rate_exponent   = 0,
                  death_rate_intercept  = 0,
                  death_rate_factor     = 0,
                  death_rate_exponent   = 0,
                  resolution            = 20,
                  rarefaction           = 0.5)
generator = generate_random_tree(parameters,max_time=max_time)
tree = generator$tree
final_total_diversity = length(tree$tip.label)+generator$Nrarefied+generator$Ncollapsed

# Calculate diversity-vs-time curve for the tree
times = seq(from=0,to=0.99*max_time,length.out=50)
tree_diversities = count_lineages_through_time(tree, times=times)$lineages

# simulate diversity curve based on deterministic model
simulation = simulate_diversification_model(times,
                                            parameters,
                                            reverse=TRUE,
                                            final_diversity=final_total_diversity,
                                            include_coalescent=TRUE)
```

```
model_diversities = simulation$coalescent_diversities

# compare diversities in the tree to the simulated ones
plot(tree_diversities,model_diversities,xlab="tree diversities",ylab="simulated diversities")
abline(a=0,b=1,col="#A0A0A0") # show diagonal for reference
```

---

simulate_dsse   *Simulate a Discrete-State Speciation and Extinction (dSSE) model.*

Description

Simulate a random phylogenetic tree in forward time based on a Poissonian speciation/extinction (birth/death) process, with optional Poissonian sampling over time, whereby birth/death/sampling rates are determined by a co-evolving discrete trait. New species are added (born) by splitting of a randomly chosen extant tip. The discrete trait, whose values determine birth/death/sampling rates over time, can evolve in two modes: (A) Anagenetically, i.e. according to a discrete-space continuous-time Markov process along each edge, with fixed transition rates between states, and/or (B) cladogenetically, i.e. according to fixed transition probabilities between states at each speciation event. Poissonian lineage sampling is assumed to lead to a removal of lineages from the pool of extant tips (as is common in epidemiology).

This model class includes the Multiple State Speciation and Extinction (MuSSE) model described by FitzJohn et al. (2009), as well as the Cladogenetic SSE (ClaSSE) model described by Goldberg and Igis (2012). Optionally, the model can be turned into a Hidden State Speciation and Extinction model (Beaulieu and O'meara, 2016), by replacing the simulated tip/node states with "proxy" states, thus hiding the original states actually influencing speciation/extinction rates.

Usage

```
simulate_dsse( Nstates,
               NPstates             = NULL,
               proxy_map            = NULL,
               parameters           = list(),
               start_state          = NULL,
               max_tips             = NULL,
               max_extant_tips      = NULL,
               max_Psampled_tips    = NULL,
               max_time             = NULL,
               max_time_eq          = NULL,
               max_events           = NULL,
               sampling_fractions   = NULL,
               reveal_fractions     = NULL,
               sampling_rates       = NULL,
               coalescent           = TRUE,
               as_generations       = FALSE,
               no_full_extinction   = TRUE,
               tip_basename         = "",
               node_basename        = NULL,
```

```
                        include_event_times    = FALSE,
                        include_rates          = FALSE,
                        include_labels         = TRUE)

    simulate_musse(Nstates,  NPstates = NULL, proxy_map = NULL,
                   parameters = list(), start_state = NULL,
                   max_tips = NULL, max_extant_tips = NULL, max_Psampled_tips = NULL,
                   max_time = NULL, max_time_eq = NULL, max_events = NULL,
               sampling_fractions = NULL, reveal_fractions = NULL, sampling_rates = NULL,
                 coalescent = TRUE, as_generations = FALSE, no_full_extinction = TRUE,
                   tip_basename = "", node_basename = NULL,
               include_event_times = FALSE, include_rates = FALSE, include_labels = TRUE)
```

Arguments

Nstates	Integer, specifying the number of possible discrete states a tip can have, influencing speciation/extinction rates. For example, if Nstates==2 then this corresponds to the common Binary State Speciation and Extinction (BiSSE) model (Maddison et al., 2007). In the case of a HiSSE model, Nstates refers to the total number of diversification rate categories, as described by Beaulieu and O'meara (2016).
NPstates	Integer, optionally specifying a number of "proxy-states" that are observed instead of the underlying speciation/extinction-modulating states. To simulate a HiSSE model, this should be smaller than Nstates. Each state corresponds to a different proxy-state, as defined using the variable proxy_map (see below). For BiSSE/MuSSE with no hidden states, NPstates can be set to either NULL or equal to Nstates, and proxy-states are equivalent to states.
proxy_map	Integer vector of size Nstates and with values in 1,..NPstates, specifying the correspondence between states (i.e. diversification-rate categories) and (observed) proxy-states, in a HiSSE model. Specifically, proxy_map[s] indicates which proxy-state the state s is represented by. Each proxy-state can represent multiple states (i.e. proxies are ambiguous), but each state must be represented by exactly one proxy-state. For non-HiSSE models, set this to NULL. See below for more details.
parameters	A named list specifying the dSSE model parameters, such as the anagenetic and/or cladogenetic transition rates between states and the state-dependent birth/death rates (see details below).
start_state	Integer within 1,..,Nstates, specifying the initial state, i.e. of the first lineage created. If left unspecified, this is chosen randomly and uniformly among all possible states.
max_tips	Integer, maximum number of tips (extant + Poissonian-sampled if coalescent==TRUE, or extant+extinct+Poissonian-sampled if coalescent==FALSE) in the generated tree, shortly before any present-day sampling. If NULL or <=0, the number of tips is not limited, so you should use another stopping criterion such as max_time and/or max_time_eq and/or max_events to stop the simulation.
max_extant_tips	
	Integer, maximum number of extant tips in the generated tree, shortly before to any present-day sampling. If NULL or <=0, this constraint is ignored.

max_Psampled_tips

        Integer, maximum number of Poissonian-sampled tips in the generated tree. If NULL or <=0, this constraint is ignored.

max_time   Numeric, maximum duration of the simulation. If NULL or <=0, this constraint is ignored.

max_time_eq   Numeric, maximum duration of the simulation, counting from the first point at which speciation/extinction equilibrium is reached, i.e. when (birth rate - death rate) changed sign for the first time. If NULL or <0, this constraint is ignored.

max_events   Integer, maximum number of speciation/extinction/transition events before halting the simulation. If NULL, this constraint is ignored.

sampling_fractions

        A single number, or a numeric vector of size NPstates, listing the sampling fractions for extant tips at the end of the simulation (i.e., at "present-day")", depending on proxy-state. sampling_fractions[p] is the probability of including an extant tip in the final tree, if its proxy-state is p. If a single number, all extant tips are sampled with the same probability, i.e. regardless of their proxy-state. If NULL, this is the same as setting sampling_fractions to 1, i.e., all extant tips are sampled at the end of the simulation.

reveal_fractions

        Numeric vector of size NPstates, listing reveal fractions of tip proxy-states, depending on proxy state. reveal_fractions[p] is the probability of knowing a tip's proxy-state, if its proxy state is p. Can also be NULL, in which case all tip proxy states will be known.

sampling_rates   Numeric vector of size NPstates, listing Poissonian sampling rates of lineages over time, depending on proxy state. Hence, sampling_rates[p] is the sampling rate of a lineage if its proxy state is p. Can also be a single numeric, thus applying the same sampling rate to all lineages regardless of proxy state. Can also be NULL, in which case Poissonian sampling is not included.

coalescent   Logical, specifying whether only the coalescent tree (i.e. the tree spanning the sampled tips) should be returned. If coalescent==FALSE and the death rate is non-zero, then the tree may include extinct tips.

as_generations   Logical, specifying whether edge lengths should correspond to generations. If FALSE, then edge lengths correspond to time.

no_full_extinction

        Logical, specifying whether to prevent complete extinction of the tree. Full extinction is prevented by temporarily disabling extinctions and Poissonian samplings whenever the number of extant tips is 1. if no_full_extinction==FALSE and death rates and/or Poissonian sampling rates are non-zero, the tree may go extinct during the simulation; if coalescent==TRUE, then the returned could end up empty, hence the function will return unsuccessfully (i.e. success will be FALSE). By default no_full_extinction is TRUE, however in some special cases it may be desirable to allow full extinctions to ensure that the generated trees are statistically distributed exactly according to the underlying cladogenetic model.

tip_basename   Character. Prefix to be used for tip labels (e.g. "tip."). If empty (""), then tip labels will be integers "1", "2" and so on.

node_basename   Character. Prefix to be used for node labels (e.g. "node."). If NULL, no node labels will be included in the tree.

include_event_times

       Logical. If TRUE, then the times of speciation and extinction events (each in order of occurrence) will also be returned.

include_rates   Logical. If TRUE, then the per-capita birth & death rates of all tips and nodes will also be returned.

include_labels   Logical, specifying whether to include tip-labels and node-labels (if available) as names in the returned state vectors (e.g. tip_states and node_states). In any case, returned states are always listed in the same order as tips and nodes in the tree. Setting this to FALSE may increase computational efficiency for situations where labels are not required.

Details

The function simulate_dsse can be used to simulate a diversification + discrete-trait evolutionary process, in which birth/death (speciation/extinction) and Poissonian sampling rates at each tip are determined by a tip's current "state". Lineages can transition between states anagenetically along each edge (according to fixed Markov transition rates) and/or cladogenetically at each speciation event (according to fixed transition probabilities). In addition to Poissonian sampling through time (commonly included in epidemiological models), extant tips can also be sampled at the end of the simulation (i.e. at "present-day") according to some state-specific sampling_fractions (common in macroevolution).

The function simulate_musse is a simplified variant meant to simulate MuSSE/HiSSE models in the absence of cladogenetic state transitions, and is included mainly for backward-compatibility reasons. The input arguments for simulate_musse are identical to simulate_dsse, with the exception that the parameters argument must include slightly different elements (explained below). Note that the standard MuSSE/HiSSE models published by FitzJohn et al. (2009) and Beaulieu and O'meara (2016) did not include Poissonian sampling through time, i.e. sampling of extant lineages was only done once at present-day.

For simulate_dsse, the argument parameters should be a named list including one or more of the following elements:

- birth_rates: Numeric vector of size Nstates, listing the per-capita birth rate (speciation rate) at each state. Can also be a single number (all states have the same birth rate).

- death_rates: Numeric vector of size Nstates, listing the per-capita death rate (extinction rate) at each state. Can also be a single number (all states have the same death rate).

- transition_matrix_A: 2D numeric matrix of size Nstates x Nstates, listing anagenetic transition rates between states along an edge. Hence, transition_matrix_A[r,c] is the probability rate for transitioning from state r to state c. Non-diagonal entries must be non-negative, diagonal entries must be non-positive, and the sum of each row must be zero.

- transition_matrix_C: 2D numeric matrix of size Nstates x Nstates, listing cladogenetic transition probabilities between states during a speciation event, seperately for each child. Hence, transition_matrix_C[r,c] is the probability that a child will have state c, conditional upon the occurrence of a speciation event, given that the parent had state r, and independently of all other children. Entries must be non-negative, and the sum of each row must be one.

For `simulate_musse`, the argument `parameters` should be a named list including one or more of the following elements:

- `birth_rates`: Same as for `simulate_dsse`.
- `death_rates`: Same as for `simulate_dsse`.
- `transition_matrix`: 2D numeric matrix of size Nstates x Nstates, listing anagenetic transition rates between states. This is equivalent to `transition_matrix_A` in `simulate_dsse`.

Note that this code generates trees in forward time, and halts as soon as one of the enabled halting conditions is met; the halting conditions chosen affects the precise probability distribution from which the generated trees are drawn (Stadler 2011). If at any moment during the simulation the tree only includes a single extant tip, and if `no_full_extinction=TRUE`, the death and sampling rate are temporarily set to zero to prevent the complete extinction of the tree. The tree will be ultrametric if `coalescent==TRUE` (or death rates were zero) and Poissonian sampling was not included.

HiSSE models (Beaulieu and O'meara, 2016) are closely related to BiSSE/MuSSE models, the main difference being the fact that the actual diversification-modulating states are not directly observed. Hence, this function is also able to simulate HiSSE models, with appropriate choice of the input variables Nstates, NPstates and proxy_map. For example, Nstates=4, NPstates=2 and proxy_map=c(1,2,1,2) specifies that states 1 and 3 are represented by proxy-state 1, and states 2 and 4 are represented by proxy-state 2. This is the original case described by Beaulieu and O'meara (2016); in their terminology, there would be 2 "hidden"" states ("0" and "1") and 2 "observed" (proxy) states ("A" and "B"), and the 4 diversification rate categories (Nstates=4) would be called "0A", "1A", "0B" and "1B", respectively. The somewhat different terminology used here allows for easier generalization to an arbitrary number of diversification-modulating states and an arbitrary number of proxy states. For example, if there are 6 diversification modulating states, represented by 3 proxy-states as 1->A, 2->A, 3->B, 4->C, 5->C, 6->C, then one would set Nstates=6, NPstates=3 and proxy_map=c(1,1,2,3,3,3).

The parameter `transition_matrix_C` can be used to define ClaSSE models (Goldberg and Igic, 2012) or BiSSE-ness models (Magnuson-Ford and Otto, 2012), although care must be taken to properly define the transition probabilities. Here, cladogenetic transitions occur at probabilities that are defined conditionally upon a speciation event, whereas in other software they may be defined as probability rates.

Value

A named list with the following elements:

success
    Logical, indicating whether the simulation was successful. If `FALSE`, an additional element `error` (of type character) is included containing an explanation of the error; in that case the value of any of the other elements is undetermined.

tree
    A rooted bifurcating tree of class "phylo", generated according to the specified birth/death model.
    If `coalescent==TRUE` or if all death rates are zero, and only if `as_generations==FALSE` and in the absence of Poissonian sampling, then the tree will be ultrametric. If `as_generations==TRUE` and `coalescent==FALSE`, all edges will have unit length.

root_time
    Numeric, giving the time at which the tree's root was first split during the simulation. Note that if `coalescent==TRUE`, this may be later than the first speciation event during the simulation.

final_time   Numeric, giving the final time at the end of the simulation. If coalescent==TRUE, then this may be greater than the total time span of the tree (since the root of the coalescent tree need not correspond to the first speciation event).

equilibrium_time

Numeric, giving the first time where the sign of (death rate - birth rate) changed from the beginning of the simulation, i.e. when speciation/extinction equilibrium was reached. May be infinite if the simulation stopped before reaching this point.

Nbirths   Integer vector of size Nstates, listing the total number of birth events (speciations) that occurred at each state. The sum of all entries in Nbirths may be lower than the total number of tips in the tree if death rates were non-zero and coalescent==TRUE.

Ndeaths   Integer vector of size Nstates, listing the total number of death events (extinctions) that occurred at each state.

NPsamplings   Integer vector of size Nstates, listing the total number of Poissonian sampling events that occurred at each state.

Ntransitions_A 2D numeric matrix of size Nstates x Nstates, listing the total number of anagenetic transition events that occurred between each pair of states. For example, Ntransitions_A[1,2] is the number of anagenetic transitions (i.e., within a species) that occured from state 1 to state 2.

Ntransitions_C 2D numeric matrix of size Nstates x Nstates, listing the total number of cladogenetic transition events that occurred between each pair of states. For example, Ntransitions_C[1,2] is the number of cladogenetic transitions (i.e., from a parent to a child) that occured from state 1 to state 2 during some speciation event. Note that each speciation event will have caused 2 transitions (one per child), and that the emergence of a child with the same state as the parent is counted as a transition between the same state (diagonal entries in Ntransitions_C).

NnonsampledExtant

Integer, specifying the number of extant tips not sampled at the end, i.e., omitted from the tree.

tip_states   Integer vector of size Ntips and with values in 1,..,Nstates, listing the state of each tip in the tree.

node_states   Integer vector of size Nnodes and with values in 1,..,Nstates, listing the state of each node in the tree.

tip_proxy_states

Integer vector of size Ntips and with values in 1,..,NPstates, listing the proxy state of each tip in the tree. Only included in the case of HiSSE models.

node_proxy_states

Integer vector of size Nnodes and with values in 1,..,NPstates, listing the proxy state of each node in the tree. Only included in the case of HiSSE models.

start_state   Integer, specifying the state of the first lineage (either provided during the function call, or generated randomly).

extant_tips   Integer vector, listing the indices of any extant tips in the tree.

extinct_tips Integer vector, listing the indices of any extinct tips in the tree. Note that if coalescent==TRUE, this vector will be empty.

Psampled_tips	Integer vector, listing the indices of any Poissonian-sampled tips in the tree.
birth_times	Numeric vector, listing the times of speciation events during tree growth, in order of occurrence. Note that if `coalescent==TRUE`, then `speciation_times` may be greater than the phylogenetic distance to the coalescent root. Only returned if `include_event_times==TRUE`.
death_times	Numeric vector, listing the times of extinction events during tree growth, in order of occurrence. Note that if `coalescent==TRUE`, then `speciation_times` may be greater than the phylogenetic distance to the coalescent root. Only returned if `include_event_times==TRUE`.
Psampling_times	
	Numeric vector, listing the times of Poissonian sampling events during tree growth, in order of occurrence. Only returned if `include_event_times==TRUE`.
clade_birth_rates	
	Numeric vector of size Ntips+Nnodes, listing the per-capita birth rate of each tip and node in the tree. Only included if `include_rates==TRUE`.
clade_death_rates	
	Numeric vector of size Ntips+Nnodes, listing the per-capita death rate of each tip and node in the tree. Only included if `include_rates==TRUE`.

Author(s)

Stilianos Louca

References

W. P. Maddison, P. E. Midford, S. P. Otto (2007). Estimating a binary character's effect on speciation and extinction. Systematic Biology. 56:701-710.

R. G. FitzJohn, W. P. Maddison, S. P. Otto (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology. 58:595-611

R. G. FitzJohn (2012). Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution. 3:1084-1092

E. E. Goldberg, B. Igic (2012). Tempo and mode in plant breeding system evolution. Evolution. 66:3701-3709.

K. Magnuson-Ford, S. P. Otto (2012). Linking the investigations of character evolution and species diversification. The American Naturalist. 180:225-245.

J. M. Beaulieu and B. C. O'Meara (2016). Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology. 65:583-601.

T. Stadler (2011). Simulating trees with a fixed number of extant species. Systematic Biology. 60:676-684.

S. Louca and M. W. Pennell (2020). A general and efficient algorithm for the likelihood of diversification and discrete-trait evolutionary models. Systematic Biology. 69:545-556.

See Also

simulate_tdsse, fit_musse

Examples

```
# Simulate a tree under a classical BiSSE model
# I.e., anagenetic transitions between two states, no Poissonian sampling through time.
A = get_random_mk_transition_matrix(Nstates=2, rate_model="ER", max_rate=0.1)
parameters = list(birth_rates     = c(1,1.5),
                  death_rates      = 0.5,
                  transition_matrix_A = A)
simulation = simulate_dsse( Nstates        = 2,
                            parameters     = parameters,
                            max_extant_tips = 1000,
                            include_rates   = TRUE)
tree      = simulation$tree
Ntips     = length(tree$tip.label)

# plot distribution of per-capita birth rates of tips
rates = simulation$clade_birth_rates[1:Ntips]
barplot(table(rates)/length(rates),
        xlab="rate",
        main="Distribution of pc birth rates across tips (BiSSE model)")
```

---

simulate_mk_model   *Simulate an Mk model for discrete trait evolution.*

Description

Given a rooted phylogenetic tree, a fixed-rates continuous-time Markov model for the evolution of a discrete trait ("Mk model", described by a transition matrix) and a probability vector for the root, simulate random outcomes of the model on all nodes and/or tips of the tree. The function traverses nodes from root to tips and randomly assigns a state to each node or tip based on its parent's previously assigned state and the specified transition rates between states. The generated states have joint distributions consistent with the Markov model. Optionally, multiple independent simulations can be performed using the same model.

Usage

```
simulate_mk_model(tree, Q, root_probabilities="stationary",
                  include_tips=TRUE, include_nodes=TRUE,
                  Nsimulations=1, drop_dims=TRUE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

Q   A numeric matrix of size Nstates x Nstates, storing the transition rates between states. In particular, every row must sum up to zero.

root_probabilities

Probabilities of the different states at the root. Either a character vector with
value "stationary" or "flat", or a numeric vector of length Nstates, where Nstates
is the number of possible states of the trait. In the later case, root_probabilities
must be a valid probability vector, i.e. with non-negative values summing up to
1. "stationary" sets the probabilities at the root to the stationary distribution of
Q (see [get_stationary_distribution](#)), while "flat" means that each state is
equally probable at the root.

include_tips   Include random states for the tips. If FALSE, no states will be returned for tips.

include_nodes   Include random states for the nodes. If FALSE, no states will be returned for
                nodes.

Nsimulations   Number of random independent simulations to perform. For each node and/or
                tip, there will be Nsimulations random states generated.

drop_dims   Logical, specifying whether the returned tip_states and node_states (see
                below) should be vectors, if Nsimulations==1. If drop_dims==FALSE, then
                tip_states and tip_nodes will always be 2D matrices.

Details

For this function, the trait's states must be represented by integers within 1,..,Nstates, where Nstates
is the total number of possible states. If the states are originally in some other format (e.g. characters
or factors), you should map them to a set of integers 1,..,Nstates. These integers should correspond
to row & column indices in the transition matrix Q. You can easily map any set of discrete states to
integers using the function [map_to_state_space](#).

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. The tree may
include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e.
nodes with only one child). The time required per simulation decreases with the total number of
requested simulations.

Value

A list with the following elements:

tip_states   Either NULL (if include_tips==FALSE), or a 2D integer matrix of size Nsim-
                ulations x Ntips with values in 1,..,Nstates, where Ntips is the number of tips
                in the tree and Nstates is the number of possible states of the trait. The [r,c]-th
                entry of this matrix will be the state of tip c generated by the r-th simulation. If
                drop_dims==TRUE and Nsimulations==1, then tip_states will be a vector.

node_states   Either NULL (if include_nodes==FALSE), or a 2D integer matrix of size Nsim-
                ulations x Nnodes with values in 1,..,Nstates, where Nnodes is the number of
                nodes in the tree. The [r,c]-th entry of this matrix will be the state of node c
                generated by the r-th simulation. If drop_dims==TRUE and Nsimulations==1,
                then node_states will be a vector.

Author(s)

Stilianos Louca

See Also

exponentiate_matrix, get_stationary_distribution, simulate_bm_model, simulate_ou_model, simulate_rou_model

Examples

```
## Not run:
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=1000)$tree

# simulate discrete trait evolution on the tree (5 states)
Nstates = 5
Q = get_random_mk_transition_matrix(Nstates, rate_model="ARD", max_rate=0.1)
tip_states = simulate_mk_model(tree, Q)$tip_states

# plot histogram of simulated tip states
barplot(table(tip_states)/length(tip_states), xlab="state")

## End(Not run)
```

---

simulate_ou_model	*Simulate an Ornstein-Uhlenbeck model for continuous trait evolution.*

Description

Given a rooted phylogenetic tree and an Ornstein-Uhlenbeck (OU) model for the evolution of a continuous (numeric) trait, simulate random outcomes of the model on all nodes and/or tips of the tree. The function traverses nodes from root to tips and randomly assigns a state to each node or tip based on its parent's previously assigned state and the specified model parameters. The generated states have joint distributions consistent with the OU model. Optionally, multiple independent simulations can be performed using the same model.

Usage

```
simulate_ou_model(tree, stationary_mean, spread, decay_rate,
                  include_tips=TRUE, include_nodes=TRUE,
                  Nsimulations=1, drop_dims=TRUE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.

stationary_mean

                 Numeric. The mean (center) of the stationary distribution of the OU model.

spread   Numeric. The standard deviation of the stationary distribution of the OU model.

decay_rate   Numeric. Exponential decay rate (stabilization rate) of the OU model (in units 1/edge_length_units).

include_tips   Include random states for the tips. If FALSE, no states will be returned for tips.

include_nodes   Include random states for the nodes. If FALSE, no states will be returned for
                    nodes.

Nsimulations   Number of random independent simulations to perform. For each node and/or
                    tip, there will be Nsimulations random states generated.

drop_dims   Logical, specifying whether the returned tip_states and node_states (see
                    below) should be vectors, if Nsimulations==1. If drop_dims==FALSE, then
                    tip_states and tip_nodes will always be 2D matrices.

Details

For each simulation, the state of the root is picked randomly from the stationary distribution of the
OU model, i.e. from a normal distribution with mean = stationary_mean and standard deviation
= spread.

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. The tree may in-
clude multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes
with only one child). The asymptotic time complexity of this function is O(Nedges*Nsimulations),
where Nedges is the number of edges in the tree.

Value

A list with the following elements:

tip_states   Either NULL (if include_tips==FALSE), or a 2D numeric matrix of size Nsim-
                    ulations x Ntips, where Ntips is the number of tips in the tree. The [r,c]-th
                    entry of this matrix will be the state of tip c generated by the r-th simulation. If
                    drop_dims==TRUE and Nsimulations==1, then tip_states will be a vector.

node_states   Either NULL (if include_nodes==FALSE), or a 2D numeric matrix of size Nsim-
                    ulations x Nnodes, where Nnodes is the number of nodes in the tree. The [r,c]-th
                    entry of this matrix will be the state of node c generated by the r-th simulation. If
                    drop_dims==TRUE and Nsimulations==1, then node_states will be a vector.

Author(s)

Stilianos Louca

See Also

[simulate_bm_model](#), [simulate_mk_model](#), [simulate_rou_model](#)

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=10000)$tree

# simulate evolution of a continuous trait
tip_states = simulate_ou_model(tree, stationary_mean=10, spread=1, decay_rate=0.1)$tip_states

# plot histogram of simulated tip states
hist(tip_states, breaks=20, xlab="state", main="Trait probability distribution", prob=TRUE)
```

---

simulate_rou_model *Simulate a reflected Ornstein-Uhlenbeck model for continuous trait evolution.*

Description

Given a rooted phylogenetic tree and a reflected Ornstein-Uhlenbeck (ROU) model for the evolution of a continuous (numeric) trait, simulate random outcomes of the model on all nodes and/or tips of the tree. The ROU process is similar to the Ornstein-Uhlenbeck process (see `simulate_ou_model`), with the difference that the ROU process cannot fall below a certain value (its "reflection point"), which (in this implementation) is also its deterministic equilibrium point (Hu et al. 2015). The function traverses nodes from root to tips and randomly assigns a state to each node or tip based on its parent's previously assigned state and the specified model parameters. The generated states have joint distributions consistent with the ROU model. Optionally, multiple independent simulations can be performed using the same model.

Usage

```
simulate_rou_model(tree, reflection_point, spread, decay_rate,
                    include_tips=TRUE, include_nodes=TRUE,
                    Nsimulations=1, drop_dims=TRUE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
reflection_point	
	Numeric. The reflection point of the ROU model. In castor, this also happens to be the deterministic equilibrium of the ROU process (i.e. if the decay rate were infinite). For example, if a trait can only be positive (but arbitrarily small), then `reflection_point` may be set to 0.
spread	Numeric. The stationary standard deviation of the corresponding unreflected OU process.
decay_rate	Numeric. Exponential decay rate (stabilization rate) of the ROU process (in units 1/edge_length_units).
include_tips	Include random states for the tips. If FALSE, no states will be returned for tips.
include_nodes	Include random states for the nodes. If FALSE, no states will be returned for nodes.
Nsimulations	Number of random independent simulations to perform. For each node and/or tip, there will be Nsimulations random states generated.
drop_dims	Logical, specifying whether the returned tip_states and node_states (see below) should be vectors, if Nsimulations==1. If drop_dims==FALSE, then `tip_states` and `tip_nodes` will always be 2D matrices.

Details

For each simulation, the state of the root is picked randomly from the stationary distribution of the ROU model, i.e. from a one-sided normal distribution with mode = reflection_point and standard deviation = stationary_std.

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. The tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child). The asymptotic time complexity of this function is O(Nedges*Nsimulations), where Nedges is the number of edges in the tree.

Value

A list with the following elements:

tip_states   Either NULL (if include_tips==FALSE), or a 2D numeric matrix of size Nsim-
                ulations x Ntips, where Ntips is the number of tips in the tree. The [r,c]-th
                entry of this matrix will be the state of tip c generated by the r-th simulation. If
                drop_dims==TRUE and Nsimulations==1, then tip_states will be a vector.

node_states   Either NULL (if include_nodes==FALSE), or a 2D numeric matrix of size Nsim-
                ulations x Nnodes, where Nnodes is the number of nodes in the tree. The [r,c]-th
                entry of this matrix will be the state of node c generated by the r-th simulation. If
                drop_dims==TRUE and Nsimulations==1, then node_states will be a vector.

Author(s)

Stilianos Louca

References

Y. Hu, C. Lee, M. H. Lee, J. Song (2015). Parameter estimation for reflected Ornstein-Uhlenbeck processes with discrete observations. Statistical Inference for Stochastic Processes. 18:279-291.

See Also

simulate_ou_model, simulate_bm_model, simulate_mk_model

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=10000)$tree

# simulate evolution of a continuous trait whose value is always >=1
tip_states = simulate_rou_model(tree, reflection_point=1, spread=2, decay_rate=0.1)$tip_states

# plot histogram of simulated tip states
hist(tip_states, breaks=20, xlab="state", main="Trait probability distribution", prob=TRUE)
```

---

simulate_sbm   *Simulate Spherical Brownian Motion on a tree.*

Description

Given a rooted phylogenetic tree and a Spherical Brownian Motion (SBM) model for the evolution of the geographical location of a lineage on a sphere, simulate random outcomes of the model on all nodes and/or tips of the tree. The function traverses nodes from root to tips and randomly assigns a geographical location to each node or tip based on its parent's previously assigned location and the specified model parameters. The generated states have joint distributions consistent with the SBM model (Perrin 1928; Brillinger 2012). This function generalizes the simple SBM model to support time-dependent diffusivities.

Usage

```
simulate_sbm(tree,
             radius,
             diffusivity,
             time_grid      = NULL,
             splines_degree = 1,
             root_latitude  = NULL,
             root_longitude = NULL)
```

Arguments

tree
: A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge. Edge lengths are assumed to represent time intervals or a similarly interpretable phylogenetic distance.

radius
: Strictly positive numeric, specifying the radius of the sphere. For Earth, the mean radius is 6371 km.

diffusivity
: Either a single numeric, or a numeric vector of length equal to that of time_grid. Diffusivity ("D") of the SBM model (in units distance^2/time). If time_grid is NULL, then diffusivity should be a single number specifying the time-independent diffusivity. Otherwise diffusivity specifies the diffusivity at each time point listed in time_grid.

  Under a planar approximation the squared geographical distance of a node from the root will have expectation $4LD$, where L is the node's phylogenetic distance from the root. Note that distance is measured in the same units as the radius (e.g., km if the radius is given in km), and time is measured in the same units as the tree's edge lengths (e.g., Myr if edge lengths are given in Myr).

time_grid
: Numeric vector of the same length as diffusivity and listing times since the root in ascending order, or NULL. This can be used to specify a time-variable diffusivity (see details below). If NULL, the diffusivity is assumed to be constant over time and equal to diffusivity (which should be a single numeric). Time is measured in the same units as edge lengths, with root having time 0.

splines_degree Integer, either 0,1,2 or 3, specifying the polynomial degree of the provided
                diffusivity between grid points in time_grid. For example, if splines_degree==1,
                then the provided diffusivity is interpreted as a piecewise-linear curve; if
                splines_degree==2 it is interpreted as a quadratic spline; if splines_degree==3
                it is interpreted as a cubic spline. The splines_degree influences the analyt-
                ical properties of the curve, e.g. splines_degree==1 guarantees a continu-
                ous curve, splines_degree==2 guarantees a continuous curve and continuous
                derivative, and so on.

root_latitude   The latitude of the tree's root, in decimal degrees, between -90 and 90. If NULL,
                the root latitude is chosen randomly according to the stationary probability dis-
                tribution of the SBM.

root_longitude The longitude of the tree's root, in decimal degrees, between -180 and 180. If
                NULL, the root longitude is chosen randomly according to the stationary prob-
                ability distribution of the SBM.

Details

For short expected transition distances this function uses the approximation formula by Ghosh et al.
(2012). For longer expected transition distances the function uses a truncated approximation of the
series representation of SBM transition densities (Perrin 1928).

The pair time_grid and diffusivity can be used to define a time-dependent diffusivity, with
time counted from the root to the tips (i.e. root has time 0) in the same units as edge lengths.
For example, to define a diffusivity that varies linearly with time, you only need to specify the
diffusivity at two time points (one at 0, and one at the time of the youngest tip), i.e. time_grid and
diffusivity would each have length 2. Note that time_grid should cover the full time range of
the tree; otherwise, diffusivity will be extrapolated as a constant when needed.

If tree$edge.length is missing, each edge in the tree is assumed to have length 1. The tree may
include multifurcations as well as monofurcations.

Value

A list with the following elements:

success   Logical, specifying whether the simulation was successful. If FALSE, then an
                additional return variable error will contain a brief description of the error that
                occurred, and all other return variables may be undefined.

tip_latitudes   Numeric vector of length Ntips, listing simulated decimal latitudes for each tip
                in the tree.

tip_longitudes Numeric vector of length Ntips, listing simulated decimal longitudes for each
                tip in the tree.

node_latitudes Numeric vector of length Nnodes, listing simulated decimal latitudes for each
                internal node in the tree.

node_longitudes

                Numeric vector of length Nnodes, listing simulated decimal longitudes for each
                internal node in the tree.

Author(s)

Stilianos Louca

References

F. Perrin (1928). Etude mathematique du mouvement Brownien de rotation. 45:1-51.

D. R. Brillinger (2012). A particle migrating randomly on a sphere. in Selected Works of David Brillinger. Springer.

A. Ghosh, J. Samuel, S. Sinha (2012). A Gaussian for diffusion on the sphere. Europhysics Letters. 98:30003.

S. Louca (2021). Phylogeographic estimation and simulation of global diffusive dispersal. Systematic Biology. 70:340-359.

See Also

simulate_ou_model, simulate_rou_model, simulate_bm_model, fit_sbm_const

Examples

```
## Not run:
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=100)$tree

# simulate SBM on the tree
simulation = simulate_sbm(tree, radius=6371, diffusivity=1e4,
                          root_latitude=0, root_longitude=0)

# plot latitudes and longitudes of the tips
plot(simulation$tip_latitudes,simulation$tip_longitudes)

## End(Not run)
```

---

simulate_tdsse	*Simulate a time-dependent Discrete-State Speciation and Extinction (tdSSE) model.*

Description

Simulate a random phylogenetic tree in forward time based on a Poissonian speciation/extinction (birth/death) process, whereby birth and death rates are determined by a co-evolving discrete trait. New species are added (born) by splitting of a randomly chosen extant tip. The discrete trait, whose values determine birth/death rates, can evolve in two modes: (A) Anagenetically, i.e. according to a discrete-space continuous-time Markov process along each edge, with fixed or time-dependent transition rates between states, and/or (B) cladogenetically, i.e. according to fixed or time-dependent transition probabilities between states at each speciation event. This model class includes the Multiple State Speciation and Extinction (MuSSE) model described by FitzJohn et al. (2009), as well as the Cladogenetic SSE (ClaSSE) model described by Goldberg and Igis (2012). Optionally, the

model can be turned into a Hidden State Speciation and Extinction model (Beaulieu and O'meara, 2016), by replacing the simulated tip/node states with "proxy" states, thus hiding the original states actually influencing speciation/extinction rates. This function is similar to simulate_dsse, the main difference being that state-specific speciation/extinction rates as well as state transition rates can be time-dependent.

Usage

```
simulate_tdsse(Nstates,
               NPstates            = NULL,
               proxy_map           = NULL,
               time_grid           = NULL,
               parameters          = list(),
               splines_degree      = 1,
               start_state         = NULL,
               max_tips            = NULL,
               max_time            = NULL,
               max_events          = NULL,
               sampling_fractions  = NULL,
               reveal_fractions    = NULL,
               coalescent          = TRUE,
               as_generations      = FALSE,
               no_full_extinction  = TRUE,
               Nsplits             = 2,
               tip_basename        = "",
               node_basename       = NULL,
               include_birth_times = FALSE,
               include_death_times = FALSE,
               include_labels      = TRUE)
```

Arguments

Nstates
: Integer, specifying the number of possible discrete states a tip can have, influencing speciation/extinction rates. For example, if Nstates==2 then this corresponds to the common Binary State Speciation and Extinction (BiSSE) model (Maddison et al., 2007). In the case of a HiSSE model, Nstates refers to the total number of diversification rate categories, as described by Beaulieu and O'meara (2016).

NPstates
: Integer, optionally specifying a number of "proxy-states" that are observed instead of the underlying speciation/extinction-modulating states. To simulate a HiSSE model, this should be smaller than Nstates. Each state corresponds to a different proxy-state, as defined using the variable proxy_map (see below). For BiSSE/MuSSE with no hidden states, NPstates can be set to either NULL or equal to Nstates, and proxy-states are equivalent to states.

proxy_map
: Integer vector of size Nstates and with values in 1,..,NPstates, specifying the correspondence between states (i.e. diversification-rate categories) and (observed) proxy-states, in a HiSSE model. Specifically, proxy_map[s] indicates which proxy-state the state s is represented by. Each proxy-state can represent

multiple states (i.e. proxies are ambiguous), but each state must be represented by exactly one proxy-state. For non-HiSSE models, set this to NULL. See below for more details.

time_grid   Numeric vector listing discrete times in ascending order, used to define the time-dependent rates of the model. The time grid should generally cover the maximum possible simulation time, otherwise it will be polynomially extrapolated (according to splines_degree).

parameters   A named list specifying the time-dependent model parameters, including optional anagenetic and/or cladogenetic transition rates between states, as well as the mandatory state-dependent birth/death rates (see details below).

splines_degree   Integer, either 0, 1, 2 or 3, specifying the polynomial degree of time-dependent model parameters (birth_rates, death_rates, transition_rates) between time-grid points. For example, splines_degree=1 means that rates are to be considered linear between adjacent grid points.

start_state   Integer within 1,..,Nstates, specifying the initial state, i.e. of the first lineage created. If left unspecified, this is chosen randomly and uniformly among all possible states.

max_tips   Maximum number of tips in the generated tree, prior to any subsampling. If coalescent=TRUE, this refers to the number of extant tips, prior to subsampling. Otherwise, it refers to the number of extinct + extant tips, prior to subsampling. If NULL or <=0, the number of tips is not limited, so you should use max_time and/or max_time_eq and/or max_events to stop the simulation.

max_time   Numeric, maximum duration of the simulation. If NULL or <=0, this constraint is ignored.

max_events   Integer, maximum number of speciation/extinction/transition events before halting the simulation. If NULL, this constraint is ignored.

sampling_fractions

A single number, or a numeric vector of size NPstates, listing tip sub-sampling fractions, depending on proxy-state. sampling_fractions[p] is the probability of including a tip in the final tree, if its proxy-state is p. If NULL, all tips (or all extant tips, if coalescent==TRUE) are included in the tree. If a single number, all tips are included with the same probability, i.e. regardless of their proxy-state.

reveal_fractions

Numeric vector of size NPstates, listing reveal fractions of tip proxy-states, depending on proxy state. reveal_fractions[p] is the probability of knowing a tip's proxy-state, if its proxy state is p. Can also be NULL, in which case all tip proxy states will be known.

coalescent   Logical, specifying whether only the coalescent tree (i.e. the tree spanning the extant tips) should be returned. If coalescent==FALSE and the death rate is non-zero, then the tree may include non-extant tips (i.e. tips whose distance from the root is less than the total time of evolution). In that case, the tree will not be ultrametric.

as_generations   Logical, specifying whether edge lengths should correspond to generations. If FALSE, then edge lengths correspond to time.

no_full_extinction

Logical, specifying whether to prevent complete extinction of the tree. Full extinction is prevented by temporarily disabling extinctions whenever the number of extant tips is 1. if `no_full_extinction==FALSE` and death rates are nonzero, the tree may go extinct during the simulation; if `coalescent==TRUE`, then the returned tree would be empty, hence the function will return unsuccessfully (i.e. `success` will be FALSE). By default `no_full_extinction` is TRUE, however in some special cases it may be desirable to allow full extinctions to ensure that the generated trees are statistically distributed exactly according to the underlying cladogenetic model.

Nsplits   Integer greater than 1. Number of child-tips to generate at each diversification event. If set to 2, the generated tree will be bifurcating. If >2, the tree will be multifurcating.

tip_basename   Character. Prefix to be used for tip labels (e.g. "tip."). If empty (""), then tip labels will be integers "1", "2" and so on.

node_basename   Character. Prefix to be used for node labels (e.g. "node."). If NULL, no node labels will be included in the tree.

include_birth_times

Logical. If TRUE, then the times of speciation events (in order of occurrence) will also be returned.

include_death_times

Logical. If TRUE, then the times of extinction events (in order of occurrence) will also be returned.

include_labels   Logical, specifying whether to include tip-labels and node-labels (if available) as names in the returned state vectors (e.g. `tip_states` and `node_states`). In any case, returned states are always listed in the same order as tips and nodes in the tree. Setting this to FALSE may increase computational efficiency for situations where labels are not required.

Details

The function `simulate_tdsse` can be used to simulate a diversification + discrete-trait evolutionary process, in which birth/death (speciation/extinction) rates at each tip are determined by a tip's current "state". Lineages can transition between states anagenetically along each edge (according to some Markov transition rates) and/or cladogenetically at each speciation event (according to some transition probabilities). The speciation and extinction rates, as well as the transition rates, may be specified as time-dependent variables, defined as piecewise polynomial functions (natural splines) on a temporal grid.

In the following, Ngrid refers to the length of the vector `time_grid`. The argument `parameters` should be a named list including one or more of the following elements:

- `birth_rates`: Numeric 2D matrix of size Nstates x Ngrid, listing the per-capita birth rate (speciation rate) at each state and at each time-grid point. Can also be a single number (same birth rate for all states and at all times).

- `death_rates`: Numeric 2D matrix of size Nstates x Ngrid, listing the per-capita death rate (extinction rate) at each state and at each time-grid point. Can also be a single number (same death rate for all states and at all times) or NULL (no deaths).

- transition_matrix_A: Either a 3D numeric array of size Nstates x Nstates x Ngrid, or a 2D numeric matrix of size Nstates x Nstates, listing anagenetic transition rates between states along an edge. If a 3D array, then transition_matrix_A[r,c,t] is the infinitesimal rate for transitioning from state r to state c at time time_grid[t]. If a 2D matrix, transition_matrix_A[r,c] is the time-independent infintesimal rate for transitioning from state r to state c. At each time point (i.e., a fixed t), non-diagonal entries in transition_matrix_A[,,t] must be non-negative, diagonal entries must be non-positive, and the sum of each row must be zero.

- transition_matrix_C: Either a 3D numeric array of size Nstates x Nstates x Ngrid, or a 2D numeric matrix of size Nstates x Nstates, listing cladogenetic transition probabilities between states during a speciation event, seperately for each child. If a 3D array, then transition_matrix_C[r,c,t] is the probability that a child emerging at time time_grid[t] will have state c, conditional upon the occurrence of a speciation event, given that the parent had state r, and independently of all other children. If a 2D matrix, then transition_matrix_C[r,c] is the (time-independent) probability that a child will have state c, conditional upon the occurrence of a speciation event, given that the parent had state r, and independently of all other children. Entries must be non-negative, and for any fixed t the sum of each row in transition_matrix[,,t] must be one.

If max_time==NULL and max_events==NULL, then the returned tree will always contain max_tips tips. If at any moment during the simulation the tree only includes a single extant tip, and if no_full_extinction=TRUE the death rate is temporarily set to zero to prevent the complete extinction of the tree. If max_tips==NULL, then the simulation is ran as long as specified by max_time and/or max_events. If neither max_time, max_tips nor max_events is NULL, then the simulation halts as soon as the time reaches max_time, or the number of tips (extant tips if coalescent is TRUE) reaches max_tips, or the number of speciation/extinction/transition events reaches max_events whichever occurs first. If max_tips!=NULL and Nsplits>2, then the last diversification even may generate fewer than Nsplits children, in order to keep the total number of tips within the specified limit. Note that this code generates trees in forward time, and halts as soon as one of the halting conditions is met; the halting condition chosen affects the precise distribution from which the generated trees are drawn (Stadler 2011).

For additional information on simulating HiSSE models see the related function simulate_dsse.

The parameter transition_matrix_C can be used to define ClaSSE models (Goldberg and Igic, 2012) or BiSSE-ness models (Magnuson-Ford and Otto, 2012), although care must be taken to properly define the transition probabilities. Here, cladogenetic transitions occur at probabilities that are defined conditionally upon a speciation event, whereas in other software they may be defined as probability rates.

Value

A named list with the following elements:

success	Logical, indicating whether the simulation was successful. If FALSE, an additional element error (of type character) is included containing an explanation of the error; in that case the value of any of the other elements is undetermined.
tree	A rooted bifurcating (if Nsplits==2) or multifurcating (if Nsplits>2) tree of class "phylo", generated according to the specified birth/death model.

	If coalescent==TRUE or if all death rates are zero, and only if as_generations==FALSE, then the tree will be ultrametric. If as_generations==TRUE and coalescent==FALSE, all edges will have unit length.
root_time	Numeric, giving the time at which the tree's root was first split during the simulation. Note that if coalescent==TRUE, this may be later than the first speciation event during the simulation.
final_time	Numeric, giving the final time at the end of the simulation. If coalescent==TRUE, then this may be greater than the total time span of the tree (since the root of the coalescent tree need not correspond to the first speciation event).
Nbirths	Numeric vector of size Nstates, listing the total number of birth events (speciations) that occurred at each state. The sum of all entries in Nbirths may be lower than the total number of tips in the tree if death rates were non-zero and coalescent==TRUE, or if Nsplits>2.
Ndeaths	Numeric vector of size Nstates, listing the total number of death events (extinctions) that occurred at each state.
Ntransitions_A	2D numeric matrix of size Nstates x Nstates, listing the total number of anagenetic transition events that occurred between each pair of states. For example, Ntransitions_A[1,2] is the number of anagenetic transitions (i.e., within a species) that occured from state 1 to state 2.
Ntransitions_C	2D numeric matrix of size Nstates x Nstates, listing the total number of cladogenetic transition events that occurred between each pair of states. For example, Ntransitions_C[1,2] is the number of cladogenetic transitions (i.e., from a parent to a child) that occured from state 1 to state 2 during some speciation event. Note that each speciation event will have caused Nsplits transitions, and that the emergence of a child with the same state as the parent is counted as a transition between the same state (diagonal entries in Ntransitions_C).
tip_states	Integer vector of size Ntips and with values in 1,..,Nstates, listing the state of each tip in the tree.
node_states	Integer vector of size Nnodes and with values in 1,..,Nstates, listing the state of each node in the tree.
tip_proxy_states	
	Integer vector of size Ntips and with values in 1,..,NPstates, listing the proxy state of each tip in the tree. Only included in the case of HiSSE models.
node_proxy_states	
	Integer vector of size Nnodes and with values in 1,..,NPstates, listing the proxy state of each node in the tree. Only included in the case of HiSSE models.
start_state	Integer, specifying the state of the first lineage (either provided during the function call, or generated randomly).
birth_times	Numeric vector, listing the times of speciation events during tree growth, in order of occurrence. Note that if coalescent==TRUE, then speciation_times may be greater than the phylogenetic distance to the coalescent root.
death_times	Numeric vector, listing the times of extinction events during tree growth, in order of occurrence. Note that if coalescent==TRUE, then speciation_times may be greater than the phylogenetic distance to the coalescent root.

Author(s)

Stilianos Louca

References

W. P. Maddison, P. E. Midford, S. P. Otto (2007). Estimating a binary character's effect on speciation and extinction. Systematic Biology. 56:701-710.

R. G. FitzJohn, W. P. Maddison, S. P. Otto (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic Biology. 58:595-611

R. G. FitzJohn (2012). Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution. 3:1084-1092

E. E. Goldberg, B. Igic (2012). Tempo and mode in plant breeding system evolution. Evolution. 66:3701-3709.

K. Magnuson-Ford, S. P. Otto (2012). Linking the investigations of character evolution and species diversification. The American Naturalist. 180:225-245.

J. M. Beaulieu and B. C. O'Meara (2016). Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology. 65:583-601.

T. Stadler (2011). Simulating trees with a fixed number of extant species. Systematic Biology. 60:676-684.

See Also

simulate_dsse, simulate_musse, fit_musse

Examples

```
## Not run:
# prepare params for time-dependent BiSSE model
# include time-dependent speciation & extinction rates
# as well as time-dependent anagenetic transition rates
Nstates         = 2
reveal_fractions = c(1,0.5)
rarefaction     = 0.5 # species sampling fraction

time2lambda1 = function(times) rep(1,times=length(times))
time2lambda2 = function(times) rep(2,times=length(times))
time2mu1     = function(times) 0.5 + 2.5*exp(-((times-8)**2)/2)
time2mu2     = function(times) 1 + 2*exp(-((times-12)**2)/2)
time_grid    = seq(from=0, to=100, length.out=1000)

time2Q12     = function(times) 1*exp(0.1*times)
time2Q21     = function(times) 2*exp(-0.1*times)
QA           = array(0, dim=c(Nstates,Nstates,length(time_grid)))
QA[1,2,]     = time2Q12(time_grid)
QA[2,1,]     = time2Q21(time_grid)
QA[1,1,]     = -QA[1,2,]
QA[2,2,]     = -QA[2,1,]

parameters = list()
```

```
parameters$birth_rates = rbind(time2lambda1(time_grid), time2lambda2(time_grid))
parameters$death_rates = rbind(time2mu1(time_grid), time2mu2(time_grid))
parameters$transition_matrix_A = QA

# simulate time-dependent BiSSE model
cat(sprintf("Simulating tMuSSE model..\n"))
sim = castor::simulate_tdsse(Nstates           = Nstates,
                             time_grid         = time_grid,
                             parameters        = parameters,
                             splines_degree    = 1,
                             max_tips          = 10000/rarefaction,
                             sampling_fractions = rarefaction,
                             reveal_fractions  = reveal_fractions,
                             coalescent        = TRUE,
                             no_full_extinction = TRUE)
if(!sim$success){
    cat(sprintf("ERROR: %s\n",sim$error))
}else{
    # print some summary info about the generated tree
    tree        = sim$tree
    Ntips       = length(tree$tip.label)
    root_age    = get_tree_span(tree)$max_distance
    root_time   = sim$final_time - root_age
    tip_states  = sim$tip_states
    Nknown_tips = sum(!is.na(tip_states))
    cat(sprintf("Note: Simulated tree has root_age = %g\n",root_age))
    cat(sprintf("Note: %d tips have known state\n", Nknown_tips));
}

## End(Not run)
```

---

split_tree_at_height   *Split a tree into subtrees at a specific height.*

Description

Given a rooted phylogenetic tree and a specific distance from the root ("height"), split the tree into subtrees at the specific height. This corresponds to drawing the tree in rectangular layout and trimming everything below the specified phylogenetic distance from the root: What is obtained is a set of separated subtrees. The tips of the original tree are spread across those subtrees.

Usage

```
split_tree_at_height(tree, height = 0, by_edge_count = FALSE)
```

Arguments

tree   A rooted tree of class "phylo". The root is assumed to be the unique node with
                no incoming edge.

height	Numeric, specifying the phylogenetic distance from the root at which to split the tree. If <=0, the original tree is returned as the sole subtree.
by_edge_count	Logical. Instead of considering edge lengths, consider edge counts as phylogenetic distance. This is the same as if all edges had length equal to 1.

Details

This function can be used to generate multiple smaller trees from one large tree, with each subtree having a time span equal to or lower than a certain threshold. The input tree may include multifurcations (i.e. nodes with more than 2 children) as well as monofurcations (i.e. nodes with only one child).

Note that while edges are cut exactly at the specified distance from the root, the cut point does not become the root node of the obtained subtree; rather, the first node encountered after the cut will become the subtree's root. The length of the remaining edge segment leading into this node will be used as root.edge in the returned subtree.

Value

A list with the following elements:

Nsubtrees	Integer, the number of subtrees obtained.
subtrees	A list of length Nsubtrees, each element of which is a named list containing the following elements:

- tree: A rooted tree of class "phylo", representing a subtree obtained from the original tree.
- new2old_clade: An integer vector of length NStips+NSnodes (where NStips is the number of tips and NSnodes the number of nodes of the subtree), mapping subtree tip and node indices (i.e., 1,..,NStips+NSnodes) to tip and node indices in the original tree.
- new2old_edge: Integer vector of length NSedges (=number of edges in the subtree), mapping subtree edge indices (i.e., 1,..,NSedges) to edge indices in the original tree.

clade2subtree	Integer vector of length Ntips+Nnodes and containing values from 1 to Nsubtrees, mapping tip and node indices of the original tree to their assigned subtree.

Author(s)

Stilianos Louca

See Also

trim_tree_at_height

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),
                            max_tips=100)$tree
```

```
# split tree halfway towards the root
root_age = get_tree_span(tree)$max_distance
splitting = split_tree_at_height(tree, height=0.5*root_age)

# print number of subtrees obtained
cat(sprintf("Obtained %d subtrees\n",splitting$Nsubtrees))
```

---

tree_distance   *Calculate the distance between two trees.*

Description

Given two rooted phylogenetic trees with identical tips, calculate their difference using a distance metric or pseudometric.

Usage

```
tree_distance(  treeA,
                treeB,
                tipsA2B       = NULL,
                metric        = "RFrooted",
                normalized    = FALSE,
                NLeigenvalues = 10)
```

Arguments

treeA   A rooted tree of class "phylo".

treeB   A rooted tree of class "phylo". Depending on the metric used, this tree may need
                to have the same number of tips as treeA (see details below).

tipsA2B   Optional integer vector of size Ntips, mapping treeA tip indices to treeB tip
                indices (i.e. tipsA2B[a] is the tip index in treeB corresponding to tip index a
                in treeA). The mapping must be one-to-one. If left unspecified, it is determined
                by matching tip labels between the two trees (this assumes that the same tip
                labels are used in both trees). Only relevant if the metric requires tip matching
                (i.e., considers labeled trees).

metric   Character, specifying the distance measure to be used. Currently the Robinson-
                Foulds metric for rooted trees ("RFrooted"), the mean-path-difference ("Mean-
                PathLengthDifference"), "WassersteinNodeAges" and "WassersteinLaplacianSpec-
                trum" are implemented. Note that these distances are not necessarily metrics in
                the strict mathematical sense; in particular, non-identical trees can sometimes
                have a zero distance.

                "RFrooted" counts the number of clusters (sets of tips descending from a node)
                in either of the trees but not shared by both trees (Robinson and Foulds, 1981;
                Day, 1985); this metric does not take into account branch lengths and depends
                on the position of the root.

"MeanPathLengthDifference" is the square root of the mean squared difference of patristic distances (shortest path lengths) between tip pairs, as described by Steel and Penny (1993); this metric takes into account path lengths and does not depend on the position of the root.

"WassersteinNodeAges" calculates the first Wasserstein distance (Ramdas et al. 2017) between the distributions of node ages in the two trees. It depends on the branch lengths and the rooting, but does not depend on tip labeling nor topology (as long as node ages are the same). Hence, this is only a 'pseudometric' in the space of unlabeled trees - any two trees with identical node ages will have distance 0.

"WassersteinLaplacianSpectrum" calculates the first Wasserstein distance between the spectra (sets of eigenvalues) of the modified graph Laplacians (Lewitus and Morlon, 2016). This distance depends on tree topology and branch lengths, but not on tip labeling nor on the rooting. Note that Lewitus and Morlon measured the distance between the Laplacian spectra in a different way than here. Also note that if `NLeigenvalues>0`, only a subset of the eigenvalues may be considered.

normalized   Logical, specifying whether the calculated distance should be normalized to be between 0 and 1. For the Robinson-Foulds distance, the distance will be normalized by dividing it by the total number of nodes in the two trees. For `MeanPathLengthDifference`, normalization is done by dividing each path-length difference by the maximum of the two path-lengths considered. For `WassersteinNodeAges`, normalization is achieved by scaling all node ages relative to the oldest node age in any of the two trees (hence times are converted to relative times). Note that normalized distances may no longer satisfy the triangle inequality required for metrics, i.e. the resulting distance function may not be a metric in the mathematical sense.

NLeigenvalues Integer, number of top eigenvalues (i.e., with largest magnitude) to consider from the Graph-Laplacian's spectrum (e.g., for the metric "WassersteinLaplacianSpectrum"). This option is mostly provided for computational efficiency reasons, because it is cheaper to compute a small subset of eigenvalues rather than the entire spectrum. If <=0, all eigenvalues are considered, which can substantially increase computation time and memory for large trees.

Details

For some metrics ("RFrooted", "MeanPathLengthDifference"), the trees must have the same number of tips and their tips must be matched one-to-one. If the trees differ in theis tips, they must be pruned down to their common set of tips. If tips have different labels in the two trees, but are nevertheless equivalent, the mapping between the two trees must be provided using `tipsA2B`.

The trees may include multi-furcations as well as mono-furcations (i.e. nodes with only one child).

Note that under some Robinson-Foulds variants the trees can be unrooted; in this present implementation trees must be rooted and the placement of the root influences the distance, following the definition by Day (1985).

Value

A single non-negative number, representing the distance between the two trees.

Author(s)

Stilianos Louca

References

Robinson, D. R., Foulds, L. R. (1981). Comparison of phylogenetic trees. Mathematical Biosciences. 53: 131-147.

Day, W. H. E. (1985). Optimal algorithms for comparing trees with labeled leaves. Journal of Classification. 2:7-28.

Steel, M. A., Penny D. (1993). Distributions of tree comparison metrics - Some new results. Systematic Biology. 42:126-141.

Ramdas, A. et al. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. Entropy. 19(2):47.

Lewitus, E. and Morlon, H. (2016). Characterizing and comparing phylogenies from their laplacian spectrum. Systematic Biology. 65:495-507.

See Also

congruent_divergence_times

Examples

```
# generate a random tree
Ntips = 1000
treeA = generate_random_tree(list(birth_rate_intercept=1),
                             max_tips=Ntips)$tree

# create a second tree with slightly different topology
treeB = treeA
shuffled_tips = sample.int(Ntips, size=Ntips/10, replace=FALSE)
treeB$tip.label[shuffled_tips] = treeB$tip.label[sample(shuffled_tips)]

# calculate Robinson-Foulds distance between trees
distance = tree_distance(treeA, treeB, metric="RFrooted")
```

---

tree_from_branching_ages

Generate a random timetree with specific branching ages.

Description

Generate a random timetree based on specific branching ages (time before present), by randomly connecting tips and nodes. The tree's root will have the greatest age provided. The tree thus corresponds to a homogenous birth-death model, i.e. where at any given time point all lineages were equally likely to split or go extinct.

Usage

```
tree_from_branching_ages(   branching_ages,
                            tip_basename       = "",
                            node_basename      = NULL,
                            edge_basename      = NULL)
```

Arguments

branching_ages	Numeric vector of size Nnodes, listing branching ages (time before present) in ascending order. The last entry will be the root age.
tip_basename	Character. Prefix to be used for tip labels (e.g. "tip."). If empty (""), then tip labels will be integers "1", "2" and so on.
node_basename	Character. Prefix to be used for node labels (e.g. "node."). If NULL, no node labels will be included in the tree.
edge_basename	Character. Prefix to be used for edge labels (e.g. "edge."). Edge labels (if included) are stored in the character vector edge.label. If NULL, no edge labels will be included in the tree.

Details

Tips in the generated tree are guaranteed to be connected in random order, i.e. this function can also be used to connect a random set of labeled tips into a tree. Nodes will be indexed in chronological order (i.e. in order of decreasing age). In particular, node 0 will be the root.

Value

A named list with the following elements:

success	Logical, indicating whether the tree was successfully generated. If FALSE, the only other value returned is error.
tree	A rooted, ultrametric bifurcating tree of class "phylo", with the requested branching ages.
error	Character, containing an explanation of the error that occurred. Only included if success==FALSE.

Author(s)

Stilianos Louca

See Also

tree_from_sampling_branching_ages

Examples

```
Nnodes              = 100
branching_intervals = rexp(n=Nnodes, rate=1)
branching_ages      = cumsum(branching_intervals)
tree                = castor::tree_from_branching_ages(branching_ages)$tree
```

---

tree_from_sampling_branching_ages

Generate a random timetree with specific tip/sampling and node/branching ages.

Description

Generate a random bifurcating timetree based on specific sampling (tip) ages and branching (node) ages, by randomly connecting tips and nodes. Age refers to time before present, i.e., measured in reverse chronological direction. The tree's root will have the greatest age provided. The tree thus corresponds to a homogenous birth-death-sampling model, i.e. where at any given time point all lineages were equally likely to split, be sampled or go extinct.

Usage

```
tree_from_sampling_branching_ages(sampling_ages,
                                  branching_ages,
                                  tip_basename  = "",
                                  node_basename = NULL,
                                  edge_basename = NULL)
```

Arguments

sampling_ages	Numeric vector of size Ntips, listing sampling ages (time before present) in ascending order.
branching_ages	Numeric vector of size Nnodes, listing branching ages (time before present) in ascending order. The last entry will be the root age. Note that Nnodes must be equal to Ntips-1.
tip_basename	Character. Prefix to be used for tip labels (e.g. "tip."). If empty (""), then tip labels will be integers "1", "2" and so on.
node_basename	Character. Prefix to be used for node labels (e.g. "node."). If NULL, no node labels will be included in the tree.
edge_basename	Character. Prefix to be used for edge labels (e.g. "edge."). Edge labels (if included) are stored in the character vector edge.label. If NULL, no edge labels will be included in the tree.

Details

Tips and nodes will be indexed in chronological order (i.e. in order of decreasing age). In particular, node 0 will be the root. Note that not all choices of sampling_ages and branching_ages are permissible. Specifically, at any given age T, the number of sampling events with age equal or smaller than T must be greater than the number of branching events with age equal or smaller than T. If this requirement is not satisfied, the function will return with an error.

Value

A named list with the following elements:

success	Logical, indicating whether the tree was successfully generated. If FALSE, the only other value returned is error.
tree	A rooted, ultrametric bifurcating tree of class "phylo", with the requested tip and node ages.
error	Character, containing an explanation of the error that occurred. Only included if success==FALSE.

Author(s)

Stilianos Louca

See Also

tree_from_branching_ages

Examples

```
sampling_ages  = c(0, 0.1, 0.15, 0.25, 0.9, 1.9, 3)
branching_ages = c(0.3, 0.35, 0.4, 1.1, 2.5, 3.5)
tree = tree_from_sampling_branching_ages(sampling_ages, branching_ages)$tree
```

---

tree_imbalance	*Calculate various imbalance statistics for a tree.*

Description

Given a rooted phylogenetic tree, calculate various "imbalance" statistics of the tree, such as Colless' Index or Sackin's Index.

Usage

```
tree_imbalance(tree, type)
```

Arguments

`tree`	A rooted tree of class "phylo".
`type`	Character, specifying the statistic to be calculated. Must be one of "Colless" (Shao 1990), "Colless_normalized" (Colless normalized by the maximum possible value in the case of a bifurcating tree) or "Sackin" (Sackin 1972).

Details

The tree may include multifurcations and monofurcations. Note that the Colless Index is traditionally only defined for bifurcating trees. For non-bifurcating trees this function calculates a generalization of the index, by summing over all children pairs at each node.

Value

Numeric, the requested imbalance statistic of the tree.

Author(s)

Stilianos Louca

References

M. J. Sackin (1972). "Good" and "Bad" Phenograms. Systematic Biology. 21:225-226.

K.T. Shao, R. R. Sokal (1990). Tree Balance. Systematic Biology. 39:266-276.

Examples

```
# generate a random tree
Ntips = 100
tree = generate_random_tree(list(birth_rate_intercept=1),Ntips)$tree

# calculate Colless statistic
colless_index = tree_imbalance(tree, type="Colless")
```

---

trim_tree_at_height   *Trim a rooted tree down to a specific height.*

Description

Given a rooted phylogenetic tree and a maximum allowed distance from the root ("height"), remove tips and nodes and shorten the remaining terminal edges so that the tree's height does not exceed the specified threshold. This corresponds to drawing the tree in rectangular layout and trimming everything beyond a specific phylogenetic distance from the root. Tips or nodes at the end of trimmed edges are kept, and the affected edges are shortened.

Usage

```
trim_tree_at_height(tree, height = Inf, by_edge_count = FALSE)
```

Arguments

tree	A rooted tree of class "phylo". The root is assumed to be the unique node with no incoming edge.
height	Numeric, specifying the phylogenetic distance from the root at which to trim.
by_edge_count	Logical. Instead of considering edge lengths, consider edge counts as phylogenetic distance. This is the same as if all edges had length equal to 1.

Details

The input tree may include multi-furcations (i.e. nodes with more than 2 children) as well as mono-furcations (i.e. nodes with only one child).

Tip labels and uncollapsed node labels of the collapsed tree are inheritted from the original tree. Labels of tips that used to be nodes (i.e. of which all descendants have been removed) will be the node labels from the original tree. If the input tree has no node names, it is advised to first add node names to avoid NA in the resulting tip names.

Value

A list with the following elements:

tree	A new rooted tree of class "phylo", representing the trimmed tree.
Nedges_trimmed	Integer. Number of edges trimmed (shortened).
Nedges_removed	Integer. Number of edges removed.
new2old_clade	Integer vector of length equal to the number of tips+nodes in the trimmed tree, with values in 1,..,Ntips+Nnodes, mapping tip/node indices of the trimmed tree to tip/node indices in the original tree. In particular,
	c(tree$tip.label,tree$node.label)[new2old_clade]
	will be equal to:
	c(trimmed_tree$tip.label,trimmed_tree$node.label).
new2old_edge	Integer vector of length equal to the number of edges in the trimmed tree, with values in 1,..,Nedges, mapping edge indices of the trimmed tree to edge indices in the original tree. In particular, tree$edge.length[new2old_edge] will be equal to trimmed_tree$edge.length (if edge lengths are available).
new_edges_trimmed	
	Integer vector, listing edge indices in the trimmed tree that we originally longer edges and have been trimmed. In other words, these are the edges that "crossed" the trimming height.

Author(s)

Stilianos Louca

See Also

split_tree_at_height

Examples

```
# generate a random tree, include node names
tree = generate_random_tree(list(birth_rate_intercept=1),
                            max_time=1000,
                            node_basename="node.")$tree

# print number of tips
cat(sprintf("Simulated tree has %d tips\n",length(tree$tip.label)))

# trim tree at height 500
trimmed = trim_tree_at_height(tree, height=500)$tree

# print number of tips in trimmed tree
cat(sprintf("Trimmed tree has %d tips\n",length(trimmed$tip.label)))
```

---

write_tree   *Write a tree in Newick (parenthetic) format.*

Description

Write a phylogenetic tree to a file or a string, in Newick (parenthetic) format. If the tree is unrooted, it is first rooted internally at the first node.

Usage

```
write_tree (tree,
            file               = "",
            append             = FALSE,
            digits             = 10,
            quoting            = 0,
            include_edge_labels  = FALSE,
            include_edge_numbers = FALSE)
```

Arguments

tree   A tree of class "phylo".

file   An optional path to a file, to which the tree should be written. The file may be
                 overwritten without warning. If left empty (default), then a string is returned
                 representing the tree.

append   Logical, specifying whether the tree should be appended at the end of the file,
                 rather than replacing the entire file (if it exists).

digits   Integer, number of significant digits for writing edge lengths.

quoting   Integer, specifying whether and how to quote tip/node/edge names, as follows:
                 0:no quoting at all, 1:always use single quotes, 2:always use double quotes, -
                 1:only quote when needed and prefer single quotes if possible, -2:only quote
                 when needed and prefer double quotes if possible.

include_edge_labels

> Logical, specifying whether to include edge labels (if available) in the output tree, inside square brackets. Note that this is an extension (Matsen et al. 2012) to the standard Newick format, as, and edge labels in square brackets may not be supported by all Newick readers.

include_edge_numbers

> Logical, specifying whether to include edge numbers (if available) in the output tree, inside curly braces. Note that this is an extension (Matsen et al. 2012) to the standard Newick format, and edge numbers in curly braces may not be supported by all Newick readers.

Details

If your tip and/or node and/or edge labels contain special characters (round brackets, commas, colons or quotes) then you should set quoting to non-zero, as appropriate.

If the tree contains edge labels (as a character vector named edge.label) and include_edge_labels==TRUE, then edge labels are written in square brackets (Matsen et al. 2012). If tree contains edge numbers (as an integer vector named edge.number) and include_edge_numbers==TRUE, then edge numbers are written in curly braces (Matsen et al. 2012).

This function is comparable to (but typically much faster than) the ape function write.tree.

Value

If file=="", then a string is returned containing the Newick representation of the tree. Otherwise, the tree is directly written to the file and no value is returned.

Author(s)

Stilianos Louca

References

Frederick A. Matsen et al. (2012). A format for phylogenetic placements. PLOS One. 7:e31009

See Also

[read_tree](#)

Examples

```
# generate a random tree
tree = generate_random_tree(list(birth_rate_intercept=1),max_tips=100)$tree

# obtain a string representation of the tree in Newick format
Newick_string = write_tree(tree)
```

Index