Package 'ML.MSBD'

November 14, 2017

Type Package

Title Maximum Likelihood Inference on Multi-State Trees

Version 1.0.0

Date 2017-10-25

Description Inference of a multi-states birth-death model from a phylogeny, comprising a number of states N, birth and death rates for each state and on which edges each state appears. Inference is done using a hybrid approach: states are progressively added in a greedy approach. For a fixed number of states N the best model is selected via maximum likelihood. Reference: J. Barido-Sottani and T. Stadler (2017) <doi:10.1101/215491>.

License GPL-3

Imports ape (>= 4.1)

RoxygenNote 6.0.1

NeedsCompilation no

Author Joelle Barido-Sottani [aut, cre]

Maintainer Joelle Barido-Sottani <joelle.barido-sottani@m4x.org>

Repository CRAN

Date/Publication 2017-11-14 16:54:53 UTC

R topics documented:

ML.MSBD-package *Maximum Likelihood Inference on Multi-State Trees*

Description

Inference of a multi-states birth-death model from a phylogeny, comprising a number of states N, birth and death rates for each state and on which edges each state appears. Inference is done using a hybrid approach: states are progressively added in a greedy approach. For a fixed number of states N the best model is selected via maximum likelihood. Reference: J. Barido-Sottani and T. Stadler (2017) <doi:10.1101/215491>.

Details

Package:	ML.MSBD
Type:	Package
Title:	Maximum Likelihood Inference on Multi-State Trees
Version:	1.0.0
Date:	2017-10-25
Authors@R:	person("Joelle", "Barido-Sottani", email = "joelle.barido-sottani@m4x.org", role = c("aut", "cre"))
Description:	Inference of a multi-states birth-death model from a phylogeny, comprising a number of states N, birth and d
License:	GPL-3
Imports:	ape (>= 4.1)
RoxygenNote:	6.0.1
Author:	Joelle Barido-Sottani [aut, cre]
Maintainer:	Joelle Barido-Sottani <joelle.barido-sottani@m4x.org>

Index of help topics:

Author(s)

NA

Maintainer: NA

References

J. Barido-Sottani and T. Stadler. Accurate detection of HIV transmission clusters from phylogenetic trees using a multi-state birth-death model, BioRXiv 2017. (https://www.biorxiv.org/content/early/2017/11/10/215491)

See Also

ape

Examples

```
# Simulate a random phylogeny
set.seed(25)
tree <- ape::rtree(10)

# Calculate the log likelihood under a multi-states model with 2 states
# and full extant & extinct sampling
likelihood_MSBD(tree, shifts = matrix(c(2,1.8,2), nrow = 1),
  gamma = 0.05, lambdas = c(10, 6), mus = c(1, 0.5), sigma = 1)

# Infer the most likely multi-states birth-death model with full extant & extinct sampling
ML_MSBD(tree, initial_values = c(0.1, 10, 1), sigma = 1, time_mode = "real")
# Infer the most likely multi-states birth-death model with exponential decay
# and full extant & extinct sampling
ML_MSBD(tree, initial_values = c(0.1, 10, 0.5, 1), sigma = 1,
  stepsize = 0.1, time_mode = "mid")
```

`likelihood_MSBD` *Likelihood calculation for constantly sampled trees*

Description

Calculates the negative log likelihood of a multi-states model given a tree. This function is designed to work with constant extant and/or extinct sampling.

Usage

```
likelihood_MSBD(tree, shifts, gamma, lambdas, mus,
                lambda_rates = NULL, stepsize = NULL,
                uniform_weights = TRUE, p_lambda = 0, p_mu = 0,
                rho = 1, sigma = 0, rho_sampling = TRUE,
                add_time = 0, unresolved = FALSE)
```

Arguments

`tree`	The phylogenetic tree (in ape format) to calculate the likelihood on
`shifts`	Matrix describing the positions (edges and times) of shifts. See 'Details'.
`gamma`	Rate of state change

lambdas	Birth rates of all states
mus	Death rates of all states
lambda_rates	Rates of decay of birth rate for all states. If provided, stepsize should also be provided.
stepsize	Size of the step to use for time discretization with exponential decay, default NULL. If provided, lambda_rates should also be provided.
uniform_weights	Whether all states are weighted uniformly in shifts, default true. If false, the weights of states are calculated from the provided distributions on lambda and mu. See 'Details'.
p_lambda	Prior probability distribution on lambda, used if uniform_weights is false.
p_mu	Prior probability distribution on mu, used if uniform_weights is false.
rho	Sampling proportion on extant tips, default 1.
sigma	Sampling probability on extinct tips (tips are sampled upon extinction), default 0.
rho_sampling	If true, the most recent tips will be considered extant tips and use the sampling proportion rho. If false, all tips will be considered extinct tips and use the sampling probability sigma. Should be true for most macroevolution datasets and false for most epidemiology datasets.
add_time	The time between the most recent tip and the end of the process (>=0). This is an internal variable used in calculations for unresolved trees.
unresolved	Whether to use the 'cutoff' sampling scheme, default false. If true, the times at cutoff and the number of lineages need to be provided for all tips. This sampling scheme is not recommended for epidemiology datasets. This sampling scheme supports extinct tips outside of the unresolved parts.

Details

It is to be noted that all times are counted backwards, with the most recent tip positioned at 0.

The 'shifts' matrix is composed of 3 columns and a number of rows. Each row describes a shift: the first column is the index of the edge on which the shift happens, the second column is the time of the shift, and the third column is the index of the new state. For example the row vector (3,0.5,2) specifies a shift on edge number 3, at time 0.5, towards the state that has parameters lambdas[2], lambda_rates[2] and mus[2].

The weights w are used for calculating the transition rates q from each state i to j: $q_{i,j} = \gamma * w_{i,j}$. If `uniform_weights` is true, $w_{i,j} = \frac{1}{N-1}$ for all i,j, where N is the total number of states. If `uniform_weights` is false, $w_{i,j} = \frac{p_\lambda(\lambda_j)p_\mu(\mu_j)}{sum_{k \neq i} p_\lambda(\lambda_k)p_\mu(\mu_k)}$ where the distributions p_λ and p_μ are provided by the inputs `plambda` and `pmu`.

Value

The value of the negative log likelihood of the model given the tree.

Author(s)

NA

Examples

```
# Simulate a random phylogeny
set.seed(25)
tree <- ape::rtree(10)

# Calculate the log likelihood under a constant birth-death model (i.e, no shifts)
# with full extant & extinct sampling
likelihood_MSBD(tree, shifts = c(), gamma = 0, lambdas = 10, mus = 1, sigma = 1)
# Calculate the log likelihood under a multi-states model with 2 states
# and full extant & extinct sampling
likelihood_MSBD(tree, shifts = matrix(c(2,1.8,2), nrow = 1),
  gamma = 0.05, lambdas = c(10, 6), mus = c(1, 0.5), sigma = 1)
# Calculate the log likelihood under a multi-states model with 2 states and exponential decay
# with full extant & extinct sampling
likelihood_MSBD(tree, shifts = matrix(c(2,1.8,2), nrow = 1),
  gamma = 0.05, lambdas = c(10, 6), mus = c(1, 0.5),
  sigma = 1, stepsize = 0.01, lambda_rates = c(0.1, 0.1))
```

---

likelihood_MSBD_unresolved

Likelihood calculation for unresolved trees

Description

Calculates the negative log likelihood of a multi-states model given a tree. This function is designed to work with unresolved trees, where tips represent collapsed clades.

Usage

```
likelihood_MSBD_unresolved(tree, shifts, gamma, lambdas, mus,
                           lambda_rates = NULL, stepsize = NULL,
                           uniform_weights = TRUE, p_lambda = 0, p_mu = 0,
                           rho = 1, sigma = 0, rho_sampling = TRUE,
                         unresolved = FALSE, lineage_counts = c(), tcut = NULL)
```

Arguments

tree	The phylogenetic tree (in ape format) to calculate the likelihood on
shifts	Matrix describing the positions (edges and times) of shifts. See 'Details'.
gamma	Rate of state change
lambdas	Birth rates of all states
mus	Death rates of all states

lambda_rates	Rates of decay of birth rate for all states. If provided, stepsize should also be provided.
stepsize	Size of the step to use for time discretization with exponential decay, default NULL. If provided, lambda_rates should also be provided.
uniform_weights	Whether all states are weighted uniformly in shifts, default true. If false, the weights of states are calculated from the provided distributions on lambda and mu. See 'Details'.
p_lambda	Prior probability distribution on lambda, used if uniform_weights is false.
p_mu	Prior probability distribution on mu, used if uniform_weights is false.
rho	Sampling proportion on extant tips, default 1.
sigma	Sampling probability on extinct tips (tips are sampled upon extinction), default 0.
rho_sampling	If true, the most recent tips will be considered extant tips and use the sampling proportion rho. If false, all tips will be considered extinct tips and use the sampling probability sigma. Should be true for most macroevolution datasets and false for most epidemiology datasets.
unresolved	Whether to use the 'cutoff' sampling scheme, default false. If true, the times at cutoff and the number of lineages need to be provided for all tips. This sampling scheme is not recommended for epidemiology datasets. This sampling scheme supports extinct tips outside of the unresolved parts.
lineage_counts	Number of lineages collapsed on each tip, used if unresolved is true. Should be set to 1 for extinct tips.
tcut	Times of collapsing for each tip (i.e time of the mrca of all collapsed lineages), used if unresolved is true. Can be a single number or a vector of length the number of tips (recommended).

Details

It is to be noted that all times are counted backwards, with the most recent tip positioned at 0.

The 'shifts' matrix is composed of 3 columns and a number of rows. Each row describes a shift: the first column is the index of the edge on which the shift happens, the second column is the time of the shift, and the third column is the index of the new state. For example the row vector (3,0.5,2) specifies a shift on edge number 3, at time 0.5, towards the state that has parameters lambdas[2], lambda_rates[2] and mus[2].

The weights w are used for calculating the transition rates q from each state i to j: $q_{i,j} = \gamma * w_{i,j}$. If `uniform_weights` is true, $w_{i,j} = \frac{1}{N-1}$ for all i,j, where N is the total number of states. If `uniform_weights` is false, $w_{i,j} = \frac{p_\lambda(\lambda_j)p_\mu(\mu_j)}{sum_{k \neq i} p_\lambda(\lambda_k)p_\mu(\mu_k)}$ where the distributions p_λ and p_μ are provided by the inputs `plambda` and `pmu`.

Value

The value of the negative log likelihood of the model given the tree.

Author(s)

NA

Examples

```
# Simulate a random phylogeny
set.seed(24)
tree <- ape::rcoal(10)

# Calculate the log likelihood under a constant birth-death model (i.e, no shifts)
# with unresolved tips
likelihood_MSBD_unresolved(tree, shifts = c(), gamma = 0, lambdas = 10, mus = 1,
  unresolved = TRUE, lineage_counts = c(2,5,1,3,1,1,1,1,2,6), tcut = 0.05)
# Calculate the log likelihood under a multi-states model with 2 states and unresolved tips
likelihood_MSBD_unresolved(tree, shifts = matrix(c(2,0.7,2), nrow = 1),
  gamma = 0.05, lambdas = c(10, 5), mus = c(1, 1),
  unresolved = TRUE, lineage_counts = c(2,5,1,3,1,1,1,1,2,6), tcut = 0.05)
```

ML_MSBD *Full Maximum Likelihood inference of birth and death rates together with their changes along a phylogeny under a multi-type birth-death model.*

Description

Infers complete model from a phylogeny, including number of states, positions and timing of state changes, and parameters associated with each state. Uses a greedy approach to add states and maximum likelihood for the rest of the model.

Usage

```
ML_MSBD(tree, initial_values,
        uniform_weights=TRUE, p_lambda=0, p_mu=0,
        rho = 1, sigma=0, rho_sampling = TRUE,
        unresolved = FALSE, lineage_counts = c(), tcut = 0,
        optim_control = list(), attempt_remove=TRUE, max_nshifts=Inf,
        stepsize=NULL, no_extinction=FALSE, fixed_gamma=NULL,
        unique_lambda = FALSE, unique_mu = FALSE,
        saved_state = NULL, save_path = NULL,
        time_mode = c("real","3pos","tip","mid","root"),
        fast_optim = FALSE)
```

Arguments

`tree`	The phylogenetic tree (in ape format) to run inference on
`initial_values`	Initial values for the optimizer, to be provided as a vector in this order: gamma (optional), lambda, lambda decay rate (optional), mu (optional). See 'Details'.

uniform_weights
: Whether all states are weighted uniformly in shifts, default true. If false, the weights of states are calculated from the provided distributions on lambda and mu. See 'Details'.

p_lambda
: Probability distribution on lambda, used if uniform_weights is false.

p_mu
: Probability distribution on mu, used if uniform_weights is false.

rho
: Sampling proportion on extant tips, default 1.

sigma
: Sampling probability on extinct tips (tips are sampled upon extinction), default 0.

rho_sampling
: If true, the most recent tips will be considered extant tips and use the sampling proportion rho. If false, all tips will be considered extinct tips and use the sampling probability sigma. Should be true for most macroevolution datasets and false for most epidemiology datasets.

unresolved
: Whether to use the 'cutoff' sampling scheme, default false. If true, the times at cutoff and the number of lineages need to be provided for all tips. This sampling scheme is not recommended for epidemiology datasets. This sampling scheme supports extinct tips outside of the unresolved parts.

lineage_counts
: Number of lineages collapsed on each tip, used if unresolved is true. Should be set to 1 for extinct tips.

tcut
: Times of collapsing for each tip (i.e time of the mrca of all collapsed lineages), used if unresolved is true. Can be a single number or a vector of length the number of tips (recommended).

optim_control
: Control list for the optimizer, corresponds to control input in optim function, see ?optim for details.

attempt_remove
: Whether to attempt to remove shifts at the end of the inference, default true. Full greedy approach if false.

max_nshifts
: Maximum number of shifts to test for, default Inf.

stepsize
: Size of the step to use for time discretization with exponential decay. Default NULL, will use exponential decay if a value is provided.

no_extinction
: Whether to use the Yule process (mu=0) for all states, default false. If true no initial value for mu is needed.

fixed_gamma
: Value to which gamma should be fixed, default NULL. If provided no initial value for gamma is needed.

unique_lambda
: Whether to use the same value of lambda for all states, default false. If true and exponential decay is active all states will also share the value of lambda rate.

unique_mu
: Whether to use the same value of mu for all states, default false.

saved_state
: If provided, the inference will be restarted from this state.

save_path
: If provided, the progress of the inference will be saved to this path after each optimization step.

time_mode
: String controlling how time positions of the shifts will be inferred. See 'Details'.

fast_optim
: Whether to use the faster mode of optimization, default FALSE. If true only rates associated with the state currently being added to the tree and its ancestor will be optimized at each step, otherwise all rates are optimized.

Details

It is to be noted that all times are counted backwards, with the most recent tip positioned at 0.

Five time modes are possible for the input `time_mode`. In `real` mode, the time positions of shifts will be estimated in addition to the other parameters. In `tip` mode, the shifts will be placed at 10% of the length of the edge. In `mid` mode, the shifts will be placed at 50% of the length of the edge. In `root` mode, the shifts will be placed at 90% of the length of the edge. In `3pos` mode, the three "tip", "mid" and "root" positions will be tested.

The weights w are used for calculating the transition rates q from each state i to j: $q_{i,j} = \gamma * w_{i,j}$. If `uniform_weights` is true, $w_{i,j} = \frac{1}{N-1}$ for all i,j, where N is the total number of states. If `uniform_weights` is false, $w_{i,j} = \frac{p_\lambda(\lambda_j)p_\mu(\mu_j)}{sum_{k \neq i} p_\lambda(\lambda_k)p_\mu(\mu_k)}$ where the distributions p_λ and p_μ are provided by the inputs `p_lambda` and `p_mu`.

Initial values for the optimization need to be provided as a vector and contain the following elements (in order): an initial value for gamma, which is required unless `fixed_gamma` is provided, an initial value for lambda which is always required, an initial value for lambda decay rate, which is required if `stepsize` is provided, and an initial value for mu, which is required unless `no_extinction` is true. An error will be raised if the number of initial values provided does not match the one expected from the rest of the settings, and the function will fail if the likelihood cannot be calculated at the initial values.

Value

Returns a list describing the most likely model found, with the following components:

`likelihood`	the negative log likelihood of the model
`shifts.edge`	the indexes of the edges where shifts happen, 0 indicates the root state
`shifts.time`	the time positions of shifts
`gamma`	the rate of state change
`lambdas`	the birth rates of all states
`lambda_rates`	if exponential decay was activated, the rates of decay of birth rate for all states
`mus`	the death rates of all states
`best_models`	a vector containing the negative log likelihood of the best model found for each number of states tested (best_models[i] corresponds to i states, i.e i-1 shifts)

All vectors are indexed in the same way, so that the state with parameters lambdas[i], lambda_rates[i] and mus[i] starts on edge shifts.edge[i] at time shifts.time[i].

Author(s)

NA

Examples

```
# Simulate a random phylogeny with extinct samples
set.seed(25)
tree <- ape::rtree(10)

# Infer the most likely multi-states birth-death model
# with full extant & extinct sampling
ML_MSBD(tree, initial_values = c(0.1, 10, 1), sigma = 1, time_mode = "mid")
# Infer the most likely multi-states birth-death model with exponential decay
# and full extant & extinct sampling
ML_MSBD(tree, initial_values = c(0.1, 10, 0.5, 1), sigma = 1,
  stepsize = 0.1, time_mode = "mid")

# Simulate a random phylogeny with extant samples
set.seed(24)
tree2 <- ape::rcoal(10)

# Infer the most likely multi-states Yule model with partial extant sampling
ML_MSBD(tree2, initial_values = c(0.1, 10), no_extinction = TRUE,
  rho = 0.5, time_mode = "mid")
# Infer the most likely multi-states birth-death model with full extant sampling
# and unresolved extant tips
ML_MSBD(tree2, initial_values = c(0.1, 10, 1), unresolved = TRUE,
  lineage_counts = c(2,5,1,3,1,1,1,1,2,6), tcut = 0.05, time_mode = "mid")
```

Index