Package: ML.MSBD (via r-universe)

October 21, 2024

Type Package

Title Maximum Likelihood Inference on Multi-State Trees

Version 1.2.1

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, T. G. Vaughan and T. Stadler (2018) <doi:10.1098/rsif.2018.0512>.

License GPL-3

Imports ape (>= 5.1), foreach

Suggests knitr, rmarkdown, doParallel

RoxygenNote 6.1.1

VignetteBuilder knitr

NeedsCompilation no

Author Joelle Barido-Sottani [aut, cre]

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

Date/Publication 2021-04-16 18:00:02 UTC

Repository https://bjoelle.r-universe.dev

RemoteUrl https://github.com/cran/ML.MSBD

RemoteRef HEAD

RemoteSha cc6b672e81b8056795eb60bcdee5895dc5d50a47

Contents

`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, T. G. Vaughan and T. Stadler (2018) <doi:10.1098/rsif.2018.0512>.

Details

Package:	ML.MSBD
Type:	Package
Title:	Maximum Likelihood Inference on Multi-State Trees
Version:	1.2.1
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
License:	GPL-3
Imports:	ape (>= 5.1), foreach
Suggests:	knitr, rmarkdown, doParallel
RoxygenNote:	6.1.1
VignetteBuilder:	knitr
Author:	Joelle Barido-Sottani [aut, cre]
Maintainer:	Joelle Barido-Sottani <joelle.barido-sottani@m4x.org>

Index of help topics:

`ML.MSBD-package`	`Maximum Likelihood Inference on Multi-State Trees`
`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.`
`likelihood_MSBD`	`Likelihood calculation for randomly sampled trees`
`likelihood_MSBD_unresolved`	`Likelihood calculation for unresolved trees`

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
## Not run: 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
## Not run: ML_MSBD(tree, initial_values = c(0.1, 10, 0.5, 1), sigma = 1,
  stepsize = 0.1, time_mode = "mid")
## End(Not run)
```

`likelihood_MSBD` *Likelihood calculation for randomly 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`	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. To use exponential decay, `stepsize` should also be provided.
`stepsize`	Size of the step to use for time discretization with exponential decay, default NULL. To use exponential decay, `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 distributions `p_lambda` and `p_mu`. See 'Details'.
`p_lambda`	Prior probability distribution on lambdas, used if `uniform_weights = FALSE`.
`p_mu`	Prior probability distribution on mus, used if `uniform_weights = FALSE`.
`rho`	Sampling proportion on extant tips, default 1.
`sigma`	Sampling probability on extinct tips (tips are sampled upon extinction), default 0.
`rho_sampling`	Whether the most recent tips should be considered extant tips, sampled with sampling proportion `rho`. If FALSE, all tips will be considered extinct tips, sampled with 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 this tree is the backbone of an unresolved tree. This is an internal variable used in calculations for unresolved trees.

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 = TRUE`, $w_{i,j} = \frac{1}{N-1}$ for all i,j, where N is the total number of states. If `uniform_weights = 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`.

Value

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

Examples

```
# Input a phylogeny
tree <- ape::read.tree(text = "(((t4:0.7293960718,(t1:0.450904974,t3:0.09259337652)
       :0.04068535892):0.4769176776,t8:0.1541864066):0.7282000314,((t7:0.07264320855,
       (((t5:0.8231869878,t6:0.3492440532):0.2380232813,t10:0.2367582193):0.5329497182,
       t9:0.1016243151):0.5929288475):0.3003101915,t2:0.8320755605):0.2918686506);")

# 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. This sampling scheme is not recommended for epidemiology datasets. The MRCA times of collapsed clades and the number of collapsed lineages need to be provided for all tips. If neither is provided the function will default to random sampling. Extinct tips can be present outside of the unresolved parts, but not below the time(s) set for `tcut`.

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, lineage_counts = c(), tcut = NULL)
```

Arguments

`tree`	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. To use exponential decay, `stepsize` should also be provided.
`stepsize`	Size of the step to use for time discretization with exponential decay, default NULL. To use exponential decay, `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 distributions `p_lambda` and `p_mu`. See 'Details'.
`p_lambda`	Prior probability distribution on lambdas, used if `uniform_weights = FALSE`.
`p_mu`	Prior probability distribution on mus, used if `uniform_weights = FALSE`.
`rho`	Sampling proportion on extant tips, default 1.
`sigma`	Sampling probability on extinct tips (tips are sampled upon extinction), default 0.
`rho_sampling`	Whether the most recent tips should be considered extant tips, sampled with sampling proportion `rho`. If FALSE, all tips will be considered extinct tips, sampled with sampling probability `sigma`. Should be TRUE for most macroevolution datasets and FALSE for most epidemiology datasets.
`lineage_counts`	Number of lineages collapsed on each tip. Should be set to 1 for extinct tips.
`tcut`	Times of clade collapsing for each tip (i.e time of the MRCA of all collapsed lineages). Can be a single number or a vector of length the number of tips.

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 = TRUE`, $w_{i,j} = \frac{1}{N-1}$ for all i,j, where N is the total number of states. If `uniform_weights = 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`.

Value

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

Examples

```
# Input a phylogeny
tree <- ape::read.tree(text = "(t3:0.9703302342,((t4:0.1999577823,(t2:0.1287530271,
     (t7:0.08853561159,(t8:0.07930237712,t9:0.07930237712):0.009233234474):0.04021741549):
```

```
        0.07120475526):0.4269919425,(((t10:0.0191876225,t5:0.0191876225):0.04849906822,
        t6:0.06768669072):0.1672340445,t1:0.2349207353):0.3920289896):0.3433805094);”)

# 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,
                           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),
                           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 a complete MSBD model from a phylogeny, including the most likely number of states, positions and times of state changes, and parameters associated with each state. Uses a greedy approach to add states and Maximum Likelihood inference for the other parameters.

Usage

```
ML_MSBD(tree, initial_values, uniform_weights = TRUE, p_lambda = 0,
  p_mu = 0, rho = 1, sigma = 0, rho_sampling = TRUE,
  lineage_counts = c(), tcut = 0, stepsize = NULL,
  no_extinction = FALSE, fixed_gamma = NULL, unique_lambda = FALSE,
  unique_mu = FALSE, optim_control = list(), attempt_remove = TRUE,
  max_nshifts = Inf, saved_state = NULL, save_path = NULL,
  time_mode = c(”3pos”, ”tip”, ”mid”, ”root”), fast_optim = FALSE,
  parallel = FALSE, ncores = getOption(”mc.cores”, 2L))
```

Arguments

`tree`	Phylogenetic tree (in ape format) to calculate the likelihood 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 distributions `p_lambda` and `p_mu`. See ’Details’.
`p_lambda`	Prior probability distribution on lambdas, used if `uniform_weights = FALSE`.
`p_mu`	Prior probability distribution on mus, used if `uniform_weights = FALSE`.
`rho`	Sampling proportion on extant tips, default 1.

`sigma`	Sampling probability on extinct tips (tips are sampled upon extinction), default 0.
`rho_sampling`	Whether the most recent tips should be considered extant tips, sampled with sampling proportion `rho`. If FALSE, all tips will be considered extinct tips, sampled with sampling probability `sigma`. Should be TRUE for most macroevolution datasets and FALSE for most epidemiology datasets.
`lineage_counts`	For trees with clade collapsing. Number of lineages collapsed on each tip. Should be set to 1 for extinct tips.
`tcut`	For trees with clade collapsing. Times of clade collapsing for each tip (i.e time of the MRCA of all collapsed lineages). Can be a single number or a vector of length the number of tips.
`stepsize`	Size of the step to use for time discretization with exponential decay, default NULL. To use exponential decay, an initial value for `lambda_rates` should also be 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 same value for `lambda_rate`.
`unique_mu`	Whether to use the same value of `mu` for all states, default FALSE.
`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. If FALSE, use a pure greedy algorithm.
`max_nshifts`	Maximum number of shifts to test for, default `Inf`.
`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 the time positions of inferred shifts. 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.
`parallel`	Whether the computation should be run in parallel, default FALSE. Will use a user-defined cluster if one is found, otherwise will define its own.
`ncores`	Number of cores to use for a parallel computation.

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 `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 = TRUE`, $w_{i,j} = \frac{1}{N-1}$ for all i,j, where N is the total number of states. If `uniform_weights = 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 = 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]`.

Examples

```
# Input a phylogeny
tree <- ape::read.tree(text = "(((t4:0.7293960718,(t1:0.450904974,t3:0.09259337652)
        :0.04068535892):0.4769176776,t8:0.1541864066):0.7282000314,((t7:0.07264320855,
        (((t5:0.8231869878,t6:0.3492440532):0.2380232813,t10:0.2367582193):0.5329497182,
        t9:0.1016243151):0.5929288475):0.3003101915,t2:0.8320755605):0.2918686506);")

# Infer the most likely multi-states birth-death model
# with full extant & extinct sampling
```

```
## Not run: 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
## Not run: ML_MSBD(tree, initial_values = c(0.1, 10, 0.5, 1), sigma = 1,
                 stepsize = 0.1, time_mode = "mid")
## End(Not run)

# Input a phylogeny with extant samples
tree2 <- ape::read.tree(text = "(t3:0.9703302342,((t4:0.1999577823,(t2:0.1287530271,
     (t7:0.08853561159,(t8:0.07930237712,t9:0.07930237712):0.009233234474):0.04021741549):
       0.07120475526):0.4269919425,(((t10:0.0191876225,t5:0.0191876225):0.04849906822,
       t6:0.06768669072):0.1672340445,t1:0.2349207353):0.3920289896):0.3433805094);")

# Infer the most likely multi-states Yule model with partial extant sampling
## Not run: ML_MSBD(tree2, initial_values = c(0.1, 10), no_extinction = TRUE,
                  rho = 0.5, time_mode = "mid")
## End(Not run)
# Infer the most likely multi-states birth-death model with full extant sampling
# and unresolved extant tips
## Not run: ML_MSBD(tree2, initial_values = c(0.1, 10, 1),
                lineage_counts = c(2,5,1,3,1,1,1,1,2,6), tcut = 0.05, time_mode = "mid")
## End(Not run)
```

Index