Package 'phybreak'

July 30, 2017

Type Package

Title Analysis of Outbreaks with Sequence Data

Version 0.2.0

Date 2017-07-26

Description Implementation the outbreak analysis method described by Klinkenberg et al (2017) <doi:10.1371/journal.pcbi.1005495>. Simulate outbreaks, analyse datasets by creating samples from the posterior distribution with a Markov-Chain Monte Carlo sampler, and summarize the output.

License GPL (>= 2)

LazyData TRUE

Depends R(>= 3.1.0), phangorn, ape

Imports Rcpp (>= 0.12.1), methods, stats, stringr, utils

Suggests coda, phytools, OutbreakTools, testthat, knitr, rmarkdown

LinkingTo Rcpp

RoxygenNote 6.0.1

VignetteBuilder knitr

NeedsCompilation yes

Author Don Klinkenberg [aut, cre]

Maintainer Don Klinkenberg <don@xs4all.nl>

Repository CRAN

Date/Publication 2017-07-30 14:49:53 UTC

R topics documented:

`AvianFluH7N7_2003` *Avian influenza (H7N7) epidemic*

Description

DNA sequences (base order randomized) of NA, HA, and PB2 genes of samples taken from farms during the Dutch avian influenza (H7N7) epidemic in 2003. The dataset contains 241 farms, of which 231 were sampled and sequenced (for the other 10, identified by the prefix UNK, only the detection day + 2 days is given, as in Klinkenberg et al (2017)). The labels of the sequences refer to the labels in GISAID (`www.gisaid.org`), where the sequences in correct base order can be found.

Usage

```
AvianFluH7N7_2003
```

Format

A list with four elements:

sequences_HA the HA gene

sequences_NA the NA gene

sequences_PB2 the PB2 gene

sampledays the days at which the samples were taken (or the detection day + 2 days, for the 10 unsampled farms)

References

Bataille et al. (2011) Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLoS Pathog*, **7**(6): e1002094.

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

`burnin.phybreak` *MCMC updating of a phybreak-object.*

Description

This function allows the MCMC chain to burn in. If used after samples have been taken (with `sample.phybreak`), these samples will be returned unchanged in the output.

Usage

```
burnin.phybreak(phybreak.object, ncycles, keepphylo = NULL,
  phylotopology_only = 0)
```

Arguments

`phybreak.object`	An object of class `phybreak`.
`ncycles`	Number of iterations to be carried out. Each iteration does one update of all parameters and tree updates with each host as focal host once.
`keepphylo`	The proportion of tree updates keeping the phylotree intact. If there is more than one sample per host, keepphylo should be 0. If set to NULL (default), this is done automatically, otherwise it is set to 0.2.
`phylotopology_only`	The proportion of tree updates in which only the within-host minitree topology is sampled, and the transmission tree as well as coalescence times are kept unchanged.

Value

The `phybreak`-object provided as input, with variables and parameters changed due to the updating.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First create a phybreak-object
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation)

MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 50)
```

FMD_2001 *Foot-and-Mouth Disease outbreak (2001)*

Description

DNA nucleotide patterns and sampling days of FMD virus from 15 farms in the UK. The nucleotides per sample are not ordered as a virus sequence, but do contain the correct numbers of each nucleotide.

Usage

```
FMD_2001
```

Format

A matrix with 15 rows (one per farm) and 8196 column (one per nucleotide); the names contain sampling days

References

Cottam et al. (2008) Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc R Soc B*, **275**: 887-895.

FMD_2007 *Foot-and-Mouth Disease outbreak (2007)*

Description

DNA sequences and sampling days of FMD virus from 10 farms in the UK (11 sequences)

Usage

```
FMD_2007
```

Format

A list with two elements:

sequences the sequences (class `DNAbin`)

dates the sampling days (class `Date`)

References

Cottam et al. (2008) Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. *PLoS Pathog*, **4**(4): e1000050.

get.phybreak *Accessing a phybreak object*

Description

Accessing a phybreak object

Usage

```
get.tree(phybreak.object, samplenr = 0)

get.parameters(phybreak.object, samplenr = 0, whichpars = "posterior")

get.mcmc(phybreak.object, thin = 1, nkeep = Inf)

get.phylo(phybreak.object, samplenr = 0, simmap = FALSE)

get.multiPhylo(phybreak.object, thin = 1, nkeep = Inf)

get.seqdata(phybreak.object)
```

Arguments

phybreak.object	An object of class phybreak.
samplenr	The posterior tree sample to choose. If `samplenr = 0`, the current state is used.
whichpars	Which parameters to return. Either a vector with parameter names, or `"all"` for all parameters, or `"posterior"` for parameters for which a posterior is sampled.
thin	Thinning interval.
nkeep	Number of samples to keep, counting from tail of the chain.
simmap	Whether to include class `"simmap"` elements (package **phytools**), colouring the branches on the tree to indicate hosts. Is used by `plotPhylo`.

Functions

- `get.tree`: A `data.frame` with current (`samplenr = 0`) or sampled infectors and infection times.
- `get.parameters`: A named vector with current (`samplenr = 0`) or sampled parameter values.
- `get.mcmc`: An object of class `"mcmc"` (package **coda**), with sampled parameters, infection times, and infectors.
- `get.phylo`: Returns an object of class `phylo` ans optionally of class `"simmap"` (package **phytools**).
- `get.multiPhylo`: Returns an object of class `multiphylo`.
- `get.seqdata`: The sequence data in class `"phyDat"` (package **phangorn**).

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

get.tree(MCMCstate)
get.parameters(MCMCstate)
codaobject <- get.mcmc(MCMCstate, thin = 2)
plot.phylo(get.phylo(MCMCstate))
get.seqdata(MCMCstate)

#function from package phangorn:
phangorn::parsimony(get.phylo(MCMCstate), get.seqdata(MCMCstate))

tree0 <- get.phylo(MCMCstate)
seqdata <- get.seqdata(MCMCstate)
phangorn::pml(tree0, seqdata,
              rate = 0.75*get.parameters(MCMCstate)["mu"])
logLik(MCMCstate, genetic = TRUE, withinhost = FALSE,
       sampling = FALSE, generation = FALSE)
              #should give the same result as 'pml'
```

---

infectorsets *Sampled infectors for each host in a phybreak-object.*

Description

The function takes a phybreak-object containing MCMC-samples, and returns for each host a table with posterior infectors, with support per infector.

Usage

```
infectorsets(phybreak.object, which.hosts = "all", percentile = 0.95,
  minsupport = 0, samplesize = Inf, infector.name = TRUE,
  support = c("proportion", "count"))
```

Arguments

phybreak.object	An object of class phybreak.
which.hosts	A vector with hosts (positions in the dataset), or "all" for all hosts.
percentile	Return infectors ordered by support, until a cumulative support indicated by percentile, support measured by proportion (value between 0 and 1).
minsupport	Only return infectors with more support than minsupport. Values in the range [0,1] are interpreted as support in "proportion" of posterior samples, values > 1 as code"count" of posterior samples.
samplesize	The number of samples to include (taken from the tail of the MCMC-chain).
infector.name	Whether to return the names of the infectors, or their position in the dataset.
support	Whether to return the support (= posterior probability) for each infector as a "proportion" or as a "count" of posterior trees in which that transmission link or transmission cluster is present.

Value

A named list with data.frames, each with a vector of infectors and a vector of supports.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object containing samples.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation$sequences, times = simulation$sample.times)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

infectorsets(MCMCstate)
```

---

logLik.phybreak *Log-likelihood of a phybreak-object.*

Description

The likelihood of a phybreak-object is calculated, with the option to include or exclude parts of the likelihood for genetic data, phylogenetic tree (within-host model), sampling times and generation times.

Usage

```
## S3 method for class 'phybreak'
logLik(object, genetic = TRUE, withinhost = TRUE,
  sampling = TRUE, generation = TRUE, ...)
```

Arguments

object	An object of class phybreak.
genetic	Whether to include the likelihood of the mutation model.
withinhost	Whether to include the likelihood of within-host (coalescent) model.
sampling	Whether to include the likelihood of the sampling model (sampling intervals).
generation	Whether to include the likelihood of the transmission model (generation intervals).
...	Some methods for this generic require additional arguments. None are used in this method.

Details

The sequence likelihood is calculated by Felsenstein's pruning algorithm, assuming a prior probability of 0.25 for each nucleotide. The within-host likelihood is the likelihood of coalescence times given the within-host model and slope. The generation interval and sampling interval likelihood are log-densities of the gamma distributions for these variables.

Value

The log-likelihood as an object of class logLik.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object containing samples.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation)
logLik(MCMCstate)

MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
logLik(MCMCstate)

tree0 <- get.phylo(MCMCstate)
seqdata <- get.seqdata(MCMCstate)
pml(tree0, seqdata, rate = 0.75*get.parameters(MCMCstate)["mu"])
```

```
logLik(MCMCstate, genetic = TRUE, withinhost = FALSE,
       sampling = FALSE, generation = FALSE) #should give the same result as 'pml'
```

`M_tuberculosis_2013` *Mycobacterium tuberculosis outbreak*

Description

DNA SNP patterns and sampling days of an M. tuberculosis outbreak with 33 cases in Canada (Didelot et al, 2013). The dataset contained only 20 SNP patterns with unspecified nucleotides, so the sequences in this dataset contain only ‘a‘ and ‘c‘ on these 20 loci. Additional loci with only ‘a‘ are added to a total sequence length of 440,000, which is 10

Usage

```
M_tuberculosis_2013
```

Format

A list with two elements:

sequences_Mtb the sequences (class `DNAbin`)

dates_Mtb the sampling days (class `Date`)

References

Didelot et al. (2013) Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol*, **31**(7): 1869-1879.

`phybreak` *Create a phybreak-object from data and prior distributions.*

Description

phybreak takes as data either an `'obkData'`-object or a `phybreakdata`-object with sequences (individuals in rows, nucleotides in columns). Both `'obkData'` and `phybreakdata` contain at least sequences and sampling times, and potentially more. Parameter values are used as initial values in the MCMC-chain or kept fixed. All variables are initialized by random samples from the prior distribution, unless a complete tree is given in the data and should be used (`use.tree = TRUE`). It is also possible to provide only sequences as data, and sampling times separately.

Usage

```
phybreak(dataset, times = NULL, mu = NULL, gen.shape = 3, gen.mean = 1,
  sample.shape = 3, sample.mean = 1, wh.model = 3, wh.slope = 1,
  est.gen.mean = TRUE, prior.mean.gen.mean = 1, prior.mean.gen.sd = Inf,
  est.sample.mean = TRUE, prior.mean.sample.mean = 1,
  prior.mean.sample.sd = Inf, est.wh.slope = TRUE, prior.wh.shape = 3,
  prior.wh.mean = 1, use.tree = FALSE)
```

Arguments

dataset	An object with sequences plus additional data. (class 'obkData' or 'phybreakdata'). All nucleotides that are not 'a', 'c', 'g', or 't', will be turned into 'n'. If the data are provided as an object of class 'obkData', these should contain sequences and sampling times as metadata with these sequences. The object may also contain infector and infection date vectors in the individuals slot, plus (at least) one tree in the 'trees' slot (class 'multiPhylo'). Data provided as an object of class 'phybreakdata' contain sequences and sampling.times, and potentially sim.infection.times, sim.infectors, and sim.tree. Prepare your data in this format by phybreakdata or by simulation with sim.phybreak. It is also possible to provide only sequences as data, (class 'DNAbin', 'phyDat', or a matrix with nucleotides, each row a host, each column a nucleotide), and corresponding sampling times in the separate times argument.
times	Vector of sampling times, needed if the data consist of only sequences. If the vector is named, these names will be used to identify the hosts.
mu	Initial value for mutation rate (defined per site per unit of time). If NULL (default), then an initial value is calculated by dividing the number of SNPs by the product: 0.75 times 'total sequence length' times 'sum of edge lengths in the initial phylogenetic tree'. NOTE: mutation is defined as assignment of a random nucleotide at a particular site; this could be the nucleotide that was there before the mutation event. Therefore, the actual rate of change of nucleotides is 0.75*mu.
gen.shape	Shape parameter of the generation interval distribution (not estimated).
gen.mean	Initial value for the mean generation interval, i.e. the interval between infection of a secondary case by a primary case.
sample.shape	Shape parameter of the sampling interval distribution (not estimated), i.e. the interval between infection and sampling of a host.
sample.mean	Initial value for the mean sampling interval.
wh.model	The model for within-host pathogen dynamics (effective pathogen population size = N*gE = actual population size * pathogen generation time), used to simulate coalescence events. Options are: 1. Effective size = 0, so coalescence occurs 'just before' transmission in the infector 2. Effective size = Inf, so coalescence occurs 'just after' transmission in the infectee

3. Effective size at time t after infection = `wh.slope * t`

`wh.slope` Initial value for the within-host slope, used if `wh.model = 3`.

`est.gen.mean` Whether to estimate the mean generation interval or keep it fixed.

`prior.mean.gen.mean`
Mean of the (gamma) prior distribution of mean generation interval `mG` (only if `est.gen.mean = TRUE`).

`prior.mean.gen.sd`
Standard deviation of the (gamma) prior distribution of mean generation interval `mG` (only if `est.gen.mean = TRUE`).

`est.sample.mean`
Whether to estimate the mean sampling interval or keep it fixed.

`prior.mean.sample.mean`
Mean of the (gamma) prior distribution of mean sampling interval `mS` (only if `est.sample.mean = TRUE`).

`prior.mean.sample.sd`
Standard deviation of the (gamma) prior distribution of mean sampling interval `mS` (only if `est.sample.mean = TRUE`).

`est.wh.slope` Whether to estimate the within-host slope or keep it fixed.

`prior.wh.shape` Shape parameter of the (gamma) prior distribution of `slope` (only if `est.wh.slope = TRUE`).

`prior.wh.mean` Mean of the (gamma) prior distribution of `slope` (only if `est.wh.slope = TRUE`).

`use.tree` Whether to use the transmission and phylogenetic tree given in data of class `'obkData'`, to create a `phybreak`-object with an exact copy of the outbreak. This requires more data in `data`: the slot `individuals` with vectors `infector` and `date`, and the slot `trees` with at least one phylogenetic tree. Such data can be simulated with `sim.phybreak`.

Value

An object of class `phybreak` with the following elements

d a `list` with data, i.e. names, sequences, sampling times, and total number of SNPs.

v a `list` with current state of all nodes in the tree: times, hosts in which they reside, parent nodes, node types (sampling, coalescent, or transmission)

p a `list` with the parameter values

h a `list` with helper information for the MCMC-method: `si.mu` and `si.wh` for efficiently proposing `mu` and `slope`, matrix `dist` with weights for infector sampling based on sequence distances, `logicals` `est.mG`, `est.mS`, and `est.wh.slope` whether to estimate mean generation interval `mG`, mean sampling interval `mS`, and within-host `slope`, and parameters for the priors of `mG`, `mS`, and `slope`.

s an empty `list` that will contain vector and matrices with the posterior samples; in matrices, the rows are nodes in the phylogenetic tree, the columns are the samples

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
simulation <- sim.phybreak(obsize = 10)
MCMCstate <- phybreak(data = simulation)

simulation <- sim.phybreak(obsize = 10)
MCMCstate <- phybreak(data = simulation, use.tree = TRUE)


sampletimedata <- c(0,2,2,4,4)
sampleSNPdata <- matrix(c("a","a","a","a","a",
                          "a","c","c","c","c",
                          "t","t","t","g","g"), nrow = 5)
dataset <- phybreakdata(sequences = sampleSNPdata, sample.times = sampletimedata)
MCMCstate <- phybreak(data = dataset)

### also possible without 'phybreakdata' as intermediate,
### but not with additional data (future implementation)
MCMCstate <- phybreak(data = sampleSNPdata, times = sampletimedata)
```

---

phybreakdata *Create a phybreakdata-object from raw data.*

Description

phybreakdata takes as data sequences and sampling times and makes a `phybreakdata` object. If no host names are provided, each sample is assumed to be associated with a separate host. The number of sequences should be equal to the length of the sampling time vector, and the position identifies the host (unless named vectors are provided). Sample names can be provided separately; otherwise it will be tried to extract them from the sequences or sampling times. It is also possible to include (otherwise unobserved) simulated data: sim.infection times, sim.infectors, and a (phylogenetic) sim.tree. This is done automatically when using `sim.phybreak`.

Usage

```
phybreakdata(sequences, sample.times, sample.names = NULL,
  host.names = sample.names, sim.infection.times = NULL,
  sim.infectors = NULL, sim.tree = NULL)
```

Arguments

`sequences`	Sequence data of class `'DNAbin'`, `'phyDat'`, or a `matrix` with nucleotides, each row a host, each column a nucleotide). In a matrix, nucleotides should be lowercase letters. All undefined nucleotides or ambiguity codes will be turned into `'n'`.

`sample.times` A vector of sampling times (`numerical` or `Date`).

`sample.names` A vector with sample names.

`host.names` A vector with host names. The vector identifies the host for each sample, so should be of the same length as `sample.times`.

`sim.infection.times` A vector with infection times (`numerical` or `Date`).

`sim.infectors` A vector with infectors, either by name or by position (use 0 for the index case).

`sim.tree` A tree of class `'phylo'`, with tip names identifying the hosts.

Value

An object of class `phybreakdata` with the following elements

sequences a `'phyDat'`-object with the sequence data.

sample.times a named `vector` with the sample times.

sample.hosts a named `vector` with the hosts from whom the samples have been taken.

sim.infection.times a named `vector` with the (simulated) infection times (if provided).

sim.infectors a named `vector` with the (simulated) infectors (if provided).

sim.tree a `'phylo'`-object with the (simulated) phylogenetic tree (if provided).

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
sampletimedata <- c(0,2,2,4,4)
sampleSNPdata <- matrix(c("a","a","a","a","a",
                          "a","c","c","c","c",
                          "t","t","t","g","g"), nrow = 5,
                          dimnames = list(LETTERS[1:5], NULL))
dataset <- phybreakdata(sequences = sampleSNPdata, sample.times = sampletimedata)
```

phylotree *Maximum clade credibility tree.*

Description

Identify the maximum clade credibility tree from a phybreak-object containing posterior samples.

Usage

```
phylotree(phybreak.object, samplesize = Inf, support = c("proportion",
  "count"), phylo.class = FALSE)
```

Arguments

phybreak.object	An object of class phybreak.
samplesize	The number of posterior samples that is used, taken from the tail.
support	Whether to return the support (= posterior probability) for each infector as a "proportion" or as a "count" of posterior trees in which that transmission link or transmission cluster is present.
phylo.class	Whether to return an object of class "phylo", in which case a single tree from the posterior is returned (not with summary infection times).

Value

A data.frame with per item (=node) its parent and support per clade, and optionally summary node times. The first n nodes are the samples, the last n-1 nodes are the internal nodes.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object containing samples.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation$sequences, times = simulation$sample.times)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

phylotree(MCMCstate)
plot(phylotree(MCMCstate, phylo.class = TRUE))
```

plot.phybreak *Plotting a phybreak object.*

Description

Plots a phybreak-object twice: (1) as transmission tree and (2) as phylogenetic tree, using the default graphical parameters of plotTrans and plotPhylo. The default is to plot the current state, but any posterior sample can be chosen, as well as various consensus trees. Consensus tree "edmonds" plots only a transmission tree, consensus tree "mcc" only a phylogenetic tree.

Usage

```
## S3 method for class 'phybreak'
plot(x, plot.which = c("sample", "edmonds", "mpc", "mtcc",
  "mcc"), samplenr = 0, ...)
```

Arguments

x	An object of class phybreak.
plot.which	Either "sample" to plot the current state or a selected posterior sample, "mpc" or "mtcc" to plot a consensus transmission tree (see transtree) or "mcc" to plot the maximum clade credibility tree (see phylotree).
samplenr	If plot.which = "sample", this indicates which posterior tree should be plotted: samplenr = 0 to plot the current state.
...	Some methods for this generic require additional arguments. None are used in this method.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object containing samples.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation$sequences, times = simulation$sample.times)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

plot(MCMCstate, plot.which = "mpc")
```

`plot.phybreakdata` *Plotting a phybreakdata object.*

Description

Plots a phybreakdata-object twice: (1) as transmission tree and (2) as phylogenetic tree, using the default graphical parameters of `plotTrans` and `plotPhylo`. The default is to plot the current state, but any posterior sample can be chosen, as well as various consensus trees. Consensus tree "edmonds" plots only a transmission tree, consensus tree "mcc" only a phylogenetic tree.

Usage

```
## S3 method for class 'phybreakdata'
plot(x, ...)
```

Arguments

x	An object of class phybreakdata.
...	Some methods for this generic require additional arguments. None are used in this method.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object containing samples.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation$sequences, times = simulation$sample.times)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

plot(MCMCstate, plot.which = "mpc")
```

plotPhylo *Plotting a phybreak object phylogenetic tree.*

Description

Plots a phybreak-object as phylogenetic tree with coloured branches indicating hosts. The default is to plot the current state, but any posterior sample can be chosen, as well as various consensus trees.

Usage

```
plotPhylo(x, plot.which = c("sample", "mpc", "mtcc", "mcc"), samplenr = 0,
  ...)
```

Arguments

x	An object of class phybreak.
plot.which	Either "sample" to plot the current state or a selected posterior sample, "mpc" or "mtcc" to plot a consensus transmission tree (see transtree) or "mcc" to plot the maximum clade credibility tree (see phylotree).
samplenr	If plot.which = "sample", this indicates which posterior tree should be plotted: samplenr = 0 to plot the current state.
...	Additional options for plotSimmap.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object containing samples.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation$sequences, times = simulation$sample.times)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

plot(MCMCstate, plot.which = "mpc")
```

plotTrans *Plotting a phybreak object transmission tree.*

Description

Plots a phybreak-object as transmission tree. The default is to plot the current state, but any posterior sample can be chosen, as well as various consensus trees; in that case, coloured arrows indicate posterior support.

Usage

```
plotTrans(x, plot.which = c("sample", "edmonds", "mpc", "mtcc"),
  samplenr = 0, mar = 0.1 + c(4, 0, 0, 0), label.cex = NULL,
  label.space = 0.15, label.adj = 0, arrow.lwd = 1, arrow.length = NULL,
  arrow.col = NULL, sample.pch = 4, sample.lwd = NULL,
  sample.cex = label.cex, polygon.col = "gray", polygon.border = NA,
  line.lty = 3, xlab = "Time", axis.cex = 1, title.cex = 1, ...)
```

Arguments

x	An object of class phybreak.
plot.which	Either "sample" to plot the current state or a selected posterior sample, "mpc" or "mtcc" to plot a consensus transmission tree (see transtree) or "mcc" to plot the maximum clade credibility tree (see phylotree).
samplenr	If plot.which = "sample", this indicates which posterior tree should be plotted: samplenr = 0 to plot the current state.
mar	Plot margins.
label.cex	Size of host names, as in par("cex"). Defaults to a value between 0.5 and 1 depending on outbreak size.
label.space	Scales the space at the right-hand side to place the host names.
label.adj	Left-right adjustment of host names.
arrow.lwd	Arrow width.
arrow.length	Arrow point length, as default automatically scaled with outbreak size.
arrow.col	Arrow colour. Defaults to "black" if plot.which = sample, and otherwise to five colours c("blue", "green", "orange", "red", "purple") indicating posterior support of infectors in bins of 0.2 width, from low to high support. Any vector of colours will be divided into equal-sized bins.
sample.pch	Character par("pch") used for sampling events.
sample.lwd	Line width of sampling event character.
sample.cex	Size of sampling event character.
polygon.col	Color of polygons indicating generation interval distributions.
polygon.border	Border of polygon.

line.lty	Line type of horizontal host lines.
xlab	X-axis title.
axis.cex	Size of tick labels.
title.cex	Size of X-axis title.
...	Further graphical parameters (see details).

Details

Graphical parameters can be added by using names in the format prefix.parameter for the different parts of the plot. The parameter will then be passed on to the appropriate graphics function, e.g. arrow.lty to change the line type of the arrows. The following prefixes can be used: label for the host labels, arrow for the arrows, sample for the sampling time indicators, polygon for the generation interval distributions, line for the horizontal host lines, axis for the X-axis, and title for the X-axis title.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object containing samples.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation$sequences, times = simulation$sample.times)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

plot(MCMCstate, plot.which = "mpc")
```

---

sample.phybreak *Sampling from a phybreak MCMC-chain.*

Description

Function to take (additional) samples from the posterior distribution of a phylogenetic and transmission tree (plus associated parameters), within a phybreak-object.

Usage

```
sample.phybreak(phybreak.object, nsample, thin = 1, keepphylo = NULL,
  phylotopology_only = 0)
```

Arguments

phybreak.object	An object of class phybreak.
nsample	The number of samples to take.
thin	The thinning to use (values after every thin'th iteration will be included in the posterior). Each iteration does one update of all parameters and tree updates with each host as focal host once.
keepphylo	The proportion of tree updates keeping the phylotree intact. If there is more than one sample per host, keepphylo should be 0. If set to NULL (default), this is done automatically, otherwise it is set to 0.2.
phylotopology_only	The proportion of tree updates in which only the within-host minitree topology is sampled, and the transmission tree as well as coalescence times are kept unchanged.

Value

The phybreak-object used to call the function, including (additional) samples from the posterior.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First create a phybreak-object
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation)

MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)
```

---

sim.phybreak *Outbreak simulation.*

Description

Simulate outbreaks of class 'phybreakdata', with the outbreak model of **phybreak**.

Usage

```
sim.phybreak(obsize = 50, popsize = NA, samplesperhost = 1, R0 = 1.5,
  shape.gen = 10, mean.gen = 1, shape.sample = 10, mean.sample = 1,
  additionalsampledelay = 0, wh.model = 3, wh.slope = 1, mu = 1e-04,
  sequence.length = 10000, output.class = c("phybreakdata", "obkData"))
```

Arguments

`obsize`	The outbreak size (number of cases) to obtain. If `obsize = NA`, `popsize` should be provided.
`popsize`	The population size in which to simulate. If it is not defined (default), an optimal population size will be chosen based on R0 and obsize. Be aware that choosing a `popsize` and an `obsize` can severely increase the simulation time, depending on `R0`.
`samplesperhost`	Number of samples to be taken per host, either a vector or a single number.
`R0`	The basic reproduction ratio used for simulation. The offspring distribution is Poisson.
`shape.gen`	The shape parameter of the gamma-distributed generation interval.
`mean.gen`	The mean generation interval.
`shape.sample`	The shape parameter of the gamma-distributed sampling interval.
`mean.sample`	The mean sampling interval (for the first sample of each host).
`additionalsampledelay`	Sampling intervals since first sampling times of each host. Values in this vector will be used first for all additional samples of host 1, then of host 2, etc.
`wh.model`	The model for within-host pathogen dynamics (effective pathogen population size = N*gE = actual population size * pathogen generation time), used to simulate coalescence events. Options are: 1. Effective size = 0, so coalescence occurs 'just before' transmission, in the infector 2. Effective size = Inf, so coalescence occurs 'just after' infection, in the infectee 3. Effective size at time t after infection = wh.slope * t
`wh.slope`	Within-host increase of effective population size, used if `wh.model = 3`.
`mu`	Expected number of mutations per nucleotide per unit of time along each lineage.
`sequence.length`	Number of available nucleotides for mutations.
`output.class`	Class of the simulation output. If package **OutbreakTools** is available, it is possible to choose class `'obkData'`

Value

The simulation output, either as an object of class 'phybreakdata' with sequences (class 'phyDat') and sampling times (which would be the observations), and infection times, infectors, and phylogenetic tree of class phylo; or as an object of class 'obkData' (package **OutbreakTools**), containing the outbreak data in the following slots:

individuals a data.frame with individual labels as row names, a vector infector, and a vector date containing the infection times (starting 01-01-2000)

dna an object of class 'obkSequences', with SNP data in dna and sampling times in meta$date

trees an object of class multiphylo, containing a single tree of class phylo

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
simulation <- sim.phybreak()
```

---

thin_phybreak *Remove posterior samples from a phybreak-object.*

Description

The function removes (all or some) posterior samples by thinning and/or removing the first part of the chain, but keeps the state of variables and parameters intact.

Usage

```
thin_phybreak(x, thin = 1, nkeep = Inf)
```

Arguments

x	An object of class phybreak.
thin	The thinning interval to use.
nkeep	the number of most recent samples to keep. If nkeep = Inf (default), the whole chain is thinned and returned. Otherwise, samples from the tail are kept.

Value

The phybreak-object provided as input, but with fewer posterior samples.

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First create a phybreak-object
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation$sequences, times = simulation$sample.times)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

MCMCstate <- thin_phybreak(MCMCstate, thin = 2)
```

---

transtree *Create a consensus transmission tree.*

Description

Various methods to create summary transmission trees from a phybreak-object containing posterior samples.

Usage

```
transtree(phybreak.object, method = c("count", "edmonds", "mpc", "mtcc"),
  samplesize = Inf, infector.name = TRUE, support = c("proportion",
  "count"), infection.times = c("all", "infector", "infector.sd"),
  time.quantiles = c(0.025, 0.5, 0.975), phylo.class = FALSE)
```

Arguments

phybreak.object	An object of class phybreak.
method	The method used to create the tree (see details).
samplesize	The number of posterior samples that is used, taken from the tail.
infector.name	Whether to return the infector names, or the position in the vector of hosts.
support	Whether to return the support (= posterior probability) for each infector as a "proportion" or as a "count" of posterior trees in which that transmission link or transmission cluster is present.

`infection.times`
: Whether to base the summary infection times on `"all"` samples, or only on samples in which the `"infector"` was the same as in the consensus tree. A third option is `"infector.sd"`, in which case only posterior trees with the same infector or transmission cluster as in the consensus tree were used to calculate a mean and standard deviation. In that case and if `method = "mpc"` or `method = "mtcc"`, also the infection times in the first sampled tree with consensus tree topology are given.

`time.quantiles`
: Used only if `infection.times = "all"` or `"infector"`.

`phylo.class`
: Whether to return an object of class `"phylo"`, in which case a single tree (`"mpc"` or `"mtcc"`) from the posterior is returned (not with summary infection times). This option is used by `plotPhylo`.

Details

Four methods are supported for transmission tree reconstruction (who infected whom).

- `"count"` gives the most frequent infector from the posterior samples. Note that this may result in an improper tree (multiple roots and/or cycles). Support is measured by the frequency of the infector in the posterior distribution.
- `"edmonds"` starts from the most frequent infector (method `"count"`), multiple roots and cycles are removed by selecting one by one the next most frequent option that minimizes the loss in support (Edmonds' algorithm). Support is measured by the frequency of the infector in the posterior distribution.
- "mpc" gives the maximum parent credibility tree as described in Hall et al (2015). This is the tree in the set of posterior samples that has maximum support = product of frequencies among all posterior samples. Support is measured by the frequency of the infector in the posterior distribution.
- "mtcc" gives the maximum transmission cluster credibility tree. This is equivalent to the maximum clade credibility (mcc) phylogenetic tree, with clusters defined as host + all progeny. Support is measured by the frequency of the cluster in the posterior distribution.

Value

If `phylo.class = FALSE`, a `data.frame` with per item (=host) its infector and support per infector (or cluster), and summary infection times. If `phylo.class = TRUE`, a class `"phylo"` object, a single tree (`"mpc"` or `"mtcc"`) from the posterior is returned (not with summary infection times).

Author(s)

Don Klinkenberg <don@xs4all.nl>

References

Klinkenberg et al. (2017) Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*, **13**(5): e1005495.

Examples

```
#First build a phybreak-object containing samples.
simulation <- sim.phybreak(obsize = 5)
MCMCstate <- phybreak(data = simulation)
MCMCstate <- burnin.phybreak(MCMCstate, ncycles = 20)
MCMCstate <- sample.phybreak(MCMCstate, nsample = 50, thin = 2)

transtree(MCMCstate, method = "edmonds")
transtree(MCMCstate, method = "mpc", infection.times = "infector.sd")
plot(MCMCstate, plot.which = "mpc")
```

Index