Package 'MSCsimtester'

October 12, 2022

Type Package

Title Tests of Multispecies Coalescent Gene Tree Simulator Output

Version 1.0.0

Maintainer Elizabeth Allman <e.allman@alaska.edu>

Description Statistical tests for validating multispecies coalescent gene tree simulators, using pairwise distances and rooted triple counts. Background is given by Allman, Banos, and Rhodes (2019) <arXiv:1908.01424>.

License MIT + file LICENSE

Imports Rdpack, kSamples, graphics, stats

RdMacros Rdpack

Depends R(>= 2.10),ape(>= 5.0)

Encoding UTF-8

RoxygenNote 7.1.2

NeedsCompilation no

Author Elizabeth Allman [aut, cre, cph],
Hector Banos [aut, cph],
John Rhodes [aut, cph]

Repository CRAN

Date/Publication 2021-12-15 00:00:02 UTC

R topics documented:

ADtest *Anderson-Darling test comparing sample and theoretical pairwise distance distributions.*

Description

Takes as input theoretical pairwise distance densities under the MSC and empirical pairwise distances from gene trees in a sample, as returned by the function `pairwiseDist`. Uses the package `kSamples` to perform either one test on the entire dataset or multiple tests on subsamples.

Usage

```
ADtest(distanceDensities, subsampleSize = FALSE)
```

Arguments

`distanceDensities`	A list containing values needed for performing Anderson-Darling test(s) on a gene tree sample and species tree, as output by `pairwiseDist`. For details, see code for `pairwiseDist`.
`subsampleSize`	A positive integer to perform multiple tests on subsamples, or `FALSE` (default) to perform one test on full sample.

Details

The Anderson-Darling test compares the empirical distance distribution for a supplied gene tree sample to a sample drawn from the theoretical distribution. The output, passed from the `kSamples` package, thus says that 2 samples are being compared, to test a null-hypothesis that they come from the same distribution. See `kSamples` documentation for function `ad.test` for more details.

Repeated runs of this function will give different results, since the sample from the theoretical distribution will vary. Under the null hypothesis p-values for different runs should be approximately uniformly distributed.

Numerical issues may result in poor performance of Anderson-Darling tests when the sample size is very large, so an optional parameter `subsampleSize` can be set to create subsamples of smaller size. If `subsampleSize` is a positive integer, Anderson-Darling tests are performed on each subset, comparing them to a random sample of the same size from the theoretical distribution. Good fit is indicated by an approximately uniform distribution of the subsample p-values.

Value

An object of type `ADtestOutput` including a sample `$Sample` from the theoretical distance distribution of the same size as the empirical one, and `$ADtest` which is of type `kSamples` and has all output from the Anderson-Darling test if only one test was performed, or the number of tests if tests were performed on subsamples.

See Also

`pairwiseDist`, `kSamples-package`

Examples

```
stree=read.tree(text="((((a:10000,b:10000):10000,c:20000):10000,d:30000):10000,e:40000);")
pops=c(15000,25000,10000,1,1,1,1,1,12000)
gts=read.tree(file=system.file("extdata","genetreeSample",package="MSCsimtester"))
distDen=pairwiseDist(stree,pops,gts,"a","b")
ADtest(distDen)
ADtest(distDen,1000)
```

`genetreeSample` *Simulated gene tree dataset.*

Description

A dataset of 10,000 gene trees on 5 taxa simulated under the MSC on a species tree.

Format

A text file with 10,000 metric Newick gene trees on the taxa a,b,c,d,e

Details

This simulated dataset was produced by SimPhy (Mallo et al. 2016), using the species tree

((((a:10000,b:10000):10000,c:20000):10000,d:30000):10000,e:40000);

with population sizes

c(15000,25000,10000,1,1,1,1,1,12000)

and edges ordered by the ape function `read.tree`.

File is accessed as `system.file("extdata","genetreeSample",package="MSCsimtester")`, for example using the ape command:

```
gts=read.tree(file=system.file("extdata","genetreeSample",package="MSCsimtester")
)
```

References

Mallo D, De Oliveira Martins L, Posada D (2016). "SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees." *Syst. Biol.*, **65**(2), 334-344. doi: 10.1093/sysbio/syv082, http://dx.doi.org/10.1093/sysbio/syv082.

`MSCsimtester` *Validity tests of simulators of the multspecies coalescent model in phylogenomics.*

Description

The package performs comparisons of certain summary statistics for simulated gene tree samples to theoretical predictions under the multispecies coalescent model. The primary functions are `rootedTriple` for comparison of frequencies of topological rooted triples on gene trees, and `pairwiseDist` and `ADtest` for comparison of the distributions of pairwise distances between taxa on gene trees.

Details

Required input is a collection of gene trees, stored as a `multiPhylo` object by the `ape` package, and a rooted species tree, as a `Phylo` object, with edge lengths in generations, together with constant population sizes for each edge.

`MSCsimtester` builds on the packages `ape` and `kSamples`.

For further examples of use and citation purposes, see (Allman et al. 2019).

References

Allman ES, Baños H, Rhodes JA (2019). "Testing Multispecies Coalescent Simulators Using Summary Statistics." *arXiv:1908.01424*.

`pairwiseDist` *Compute and plot sample and theoretical pairwise distance densities.*

Description

Computes theoretical pairwise distance densities under the MSC on a species tree and empirical pairwise distances from gene trees in a sample. A histogram of empirical values is plotted over the theoretical pdf.

Usage

```
pairwiseDist(
  stree,
  popSizes,
  gtSample,
  taxon1,
  taxon2,
  numSteps = 1000,
  tailProb = 0.01
)
```

Arguments

stree	An object of class phylo containing a rooted metric species tree. Edge lengths are in number of generations.
popSizes	A vector containing constant population sizes, one entry for each edge/population in the species tree, for a haploid population. Sizes should be doubled for diploids. If stree has k edges, then popSizes must have k+1 elements, with final entry the size of the population ancestral to the root.
gtSample	An object of class multiPhylo holding a sample of gene trees from a simulation. Taxon labels on gene trees must be identical to those on stree.
taxon1	A string specifying one taxon on stree.
taxon2	A string specifying a second taxon on stree, distinct from taxon1.
numSteps	A positive integer number of values to be computed for graphing the theoretical pairwise distance density. Default is numSteps = 1000. A larger value produces a smoother plot.
tailProb	A cutoff value, between 0 and 1, for the theoretical density, with a default of 0.01. The theoretical pairwise distance is plotted from (0, xMax), where xMax is the larger of the maximum pairwise distance in the gene tree sample and the value cutting off a tail of area tailProb under the pdf. A message returns the proportion of sample distances in this tail.

Value

A list of items needed for Anderson-Darling test(s), for use by ADtest, returned invisibly. See function code for more details.

See Also

plotEdgeOrder, plotPops, ADtest

Examples

```
stree=read.tree(text="((((a:10000,b:10000):10000,c:20000):10000,d:30000):10000,e:40000);")
pops=c(15000,25000,10000,1,1,1,1,1,12000)
gts=read.tree(file=system.file("extdata","genetreeSample",package="MSCsimtester"))
pairwiseDist(stree,pops,gts,"a","b")
```

---

plotEdgeOrder *Plot species tree, with edge numbers on edges.*

Description

Under the MSC, each edge in the species tree must be assigned a population size. This function displays the species tree with the edges numbered, to aid the user in entering constant population sizes as an appropriately ordered list.

Usage

```
plotEdgeOrder(stree)
```

Arguments

stree	An object of class phylo containing a rooted metric species tree.

Value

NONE

See Also

pairwiseDist, rootedTriple, plotPops

Examples

```
stree=read.tree(text="(((a:10000,b:10000):10000,c:20000):10000,d:30000);")
plotEdgeOrder(stree)
pops=c(30000,20000,1,1,1,1,10000)
plotPops(stree,pops)
```

---

plotPops *Plot species tree, with population sizes on edges.*

Description

Plot species tree, with population sizes on edges.

Usage

```
plotPops(stree, populations)
```

Arguments

stree	An object of class phylo containing a rooted metric species tree.
populations	A vector containing constant population sizes, one entry for each edge/population in the species tree, with last entry for the population ancestral to the root.

Value

NONE

See Also

pairwiseDist, rootedTriple, plotEdgeOrder

Examples

```
stree=read.tree(text="(((a:10000,b:10000):10000,c:20000):10000,d:30000);")
plotEdgeOrder(stree)
pops=c(30000,20000,1,1,1,1,10000)
plotPops(stree,pops)
```

`print.ADtestOutput` *Print function for objects of class* `ADtestOutput`.

Description

Print function for objects of class `ADtestOutput`.

Usage

```
## S3 method for class 'ADtestOutput'
print(x, ...)
```

Arguments

`x`	An object of class `ADtestOutput`, as produced by `ADtest` function.
`...`	Further arguments to be passed to or from other methods.

Value

NONE

`print.rootedTripleOutput` *Print function for objects of class* `rootedTripleOutput`.

Description

Print function for objects of class `rootedTripleOutput`.

Usage

```
## S3 method for class 'rootedTripleOutput'
print(x, ...)
```

Arguments

`x`	An object of class `rootedTripleOutput`, as produced by `rootedTriple` function.
`...`	Further arguments to be passed to or from other methods.

Details

Print function for objects of class `rootedTripleOutput`.

Value

NONE

`rootedTriple` *Compare expected and sample frequencies of topological rooted triples.*

Description

For a given species tree with population sizes, compares the expected frequencies of rooted triples to empirical frequencies in a sample of gene trees, using Chi-squared tests with 2 d.f. The exact and estimated internal branch length (in coalescent units) of the rooted triple in the species tree are also computed for comparison. A single test can be performed on the entire gene tree sample, or multiple tests on subsamples.

Usage

```
rootedTriple(
  stree,
  popSizes,
  gtSample,
  taxon1,
  taxon2,
  taxon3,
  subsampleSize = FALSE
)
```

Arguments

`stree`	An object of class `phylo` containing a rooted metric species tree. Edge lengths are in number of generations.
`popSizes`	An ordered list containing constant population sizes for each species tree edge, for a haploid organism. Sizes should be doubled for diploids. If `stree` has k edges, then `popSizes` must have k+1 elements, with the final entry for the population ancestral to the root.
`gtSample`	An object of class `multiPhylo` holding a sample of gene trees from a simulation. Taxon labels on gene trees must be identical to those on `stree`.
`taxon1`	A string specifying one taxon on `stree`.
`taxon2`	A string specifying a second taxon on `stree`, distinct from `taxon1`.
`taxon3`	A string specifying a third taxon on `stree`, distinct from `taxon1`, `taxon2`.
`subsampleSize`	A positive integer or `FALSE`, giving size of subsamples of `gtSample` to analyze.

Details

When subsampleSize is FALSE the Chi-squared test is performed using all gene trees in gtSample. Results are reported in tabular form in the console.

When subsampleSize is positive, the N trees in gtSample will be partitioned into N/subsampleSize subsamples, with a Chi-squared test performed for each. Histograms are plotted for (1) the p-values for the Chi-squared tests on subsamples, and (2) subsample estimates of the internal branch length for the rooted triple on the species tree, with the true value marked.

Three distinct taxon names must be supplied, all of which must occur on stree and in each of the gene trees in the sample.

Value

If subsampleSize is FALSE, returns an object of type rootedTripleOutput which contains a table $TripletCounts of empirical and expected rooted triple counts, a p-value $pv from the Chi-squared test, and a column $InternalEdge of estimated and exact internal edge lengths. If subsampleSize is TRUE, returns NULL but produces several plots.

See Also

plotEdgeOrder, plotPops

Examples

```
stree=read.tree(text="((((a:10000,b:10000):10000,c:20000):10000,d:30000):10000,e:40000);")
pops=c(15000,25000,10000,1,1,1,1,1,12000)
gts=read.tree(file=system.file("extdata","genetreeSample",package="MSCsimtester"))
rootedTriple(stree,pops,gts,"a","b","c")
rootedTriple(stree,pops,gts,"a","b","c",1000)
```

Index