Package 'CoalescentPhylo'

May 27, 2026

Title Phylogenetics via Root Distances Method Under the Coalescent

Version 0.1.0

Description Estimates phylogenetic trees from allele count data using the root distance method under the Coalescent Model. Given a matrix of allele counts across taxa and loci, the package estimates pairwise root distances under the Coalescent Model using maximum likelihood estimation. Then, it estimates a labeled phylogenetic tree from the estimated root distances. See Peng et al. (2021) <doi:10.1016/j.ympev.2021.107142>.

License AGPL-3

URL https://github.com/ArindamRoyChoudhury/CoalescentPhylo

BugReports https://github.com/ArindamRoyChoudhury/CoalescentPhylo/issues

Depends R (>= 3.5.0)

Imports ape, parallel, pbapply, phangorn, rootSolve

Encoding UTF-8

LazyData true

RoxygenNote 7.3.3

NeedsCompilation no

Author Arindam RoyChoudhury [aut, cre, cph],
Ying Li [aut]

Maintainer Arindam RoyChoudhury <arr2014@med.cornell.edu>

Repository CRAN

Date/Publication 2026-05-27 09:50:02 UTC

Contents

Human_Allele_Count_Data *Example allele count data for CoalescentPhylo analyses:*

Description

We provide a matrix of allele counts used to demonstrate the estimation of phylogenetic tree with the `CoalescentPhylo` package. The dataset contains allele counts from 8 ingroup human populations and 1 outgroup human population (San), measured across 2,000 loci.

Usage

```
data(Human_Allele_Count_Data)
```

Format

A numeric matrix with 9 rows and 2,000 columns, where:

Rows Populations (8 ingroup + 1 outgroup). Row names correspond to population labels from the ALFRED database, with ALFRED sample IDs given in parentheses:

- Papuan New Guinean (SA001501H)
- Uyghur (SA001492Q)
- Hazara (SA001477T)
- Yi (SA001485S)
- Dai (SA001493R)
- Japanese (SA002260K)
- Mongolian (SA001489W)
- Karitiana (SA001514L)
- San (SA001469U) — outgroup

Columns Loci (2,000 SNP sites). Each entry is a non-negative integer allele count not exceeding `fixed.n.at.tips`.

Details

The populations were selected to represent diverse geographic regions across Africa, Central Asia, East Asia, Oceania, and the Americas, providing a broad test case for coalescent-based phylogenetic inference. San is included as the outgroup, consistent with the early divergence of Southern African populations in human evolutionary history.

Source

ALFRED — The ALlele FREquency Database (https://alfred.med.yale.edu). A resource of gene frequency data on human populations supported by Biomedical Informatics and Data Science, Yale University.

Examples

```
data(Human_Allele_Count_Data)

# Check dimensions: 9 populations x 2000 loci
dim(Human_Allele_Count_Data)

# View population names
rownames(Human_Allele_Count_Data)


# Estimate phylogenetic tree
tree <- RD(mat_allele_count = Human_Allele_Count_Data, n.cores = 1)
plot(tree$labeled_tree)
```

RD *Estimates a phylogenetic tree from allele count data using The Coalescent Model*

Description

Estimates a phylogenetic tree from a matrix of allele counts using root distance method under The Coalescent Model.

Usage

```
RD(
  mat_allele_count,
  theta = 1,
  fixed.n.at.tips = 4,
  newick_br_length_digits = 3,
  n.cores = NULL
)
```

Arguments

`mat_allele_count`
: A numeric matrix of allele counts where rows represent populations or samples and columns represent alleles or loci. Input as allele counts, each should be non-negative integer $\leq$ `fixed.n.at.tips`.

`theta`
: A positive numeric scalar representing the scaled mutation rate ($4N\mu$). Used to scale genetic distances when estimating branch lengths. Defaults to 1.

`fixed.n.at.tips`
: A positive integer specifying the fixed sample size assumed at each tip of the tree. Used to correct for sampling effects when computing distances. Defaults to 4.

`newick_br_length_digits`
: A non-negative integer controlling the number of decimal places used when formatting branch lengths in the Newick string output. Defaults to `3`.

`n.cores`
: Optional integer specifying the number of CPU cores used for parallel computation. Defaults to `detectCores() - 1`.

Value

A named list with two elements:

unlabeled_tree A phylogenetic tree of class `"phylo"` with estimated branch lengths but no tip labels assigned.

labeled_tree A phylogenetic tree of class `"phylo"` with estimated branch lengths and tip labels derived from the row names of `mat_allele_count`.

References

Peng J, Rajeevan H, Kubatko L, RoyChoudhury A (2021). A fast likelihood approach for estimation of large phylogenies from continuous trait data. *Molecular Phylogenetics and Evolution*, **161**, 107142. doi:10.1016/j.ympev.2021.107142

Examples

```
# Load built-in example dataset (9 taxa x 2000 loci)
data(Human_Allele_Count_Data)

# Inspect dimensions: rows = taxa, columns = loci
dim(Human_Allele_Count_Data)

# Preview first few rows and columns
Human_Allele_Count_Data[1:3, 1:5]

# Check for missing data
anyNA(Human_Allele_Count_Data)


# NOTE: For CRAN testing, we use n.cores = 1 for compatibility.
# In practice, users may set n.cores = NULL to use all available cores
# and speed up computation.

# Estimate phylogenetic tree using the Coalescent Model
tree <- RD(mat_allele_count = Human_Allele_Count_Data, n.cores = 1)

# Summarize the result
print(tree)

# Plot the labeled phylogenetic tree
plot(tree$labeled_tree)
```

Index