Package 'DiscML'

July 26, 2014

Type Package

Title DiscML: An R package for estimating evolutionary rates of discrete characters using maximum likelihood

Version 1.0.1

Date 2014-04-28

Author Tane Kim, Weilong Hao

Maintainer Weilong Hao <haow@wayne.edu>

Description DiscML performs rate estimation using maximum likelihood with the options to correct for unobservable data, to implement a Gamma-distribution for rate variation, and to estimate the prior root probabilities from the empirical data.

Depends ape

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2014-07-26 08:27:41

R topics documented:

DiscML-package 2

Index **6**

DiscML-package *DiscML: An R package for estimating evolutionary rates of discrete characters using maximum likelihood*

Description

DiscML was developed as a unified R program for estimating the evolutionary rates of discrete characters with no restriction on the number of character states and having a great flexibility on transition models. DiscML performs maximum likelihood estimation with the options to correct for unobservable data, to implement a gamma distribution for rate variation, and to estimate the prior root probabilities from the empirical data. It gives users the ability to customize the instantaneous rate transition matrices, and to choose a variety of pre-determined matrices. DiscML is ideal for the analysis of binary (1s/0s) patterns, gene families, and multistate discrete morphological characteristics.

Arguments

x	a vector, a matrix (in which each row is a vector for a gene family), a data frame (in which each row is a vector for a gene family).
phy	phylogenetic information; an object of class "phylo" in "ape".
CI	a logical specifying whether to return the 95% confidence intervals the likelihood of the different states.
model	a customized numeric matrix, or one of the pre-determined ones, "ER", "ARD", "SYM", "BDER", "BDARD","BDSYM", "BDBI", "BDIER", "BDIARD", "BDISYM"
ip	the initial values of the entries in the rate matrix used in the maximum likelihood estimation.
alpha	a logical specifying whether to estimate the alpha parameter of a discrete gamma distribution in the maximum likelihood estimation, or a numeric value specifying an alpha value.
ialpha	the initial alpha value in a gamma distribution used in the maximum likelihood estimation.
rootprobability	a logical specifying whether to estimate the prior root probabilities in the maximum likelihood estimation.
irootprobability	an initial numeric vector for the prior root probabilities optimization. Else it will automatically set to an even numeric vector.
zerocorrection	a logical specifying whether to correct for unobservable data; see details in Felsenstein (1992).
simplify	a logical specifying whether to convert all nonzero character states to character state "1" and perform binary analysis.
individualrates	a logical specifying whether to calculate likelihoods of gene families individually, rather than multiplying them together, given that there are more than one gene family.

characters	a logical specifying whether to automatically find minimum number of character states from given data, or a non-negative integer vector to manually specify all possible character states.
plotloglik	a logical specifying whether to print the plot of 'Log likelihoods Vs Gene families', when 'inividualrates = TRUE'.
plotmu	a logical specifying whether to print the plot of 'Mu vs Gene families', when 'individualrates = TRUE', or a string specifying wich mu's to be plotted.

Details

Package:	DiscML
Type:	Package
Version:	1.0
Date:	2014-04-28
License:	GPL (>= 2)

DiscML is flexible on both the size and type of the rate transition matrix. The argument, 'model', specifies the rate transition matrix, which can be customized by the user or chosen from pre-determined matrices. The pre-determined matrices in DiscML are:

1. 'ER': an equal-rate matrix, in which all non-diagonal entries are equal, e.g., matrix(c(0,1,1,1,0,1,1,1,0),ncol = 3, nrow = 3).

2. 'SYM': a symmetric matrix, which is identical with its transpose, e.g., matrix(c(0,1,2,1,0,3,2,3,0), ncol= 3, nrow = 3).

3. 'ARD': an all-rates-different matrix, in which all non-diagonal entries are free to vary, e.g., matrix(c(0,1,2,3,0,4,5,6,0), ncol = 3, nrow = 3).

4. 'GTR': a general time reversible matrix, which is the same as a combination of arguments, model ="SYM", reversible = TRUE, and rootprobability = TRUE.

5. 'BDER': a birth-and-death matrix, in which all non-zero entries are equal, e.g., matrix(c(0,1,0,0,1,0,1,0,0,1,0,1,0,0,1,0), ncol= 4, nrow = 4)

6. 'BDSYM': a birth-and-death matrix with symmetric entries, e.g., matrix(c(0,1,0,0,1,0,2,0,0,2,0,3,0,0,3,0), ncol= 4, nrow = 4).

7. 'BDARD': a birth-and-death matrix with all non-zero entries free to vary, e.g., matrix(c(0,1,0,0,2,0,3,0,0,4,0,5,0,0,6,0), ncol= 4, nrow = 4).

8. 'BDIER': a birth-death-and-innovation matrix with equal entries, ultimately the same as 'BDER'.

8. 'BDISYM': a birth-death-and-innovation matrix with symmetric entries, e.g., matrix(c(0,1,0,0,1,0,2,0,0,2,0,2,0,0,2,0), ncol= 4, nrow = 4).

9. 'BDIARD': a birth-death-and-innovation matrix with variable entries, e.g., matrix(c(0,1,0,0,3,0,2,0,0,4,0,2,0,0,4,0), ncol= 4, nrow = 4).

10. 'BDBI' : a birth-and-death matrix with birth entries and death entries being equal respectively. e.g., matrix(c(0,1,0,0,2,0,1,0,0,2,0,1,0,0,2,0), ncol= 4, nrow = 4).

When the prior root probabilities are estimated, the argument 'reversible' allows reversibility for the pre-determined symmetric matrices, namely, ER, SYM, BDER, BDIER, BDSYM, and BDISYM, by multiplying the entries by the corresponding root probabilities.

The user can customize entries in the rate matrix by assigning a matrix to the argument 'model'. The diagonal of the matrix will be ignored, and the remaining entries are assigned non-negative integers. The entry with the number '0' will be translated into a value of 0 in the rate matrix, non-negative integers in the entries represent rate index. For example, setting model = matrix(c(NA,1,0,0,1,NA,2,0,0,2,NA,3,0,0,3,NA) ,nrow= 4, ncol= 4), will be as same as setting model= "BDSYM".

Moreover, DiscML considers rate variation among the character sites by implementing a discrete gamma distribution using alpha = TRUE.

Author(s)

Tane Kim, Weilong Hao

Maintainer: Weilong Hao <haow@wayne.edu>

References

Felsenstein, J. (1992). Phylogenies from restriction sites: A maximum-likelihood approach. Evolution, 46, 159–173.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics, 20, 289–290.

Schliep K.P. (2011) phangorn: Phylogenetic analysis in R. Bioinformatics, 27(4) 592-593.

Yang, Z. (1994), Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol, 39, 306–314.

Examples

```
# The default arguments in DiscML are:
# model = "ER"
# reversible = FALSE
# alpha = FALSE
# rootprobability = FALSE

x<- c(1,2,1,0,1)
phy <- rtree(length(x))

# x is a vector with 5 elements, and phy is a randomly generated
# 5-taxon tree (using rtree from the 'ape' package).
DiscML(x, phy)

# x here is a matrix
x <- matrix(c(1,2,0,2,1,0),2,3, byrow = TRUE)

# phy is a randomly generated tree containing 3 tips.
phy <- rtree(3)

# a symmetric rate transition matrix is used in the estimation.
DiscML(x, phy, model = "SYM")

# the prior root probabilities will be estimated.
```

```
DiscML(x, phy, rootprobability = TRUE)

# the prior root probabilities are fixed to be 1/16, 5/16, and 10/16.
DiscML(x, phy, rootprobability = c(1/16,5/16,10/16))

# the alpha value in a gamma distribution will be estimated.
DiscML(x, phy, alpha = TRUE)

# DiscML allows the reversibility for the symmetric matrices, e.g., ER, SYM..
DiscML(x, phy, rootprobability = TRUE, reversible = TRUE)

# DiscML can convert all non-zero character states to be '1's to perform
# binary analysis.
DiscML(x, phy, simplify = TRUE)

# DiscML can compute for each gene family of 'x' individually,
DiscML(x, phy, individualrates = TRUE)

# this is equivalent to:
# phy <- "(A$mu2:0.1,(B$mu0:0.3,C$mu1:0.4)$mu2:0.5);"
phy <- "(A$mu2:0.1,(B:0.3,C$mu1:0.4)$mu2:0.5);"

# DiscML can optimize different mu's for each branches.
DiscML(x, phy)

# DiscML can plot each mu vs Gene families when individualrates =TRUE.
DiscML(x, phy, individualrates = TRUE, plotmu = TRUE)
```

Index

DiscML *(*DiscML-package*)*, 2
DiscML-package, 2

print.DiscML *(*DiscML-package*)*, 2

read.tree2 *(*DiscML-package*)*, 2