Package 'TKF'

April 18, 2015

Version 0.0.8

Date 2015-04-17

Title Pairwise Distance Estimation with TKF91 and TKF92 Model

Description Pairwise evolutionary distance estimation between protein sequences with the TKF91 and TKF92 model, which consider all the possible paths of transforming from one sequence to another.

Author Ge Tan <ge.tan09@imperial.ac.uk>

Maintainer Ge Tan <ge.tan09@imperial.ac.uk>

Imports methods, expm, numDeriv, ape (>= 3.2), phytools (>= 0.4-45), phangorn (>= 1.99-12)

Depends R (>= 3.0.2)

Suggests RUnit, seqinr

SystemRequirements gsl

License GPL-2

Type Package

NeedsCompilation yes

LazyData yes

Repository CRAN

Date/Publication 2015-04-18 14:29:04

R topics documented:

AAToInt *AA, DNA, RNA character set*

Description

Some AA, DNA, RNA character set defined in this package and functions to convert them into integers.

Usage

```
AAToInt(AA)
```

Arguments

AA	A vector of character.

Details

Each AA is converted to the position of that AA in AACharacterSet.

Value

A integer vector.

Author(s)

Ge Tan

Examples

```
library(seqinr)
fasta <- read.fasta(file.path(system.file("extdata", package="TKF"),
                    "pair1.fasta"),
                    seqtype="AA", set.attributes=FALSE)
AAToInt(fasta[[1]])

AACharacterSet
```

GONNET *The GONNET AA matrix*

Description

The GONNET mutation matrix and background frequency.

Usage

```
data(GONNET)
data(GONNETBF)
```

Format

A 20*20 numeric matrix.

A 20 numeric vector.

Source

Gonnet, G.H., Cohen, M.A., and Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. Science 256, 1443-1445.

Examples

```
data(GONNET)
data(GONNETBF)
```

optim.phylo.wls *Phylogeny inference using the weighted least squares method*

Description

This function performs phylogeny inference using weighted least-squares.

Usage

```
optim.phylo.wls(Dist, Var=NULL, stree=NULL, set.neg.to.zero=TRUE,
                fixed=FALSE, tol=1e-10, collapse=TRUE)
```

Arguments

Dist	a distance matrix.
Var	a covariance matrix of the distance. When it is NULL, ordinary least squares tree will be built.
stree	an optional starting tree for the optimization.
set.neg.to.zero	a logical value indicating whether to set negative branch lengths to zero (default "TRUE").
fixed	a logical value indicating whether to estimate the topology - if "TRUE" only the branch lengths will be computed.
tol	a tolerance value used to assess whether the optimization has converged.
collapse	a logical indicating whether to collapse branches with zero length.

Details

This function extends the function `optim.phylo.ls` in package `phytools` to support weighted least squares tree reconstruction. For more details, please check the help page of `optim.phylo.ls`.

Value

An objec of class "phylo" that (may be) the least-squares tree with branch lengths; also returns the sum of squares in 'attr(tree,"Q-score")'.

Author(s)

Ge Tan

Examples

```
Dist <- matrix(c(0.00000, 27.78202, 29.54125, 29.06183, 40.63082, 41.20910,
                 27.78202, 0.00000, 14.82329, 24.26988, 47.40101, 43.76202,
                 29.54125, 14.82329, 0.00000, 26.82772, 48.17819, 41.27872,
                 29.06183, 24.26988, 26.82772, 0.00000, 44.66941, 44.39078,
                 40.63082, 47.40101, 48.17819, 44.66941, 0.00000, 45.63394,
                 41.20910, 43.76202, 41.27872, 44.39078, 45.63394, 0.00000),
                 ncol=6, dimnames=list(c("YARLI", "KLULA", "CANGA", "DEBHA",
                 "CRYNE", "ASPFU"), c("YARLI", "KLULA", "CANGA", "DEBHA",
                 "CRYNE", "ASPFU")))
Var <- matrix(c(0.000000, 6.261368, 6.816608, 6.660132, 11.361800, 12.037978,
                6.261368, 0.000000, 2.877505, 5.054447, 14.315551, 12.734813,
                6.816608, 2.877505, 0.000000, 5.699967, 12.638321, 11.598558,
                6.660132, 5.054447, 5.699967, 0.000000, 12.189609, 13.185733,
                11.36180, 14.31555, 12.63832, 12.18961, 0.00000, 15.19872,
                12.03798, 12.73481, 11.59856, 13.18573, 15.19872, 0.00000),
                ncol=6, dimnames=list(c("YARLI", "KLULA", "CANGA", "DEBHA",
                 "CRYNE", "ASPFU"), c("YARLI", "KLULA", "CANGA", "DEBHA",
                 "CRYNE", "ASPFU")))
tree <- optim.phylo.wls(Dist, Var)
plot(tree, type="unrooted")
```

PAMn *PAM and Dayhoff matrices calculation*

Description

These functions calculate the mutation matrix or Dayhoff matrix from the mutation matrix at PAM 1 and base backrgound frequency.

Usage

```
PAMn(PAM1, n)
Dayhoffn(PAM1, BF, n)
```

Arguments

PAM1	A matrix of numeric: the mutation probability from one AA to another AA at PAM distance 1. The order of AA in the matrix should be identical to AACharacterSet.
n	A numeric: the PAM distance.
BF	A numeric vector: the backrgound frequency of AAs. The order of AA in the vector should also be identical to AACharacterSet.

Details

Calculate the n-PAM matrices from PAM1 mutation matrix and n. To compute n-PAM matrices, we multiply the PAM1 matrix through itself N times, which is most efficiently achieved through n additions in log space.

Computing Dayhoff matrices from PAM mutation matrices and AA frequency. Dayhoff matrices are the ratios P("alignment i and j arose through evolution") / P("alignment i and j arose by chance")

Value

A numeric matrix is returned.

Author(s)

Ge Tan

References

Dayhoff, M.O., and Schwartz, R.M. (1978). A model of evolutionary change in proteins. In In Atlas of Protein Sequence and Structure,.

Gonnet, G.H., and Scholl, R. (2009). Scientific Computation (Cambridge, UK; New York: Cambridge University Press).

Examples

```
data(GONNET)
data(GONNETBF)
## PAM 250 mutation matrix
PAM250 <- PAMn(GONNET, 250)

## Dayhoff 250 matrix
Dayhoff250 <- Dayhoffn(GONNET, GONNETBF, 250)
```

TKF91 *Evolutionary distance estimation with TKF91 model*

Description

This function implements the TKF91 model to estimate the pairwise distance from protein sequences.

Usage

```
TKF91(fasta, mu=NULL, expectedLength=362,
      substModel, substModelBF)
TKF91Pair(seq1, seq2, mu=NULL, distance=NULL,
          expectedLength=362, substModel, substModelBF)
```

Arguments

`fasta`	A named list of sequences in vector of characters format. `read.fasta` from package `seqinr` outputs this format when reading from a fasta file.
`mu`	A numeric value or NULL. It is the death rate per normal link in TKF91 model. When it is NULL, a joint estimation of `mu` and `distance` will be done. When it is given, only the distance will be estimated.
`distance`	A numeric value: the PAM distance between two protein sequences. When it is given, TKF91Pair only calculates the negative log-likelihood.
`expectedLength`	A numeric object: the expected length of input protein sequences. By default, the average sequence length, 362, from OMA browser is used.
`substModel`	A numeric matrix: the mutation probability from one AA to another AA at PAM distance 1. The order of AA in the matrix should be identical to AACharacterSet.
`substModelBF`	A vector of numeric: the backrgound frequency of AAs. The order of AA in the vector should also be identical to AACharacterSet.
`seq1, seq2`	A vector of character: the sequences of two proteins to compare.

Details

Currently this implementation only supports the normal 20 AAs. Missing or Ambiguous characters are not supported.

Value

A list of matrices are returned: the matrix of estimated distances, the matrix of estimated distance variances, the matrix of negative log-likelihood between the sequences.

Author(s)

Ge Tan

References

Thorne, J.L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33, 114-124.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. Science 256, 1443-1445.

See Also

AACharacterSet, GONNET, GONNETBF

Examples

```
  ## This example is not tested due to running time > 5s
data(GONNET)
data(GONNETBF)
library(seqinr)
fasta <- read.fasta(file.path(system.file("extdata", package="TKF"),
                   "pair1.fasta"),
                   seqtype="AA", set.attributes=FALSE)
## 1D estimation: only distance
TKF91(fasta, mu=5.920655e-04,
      substModel=GONNET, substModelBF=GONNETBF)
## 2D estimation: joint estimation of distance and mu
TKF91(fasta, substModel=GONNET, substModelBF=GONNETBF)
## only apply to a pair of sequences
seq1 <- fasta[[1]]
seq2 <- fasta[[2]]
TKF91Pair(seq1, seq2, mu=5.920655e-04,
          substModel=GONNET, substModelBF=GONNETBF)
```

TKF92 *Evolutionary distance estimation with TKF92 model*

Description

This function implements the TKF92 model to estimate the pairwise distance from protein sequences.

Usage

```
TKF92(fasta, mu=NULL, r=NULL, expectedLength=362,
      substModel, substModelBF)
TKF92Pair(seq1, seq2, mu=NULL, r=NULL, distance=NULL,
          expectedLength=362, substModel, substModelBF)
```

Arguments

`fasta`	A named list of sequences in vector of characters format. `read.fasta` from package `seqinr` outputs this format when reading from a fasta file.
`mu`	A numeric value between 0 and 1 or NULL. It is the death rate per normal link in TKF92 model. When it is NULL, a joint estimation of `mu`, `r` and `distance` will be done. When it is given, only the distance will be estimated.
`r`	A numeric value between 0 and 1 or NULL. It is the success probability of the geometric distribution for modeling the fragment length in TKF92 model. When it is NULL, a joint estimation of `mu`, `r` and `distance` will be done. When it is given, only the distance will be estimated.
`distance`	A numeric value: the PAM distance between two protein sequences. When it is given, TKF92Pair only calculates the negative log-likelihood.
`expectedLength`	A numeric object: the expected length of input protein sequences. By default, the average sequence length, 362, from OMA browser is used.
`substModel`	A numeric matrix: the mutation probability from one AA to another AA at PAM distance 1. The order of AA in the matrix should be identical to AACharacterSet.
`substModelBF`	A vector of numeric: the backrgound frequency of AAs. The order of AA in the vector should also be identical to AACharacterSet.
`seq1, seq2`	A vector of character: the sequences of two proteins to compare.

Details

Currently this implementation only supports the normal 20 AAs. Missing or Ambiguous characters are not supported.

Value

A list of matrices are returned: the matrix of estimated distances, the matrix of estimated distance variances, the matrix of negative log-likelihood between the sequences.

Author(s)

Ge Tan

References

Thorne, J.L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34, 3-16.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. Science 256, 1443-1445.

See Also

AACharacterSet, GONNET, GONNETBF

Examples

```
  ## This example is not tested due to running time > 5s
data(GONNET)
data(GONNETBF)
library(seqinr)
fasta <- read.fasta(file.path(system.file("extdata", package="TKF"),
                    "pair1.fasta"),
                    seqtype="AA", set.attributes=FALSE)
## 1D estimation: only distance
TKF92(fasta, mu=0.0006137344, r=0.7016089061,
      substModel=GONNET, substModelBF=GONNETBF)

## 2D estimation: joint estimation of distance, mu and r
TKF92(fasta, substModel=GONNET, substModelBF=GONNETBF)

## only apply to a pair of sequences
seq1 <- fasta[[1]]
seq2 <- fasta[[2]]
TKF92Pair(seq1, seq2, mu=0.0006137344, r=0.7016089061,
          substModel=GONNET, substModelBF=GONNETBF)
```

TKF92HG *Evolutionary distance estimation with TKF92 model considering the regional heterogeneity of substitution rates*

Description

This function implements the TKF92 model to estimate the pairwise distance from protein sequences. An additional simple model of regional heterogeneity of substitution rates is used.

Usage

```
TKF92HG(fasta, mu=NULL, r=NULL, Ps=NULL, Kf=NULL, expectedLength=362,
        substModel, substModelBF)
TKF92HGPair(seq1, seq2, mu=NULL, r=NULL, Ps=NULL, Kf=NULL, distance=NULL,
            expectedLength=362, substModel, substModelBF)
```

Arguments

fasta	A named list of sequences in vector of characters format. read.fasta from package seqinr outputs this format when reading from a fasta file.

mu	A numeric value between 0 and 1 or NULL. It is the death rate per normal link in TKF92 model. When it is NULL, a joint estimation of mu, r, Ps, Kf and distance will be done. When it is given, only the distance will be estimated.
r	A numeric value between 0 and 1 or NULL. It is the success probability of the geometric distribution for modeling the fragment length in TKF92 model. When it is NULL, a joint estimation of mu, r, Ps, Kf and distance will be done. When it is given, only the distance will be estimated.
Ps	A numeric value between 0 and 1 or NULL. It is the equilibrium frequency of slow fragments. Hence the equilibrium frequency of fast fragements is Pf= 1-Ps. When it is NULL, a joint estimation of mu, r, Ps, Kf and distance will be done. When it is given, only the distance will be estimated.
Kf	A numeric value larger than 1 or NULL. It is the ratio of substitutions rates between fast fragments and slow fragments. When it is NULL, a joint estimation of mu, r, Ps, Kf and distance will be done. When it is given, only the distance will be estimated.
distance	A numeric value: the PAM distance between two protein sequences. When it is given, TKF92HGPair only calculates the negative log-likelihood.
expectedLength	A numeric object: the expected length of input protein sequences. By default, the average sequence length, 362, from OMA browser is used.
substModel	A numeric matrix: the mutation probability from one AA to another AA at PAM distance 1. The order of AA in the matrix should be identical to AACharacterSet.
substModelBF	A vector of numeric: the backrgound frequency of AAs. The order of AA in the vector should also be identical to AACharacterSet.
seq1, seq2	A vector of character: the sequences of two proteins to compare.

Details

Currently this implementation only supports the normal 20 AAs. Missing or Ambiguous characters are not supported.

This is a very simple model of substitution rate heterogeneity. This model assumes that there are only two varieties of fragments: one with relatively fast substitution rates and the other with slow substitution rates. This model also assumes the fragment size distribution of these two fragments is identical.

Value

A list of matrices are returned: the matrix of estimated distances, the matrix of estimated distance variances, the matrix of negative log-likelihood between the sequences.

Author(s)

Ge Tan

References

Thorne, J.L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34, 3-16.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. Science 256, 1443-1445.

See Also

AACharacterSet, GONNET, GONNETBF

Examples

```
  ## This example is not tested due to running time > 5s
data(GONNET)
data(GONNETBF)
library(seqinr)
fasta <- read.fasta(file.path(system.file("extdata", package="TKF"),
                   "pair1.fasta"),
                   seqtype="AA", set.attributes=FALSE)
## 1D estimation: only distance
TKF92HG(fasta, mu=5.920655e-04, r=0.8, Ps=1, Kf=1.2,
        substModel=GONNET, substModelBF=GONNETBF)

## 2D estimation: joint estimation of distance, mu and r
TKF92HG(fasta, substModel=GONNET, substModelBF=GONNETBF)

## only apply to a pair of sequences
seq1 <- fasta[[1]]
seq2 <- fasta[[2]]
TKF92HGPair(seq1, seq2, mu=5.920655e-04, r=0.8, Ps=1, Kf=1.2,
          substModel=GONNET, substModelBF=GONNETBF)
```

Index