Package ‘TKF’

February 19, 2015

Version 0.0.3
Date 2015-02-10
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
Depends R (>= 3.0.2)
Suggests RUnit, seqinr
SystemRequirements gsl
License GPL-2
Type Package
NeedsCompilation yes
LazyData yes
Repository CRAN
Date/Publication 2015-02-11 01:02:49

R topics documented:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAToInt</td>
<td>2</td>
</tr>
<tr>
<td>GONNET</td>
<td>3</td>
</tr>
<tr>
<td>PAMn</td>
<td>3</td>
</tr>
<tr>
<td>TKF91</td>
<td>4</td>
</tr>
<tr>
<td>TKF92</td>
<td>6</td>
</tr>
</tbody>
</table>

Index 9
## 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

```r
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**

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

**Format**

- A 20x20 numeric matrix.
- A 20 numeric vector.

**Source**


**Examples**

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

---

**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 background frequency.

**Usage**

```r
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 background 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


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 background 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


See Also

`AACharacterSet`, `GONNET`, `GONNETBF`
Examples

```r
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

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

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>fasta</td>
<td>A named list of sequences in vector of characters format. read.fasta from package seqinr outputs this format when reading from a fasta file.</td>
</tr>
<tr>
<td>mu</td>
<td>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.</td>
</tr>
<tr>
<td>r</td>
<td>A numeric value of 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.</td>
</tr>
<tr>
<td>distance</td>
<td>A numeric value: the PAM distance between two protein sequences. When it is given, TKF91Pair only calculates the negative log-likelihood.</td>
</tr>
</tbody>
</table>
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 background 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.
Support for the heterogeneous evolutionary rate will be added in the near future.

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

See Also
AACharacterSet, GONNET, GONNETBF

Examples
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=0.701608861,
    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)
# Index

*Topic *textasciitildekwd1

<table>
<thead>
<tr>
<th>AAToInt</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAMn</td>
<td>3</td>
</tr>
<tr>
<td>TKF91</td>
<td>4</td>
</tr>
<tr>
<td>TKF92</td>
<td>6</td>
</tr>
</tbody>
</table>

*Topic *textasciitildekwd2

<table>
<thead>
<tr>
<th>AAToInt</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAMn</td>
<td>3</td>
</tr>
<tr>
<td>TKF91</td>
<td>4</td>
</tr>
<tr>
<td>TKF92</td>
<td>6</td>
</tr>
</tbody>
</table>

*Topic datasets

| GONNET | 3 |

AACharacterSet | 3, 5, 7 |
AACharacterSet (AAToInt) | 2 |
AAGapCharacterSet (AAToInt) | 2 |
AAToInt | 2 |
AmbiguousAACharacterSet (AAToInt) | 2 |
AmbiguousAAGapCharacterSet (AAToInt) | 2 |

character | 2, 5, 7 |

Dayhoffn (PAMn) | 3 |
DNACharacterSet (AAToInt) | 2 |
DNAGapCharacterSet (AAToInt) | 2 |

GONNET | 3, 5, 7 |
GONNETBF | 5, 7 |
GONNETBF (GONNET) | 3 |

matrix | 5, 7 |
numeric | 3–7 |

PAMn | 3 |

RNACharacterSet (AAToInt) | 2 |
RNAGapCharacterSet (AAToInt) | 2 |

TKF91 | 4 |
TKF91Pair (TKF91) | 4 |
TKF92 | 6 |
TKF92Pair (TKF92) | 6 |