Package ‘DNAseqtest’

March 24, 2016

Type Package

Title Generating and Testing DNA Sequences

Version 1.0

Date 2016-03-26

Author Faisal Ababneh, John Robinson, Lars S Jermiin and Hasinur Rahaman Khan

Maintainer Hasinur Rahaman Khan <`hasinurkhan@gmail.com`>

Description Generates DNA sequences based on Markov model techniques for matched sequences. This can be generalized to several sequences. The sequences (taxa) are then arranged in an evolutionary tree (phylogenetic tree) depicting how taxa diverge from their common ancestors. This gives the tests and estimation methods for the parameters of different models. Standard phylogenetic methods assume stationarity, homogeneity and reversibility for the Markov processes, and often impose further restrictions on the parameters.

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2016-03-24 23:32:09

R topics documented:

`DNAseqtest-package` *Generating and Testing DNA Sequences*

Description

Generates DNA sequences based on Markov model techniques for matched sequences. This can be generalized to several sequences. The sequences (taxa) are then arranged in an evolutionary tree (phylogenetic tree) depicting how taxa diverge from their common ancestors. This gives the tests and estimation methods for the parameters of different models. Standard phylogenetic methods assume stationarity, homogeneity and reversibility for the Markov processes, and often impose further restrictions on the parameters.

Details

Package:	DNAseqtest
Type:	Package
Version:	1.0
Date:	2016-03-26
License:	GPL-2

Author(s)

Faisal Ababneh, John Robinson, Lars S Jermiin and Hasinur Rahaman Khan Maintainer: Hasinur Rahaman Khan <hasinurkhan@gmail.com>

References

Lars Sommer Jermiin, Vivek Jayaswal, Faisal Ababneh, John Robinson (2008). Phylogenetic model evaluation. Bioinformatics, Volume 452 of the series Methods in Molecular Biology, 331-364.

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

Faisal Ababneh, Lars S Jermiin, John Robinson (2006). Generation of the Exact Distribution and Simulation of Matched Nucleotide Sequences on a Phylogenetic Tree. Journal of mathematical modelling and algorithms, 5(3), 291-308.

Examples

```
#To generate a 4^5 gene array
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3),4,2)
theta<-c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47), 3,.1,.5,.8)
gn.sec<-gn(theta, merge2)
```

`gn.sec`

`artomat` *Transforming $4\hat{}K$ Array to $m \times K$ Matrix*

Description

This function transfers any array to a matrix.

Usage

```
artomat(fobs)
```

Arguments

`fobs`	a 4^K array, containing the observed divergent frequencies for K matched sequences

Details

This function transfers any 4^K array containing the observed divergent frequencies of K matched sequences to an m x K matrix, where m is the sum of the frequencies in the 4^K observed divergence array.

Value

An m x K matrix

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

gn2, gn, Fmatrix

`Distance` *Paralinear Distances*

Description

This function calculates the paralinear distance between K matched DNA sequences.

Usage

```
Distance(F4)
```

Arguments

`F4`	a 4^K array containing the joint distribution array $F(t)$ or the observed array N

Details

This function calculates the paralinear distances between K matched DNA sequences, depending on the joint distribution array for these K sequences or on the observed divergence array N.

Value

A K x K symmetric matrix distances between the K sequences

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

gn2, gn, Fmatrix, Ntml

Examples

```
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
theta<-c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47), 3,.1,.5,.8)
F1<-gn(theta,merge2)
dn<-Distance(F1)
dn
```

`Fmatrix` *Joint Distribution for Two Matched Sequences*

Description

This function calculates the joint distribution function for two edge tree.

Usage

```
Fmatrix(t1, t2, f0, Sx2, Sy2, Pix, Piy)
```

Arguments

`t1`	represents the length from the tree root to the first node
`t2`	represents the length from the tree root to the second node
`f0`	the initial distribution for the four nucleotides
`Sx2`	a 4 x 4 symmetric matrix related to the first edge
`Sy2`	a 4 x 4 symmetric matrix related to the second edge
`Pix`	a diagonal matrix for the stationary distribution of the first edge
`Piy`	a diagonal matrix for the stationary distribution of the second edge

Details

This function calculates the joint distribution function for a two edge tree with different edge lengths, stationary distributions and differentS matrices.

Value

A 4 x 4 matrix containing the joint edges

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

gn, Smatrix

Examples

```
f0<-c(.25,.25,.25,.25)
Pi1<-diag(c(.2,.2,.2,.4))
Pi2<-diag(c(.1,.1,.1,.7))
S1<-Smatrix(c(.2,.2,.2,.2,.2,.2),diag(Pi1))
S2<-Smatrix(c(.3,.3,.3,.3,.3,.3),diag(Pi2))
fm<-Fmatrix(1, .5, f0, S1, S2, Pi1, Pi2)
fm
```

`gn` *Joint Distribution for K Matched Sequences*

Description

This function calculates the joint distribution array for K matched sequences.

Usage

```
gn(theta, merge2)
```

Arguments

`theta`	a vector of variables containing the following parameters in this order–1. the first three parameters from π_X vector, 2. the first three parameters from π_Y vector, 3. the first three parameters from f_0 vector, 4. the six off-diagonal free parameters in the S matrix, 5. a scalar ρ, 6. a vector of lengths containing K-2 values
`merge2`	(K-1) x 2 matrix describing the tree topology

Details

This function calculates the joint distribution array for a tree with K matched sequences. it uses the following functions– Pt, Fmatrix and Smatrix.

Value

A 4^K array containing the joint distribution for the K edges

References

Lars Sommer Jermiin, Vivek Jayaswal, Faisal Ababneh, John Robinson (2008). Phylogenetic model evaluation. Bioinformatics, Volume 452 of the series Methods in Molecular Biology, 331-364.

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Fmatrix, Pt, Smatrix

Examples

```
#To generate a 4^5 gene array
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3),4,2)
theta<-c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47), 3,.1,.5,.8)
gn.sec<-gn(theta, merge2)
gn.sec
```

`gn2` *Joint Distribution for K Matched Sequences (2)*

Description

This function calculates the joint distribution array for K matched sequences (second option).

Usage

```
gn2(theta, merge2)
```

Arguments

`theta`	a vector of variables containing the following parameters in this order–1. the first three parameters from π_X vector, 2. the first three parameters from π_Y vector, 3. the first three parameters from f_0 vector, 4. the six off-diagonal free parameters in the S matrix, 5. a scalar ρ, 6. a vector of lengths containing K-2 values
`merge2`	(K-1) x 2 matrix describing the tree topology

Details

This function calculates the joint distribution array for a tree with K matched sequences. it uses the following functions– Pt, Fmatrix and Smatrix.

Value

A 4^K array containing the joint distribution for the K edges

References

Lars Sommer Jermiin, Vivek Jayaswal, Faisal Ababneh, John Robinson (2008). Phylogenetic model evaluation. Bioinformatics, Volume 452 of the series Methods in Molecular Biology, 331-364.

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Fmatrix, Pt, Smatrix

Examples

```
#To generate a 4^5 gene array
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
rho2<-matrix(c(.3,.5,.3,.2,.3,.5,.8,2.7),4,2)
theta<-c(rep(.25,3), rep(.25,3),rep(.25,3), c(.2,.35,.79,.01,.93,.47),rho2)
gn2<-gn2(theta, merge2)
gn2
```

`gn3sim` *Generating Random DNA Samples Using the Rambaut and Grassly Method*

Description

This function generates random DNA samples using Rambaut and Grassly method.

Usage

```
gn3sim(theta, seqLength, merge2)
```

Arguments

`theta`	a vector of variables containing the following parameters in this order–1. the first three parameters from π_X vector, 2. the first three parameters from π_Y vector, 3. the first three parameters from f_0 vector, 4. the six off-diagonal free parameters in the S matrix, 5. a scalar ρ, 6. a vector of lengths containing K-2 values
`seqLength`	the length of sequences we need to generate
`merge2`	(K-1) x 2 matrix describing the tree topology

Details

This function generates a 4^K DNA array using Rambaut and Grassly, (1997) method. It depends on a set of variables theta, the sequence length and a merge matrix describing the tree topology.

Value

A n x K observed divergence matrix

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Ntml, simapp, simemb, gn, gn2, Fmatrix

Examples

```
# This will give 4^5 observed divergence array
theta<-(c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47),
3,.1,.5,.8))
n<-1000
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
gn3<-gn3sim(theta, n, merge2)
gn3
```

`likelihood` *Negative Log Likelihood Ratio*

Description

This function calculates log likelihood ratio value.

Usage

```
likelihood(thetast, fobs, merge2)
```

Arguments

`thetast`	a starting values for the parameter we need to estimate
`fobs`	the 4^K joint distribution array for K edge tree
`merge2`	a (K-1) x 2 matrix describing the tree topology

Details

This function calculates the log likelihood ratio value for F(t). It needs a vector of starting values for the parameters estimate, 4^K observed divergence array and merge matrix describing the tree topology.

Value

The value of the log likelihood ratio

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

gn, gn2

Examples

```
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
theta<-c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47),3,.1,.5,.8)
F1<-gn(theta, merge2)
lh<-likelihood(theta, F1, merge2)
lh
```

`Ntml` *Generating Samples from a Multinomial Distribution*

Description

Generating random DNA samples from a multinomial distribution.

Usage

```
Ntml(N, Ft)
```

Arguments

`N`	sample size
`Ft`	a 4^K array, containing the joint distribution probabilities for K matched sequences.

Details

This function generates a 4^K DNA array from a multinomial distribution. It depends on the sample size we need to generate and the 4^K joint distribution array of K matched sequences.

Value

A 4^K observed divergence array

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

simemb, simapp, gn3sim, gn, gn2, Fmatrix

Examples

```
#This will give a 4^K observed divergence array
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
theta<-c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47),
3,.1,.5,.8)
F1<-gn(theta,merge2)
Nt<-Ntml(1000, F1)
Nt
```

`Pt` *Transition Probability Function*

Description

This function calculates the transition probability function for a process during a period of time.

Usage

```
Pt(S, Pi, t)
```

Arguments

`S`	a 4 x 4 symmetric matrix
`Pi`	a diagonal matrix containing the stationary distribution for the process
`t`	a period of time describing the length of the process

Details

This function needs the 4 x 4 symmetric matrix S, II and the process length t in order to find the transition probability over that process, where $P_{ij}(t)$ is the probability that the ith nucleotide changes to the j-th nucleotide during the period of t.

Value

A 4 x 4 matrix containing the transition probabilities for a process.

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Smatrix

Examples

```
Pi<-diag(c(.1,.1,.1,.7))
S<-Smatrix(c(.3,.3,.3,.3,.3,.3),diag(Pi))
t<-1
p<-Pt(S, Pi, t)
p
```

`simapp` *Generating Random DNA Samples Using Approximation Method*

Description

This function generates random DNA samples using an approximation method

Usage

```
simapp(theta, seqLength, merge1)
```

Arguments

`theta`	a vector of variables containing the following parameters in this order–1. the first three parameters from π_X vector, 2. the first three parameters from π_Y vector, 3. the first three parameters from f_0 vector, 4. the six off-diagonal free parameters in the S matrix, 5. a scalar ρ, 6. a vector of lengths containing K-2 values
`seqLength`	the length of sequences we need to generate
`merge1`	(K-1) x 2 matrix describing the tree topology

Details

This function generates a 4^K DNA array using an approximation method. It depends on a set of variables theta, the sequence length and a merge matrix describing the tree topology.

Value

A n x K observed divergence matrix

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Ntml, simemb, gn3sim, gn, gn2, Fmatrix

Examples

```
# This will give 4^5 observed divergence array
theta<-(c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.2,.2,.2,.2,.2),
3,.1,.5,.8))
n<-1000
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
sa<-simapp(theta, n, merge2)
sa
```

`simemb` *Generating Random DNA Samples Using Embedded Markov Chain*

Description

This function generates random DNA samples using embedded chain.

Usage

```
simemb(theta, seqLength, merge2)
```

Arguments

`theta`	a vector of variables containing the following parameters in this order–1. the first three parameters from π_X vector, 2. the first three parameters from π_Y vector, 3. the first three parameters from f_0 vector, 4. the six off-diagonal free parameters in the S matrix, 5. a scalar ρ, 6. a vector of lengths containing K-2 values
`seqLength`	the length of sequences we need to generate
`merge2`	(K-1) x 2 matrix describing the tree topology

Details

This function generates 4^K DNA array using embedded Markov chain. It depends on a set of variables theta, the sequence length and a merge matrix describing the tree topology.

Value

A n x K observed divergence matrix

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Ntml, simapp, gn3sim, gn, gn2, Fmatrix

Examples

```
# This will give 4^5 observed divergence array
theta<-(c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47),
3,.1,.5,.8))
n<-1000
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
sm<-simemb(theta, n, merge2)
sm
```

`Smatrix` *Symmetric Matrix S*

Description

This function calculates the symmetric matrix S.

Usage

```
Smatrix(s, pix)
```

Arguments

`s`	a vector of variables containing the six free parameters in the S matrix
`pix`	a vector giving the stationary probabilities for the four nucleotides A, C, G and T

Details

This function calculates the matrix S, which we used to calculate the rate matrix R.

Value

A 4 x 4 symmetric matrix

References

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Pt, Fmatrix, gn ,gn2

Examples

```
s<-c(.1,.2,.3,.4,.5,.6)
pi<-c(.1,.1,.1,.7)
sm<-Smatrix(s, pi)
sm
```

`TEST2` *Test for Symmetry of Matched DNA Sequences*

Description

This function tests for symmetry between all the pairs of K matched DNA sequences.

Usage

```
TEST2(f)
```

Arguments

`f`	a 4^K array containing the observed divergence array N

Details

This function calculates Bowker's test for symmetry, Stuart's test for marginal symmetry and the test for internal symmetry. It depends on the 4^K observed divergence array N.

Value

A list of three lower triangle matrices

`first`	the lower triangle of the matrix contains (K-1) x (K-1) values shows Bowker's test between all the possible pairs of the K sequences
`second`	the lower triangle of the matrix contains (K-1) x (K-1) values shows Stuart's test between all the possible pairs of the K sequences
`third`	the lower triangle of the matrix contains (K-1) x (K-1) values shows the internal test between all the possible pairs of the K sequences

References

Lars Sommer Jermiin, Vivek Jayaswal, Faisal Ababneh, John Robinson (2008). Phylogenetic model evaluation. Bioinformatics, Volume 452 of the series Methods in Molecular Biology, 331-364.

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Ntml, simapp, simemb, TEST3

Examples

```
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
theta<-c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47),
3,.1,.5,.8)
F1<-gn(theta,merge2)
N1<-Ntml(1000,F1)
t2<-TEST2(N1)
t2
```

`TEST3` *Overall Test for Marginal Symmetry*

Description

This function tests for symmetry between K matched DNA sequences.

Usage

```
TEST3(Farray)
```

Arguments

`Farray`	a 4^K array containing the observed divergence array N

Details

This function calculates overall test for marginal symmetry. It depends on the 4^K observed divergence array N.

Value

A single value gives the overall test for marginal symmetry between K matched sequences

References

Lars Sommer Jermiin, Vivek Jayaswal, Faisal Ababneh, John Robinson (2008). Phylogenetic model evaluation. Bioinformatics, Volume 452 of the series Methods in Molecular Biology, 331-364.

Faisal Ababneh, Lars S Jermiin, Chunsheng Ma, John Robinson (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics, 22(10), 1225-1231.

See Also

Ntml, simapp, simemb, TEST2

Examples

```
merge2<-matrix(c(-1,-4,-3,2,-2,-5,1,3), 4, 2)
theta<-c(rep(.25,3), rep(.25,3), rep(.25,3), c(.2,.35,.79,.01,.93,.47),
3,.1,.5,.8)
F1<-gn(theta,merge2)
N1<-Ntml(1000,F1)
t3<-TEST3(N1)
t3
```

Index