Package 'alignmentStat'

November 20, 2019

Type Package

Title Statistical significance for alignment score.

Version 0.1.0

Author Giovana Millan and Paul Rognon

Maintainer Paul Rognon <paul.rognon@gmail.com>

Description The package is a wrapper for alignmentStat function. That function uses shuffling to produce a range of statistics and graphical outputs that help assess the statistical signficance of the alignment score of two protein or DNA sequences. The sequences are read from FASTA files. One of the sequences is iteratively shuffled and scored against the other sequence to obtain a sample distribution of scores. The function estimates empirical p-value and e-value out of the obtained distibution. A Gumbel distribution is fitted on the sample and used to provide standardized Gumbel score and p-value.

Imports Biostrings,
ggplot2,
seqinr

License MIT License

Encoding UTF-8

LazyData true

R topics documented:

`alignment.stat` *Statistical significance of alignment score by shuffling*

Description

The function uses shuffling to produce a range of statistics and graphical outputs that help assess the statistical signficance of the alignment score of two protein or DNA sequences. The sequences are read from FASTA files. One of the sequences is iteratively shuffled and scored against the other sequence to obtain a sample distribution of scores. The function estimates empirical p-value and e-value out of the obtained distibution. A Gumbel distribution is fitted on the sample and used to provide standardized Gumbel score and p-value.

Usage

```
alignmentStat(sequence1, sequence2, type_seq, alignment,
              submatrix,gapOpenPenal, gapExtPenal, N,
              shuffled, summary.dist.score = TRUE)
```

Arguments

sequence1	A character vector containing the path to the FASTA file of the first sequence to align.
sequence2	A character vector containing the path to the FASTA file of the second sequence to align.
type_seq	A character vector indicating the type of sequences. One of "DNA" and "protein".
alignment	A character vector indicating the type of alignment. One of "global" and "local", where "global" = align whole strings with end gap penalties, "local" = align string fragments.
submatrix	A character vector indicating the substitution matrix representing the fixed substitution scores for an alignment.
gapOpenPenal	A numeric indicating the cost for opening a gap in the alignment.
gapExtPenal	A numeric indicating the incremental cost incurred along the length of the gap in the alignment..
N	A numeric indicating the number of times to shuffle.
shuffled	A numeric indicating the sequence to shuffle. One of 1 and 2.
summary.dist.score	A logical value indicating whether to show the summary statistics of the sample score distribution obtained through shuffling.

Details

The pairwise alignment is performed as implement in pairwiseAlignment. See the documentation for more details.

Value

The function returns a list containing the following components:

score	A numeric giving the raw score from the pairwise alignment.
std.score	A numeric giving the score standardized according to the Gumbel fit.
plot	The ggplot displayed when running the function.
gumbel.fit	A named numeric vector of the fitted Gumbel parameters: lambda, u and K.
emp.p.value	A numeric giving the p-value estimated on the empirical distribution of scores obtained by shuffling.
gumbel.p.value	A numeric giving the p-value using the Gumbel distribution fitted on the sample of scores obtained by shuffling.

Author(s)

Giovana Millan and Paul Rognon

Examples

```
#############################################################################################################
#Example of local alignment of two close DNA sequences with PAM50 and shuffling on the second sequence
#############################################################################################################

#Define the filepaths to FASTA files:
filepath1 <- system.file("extdata", "gi32141095_N_0.fa", package="alignmentStat")
filepath2 <- system.file("extdata", "gi32141095_N_1.fa", package="alignmentStat")


## Run the function
as1<-alignment.stat(
  sequence1=filepath1,
  sequence2=filepath2,
  type_seq="DNA",
  alignment="local",
  submatrix="BLOSUM50",
  gapOpenPenal=-5,
  gapExtPenal=-1,
  N=1000,
  shuffled=1)

#############################################################################################################
#Example of local alignment of two close DNA sequences with PAM250 and shuffling on the second sequence
#############################################################################################################

#Define the filepaths to FASTA files:
filepath1 <- system.file("extdata", "B8D9R6.fasta", package="alignmentStat")
filepath2 <- system.file("extdata", "Q96IY4.fasta", package="alignmentStat")


## Run the function
as2<-alignment.stat(
  sequence1=filepath1,
  sequence2=filepath2,
  type_seq="protein",
  alignment="global",
  submatrix="PAM250",
  gapOpenPenal=-3,
  gapExtPenal=-0.5,
  N=1000,
  shuffled=2)

#############################################################################################################
#Example of plot
#############################################################################################################

\Figure{Example_alignmentStat_plot.png}{}
```

---

gumbel.score.fit *Gumbel distribution fit on score sample*

Description

The function is a support function for alignmentStat. It fits a Gumbel distribution on the alignment score distribution obtained through shuffling. Paremeters are estimated trhough the method of moments.

Usage

```
gumbelScoreFit(score.sample,m,n)
```

Arguments

score.sample	A numeric vector of sample scores.
m	A numeric indicating the length of sequence 1.
score.sample	A numeric indicating the length of sequence 2.

Value

The functions returns a named numeric vector containing: lambda=lambda.hat,u=mode.hat,K=K.hat

lambda	A numeric giving the method of moments estimation of parameter lambda.
u	A numeric giving the method of moments estimation of parameter u.
K	A numeric giving the method of moments estimation of parameter K.

Author(s)

Giovana Millan and Paul Rognon.

Examples

```
##Example values
sample<-seq(20,200,1)
m<-25
n<-25

#Running the function
gumbel_param<-gumbel.score.fit(sample,m,n)
```

Index