Statistical Significance of a Pairwise Alignment of Sequences.

Alex-Alex-Helena

10.11.2021

```
library(StatSignfPairSeqAlign)
help(statSignf)
```

Output from Examples in help()-file

```
## Two random sequences in the specified format.
seq1 <- ">random sequence 1 consisting of 20 residues.
KMMIDIHWGMWWYEYMMCLD"
seq2 <- ">random sequence 1 consisting of 20 residues.
DVYRVCQNVFRYHHFCKRTI"

# Simple alignment. Only output in text.
statSignf(seq1, seq2, plot = F)
```

```
## The names of the sequences are 'random sequence 1 consisting of 20 residues' and
## 'random sequence 1 consisting of 20 residues'.
## These sequences are of type 'Protein'.
## The type of alignment that has been done here is 'global'.
## The substitution matrix used is 'BLOSUM62'.
## The gap scores are -3 for open penalty and -1 for extended penalty.
## The number of shuffles done are 1000.
## The shuffling was done on the first sequence mentioned above.
## The original score is 1.
## The estimations of the parameters of the Gumbel distribution are:
##  Scale parameter lambda = 0.2339798 and mode u = -1.721559.
## The p-value estimated using the estimated Gumbel distribution is 0.410799.
## The p-value estimated empirically by counting in the histogram is 0.416.
## The estimated K is 0.001671098.
## The standardized score is S' = 0.6367898.
## Summary of the 1000 scores calculated after shuffling:
## Min.     1st Qu.  Median   Mean     3rd Qu.  Max
##  -14.000   -3.000    0.000    0.745    4.000   25.000
```

```
# More specific alignment. Only output in text.
statSignf(seq1, seq2, num.shuffles = 500, plot = F,
      subst.matrix = "PAM30", kind.align = "local")
```

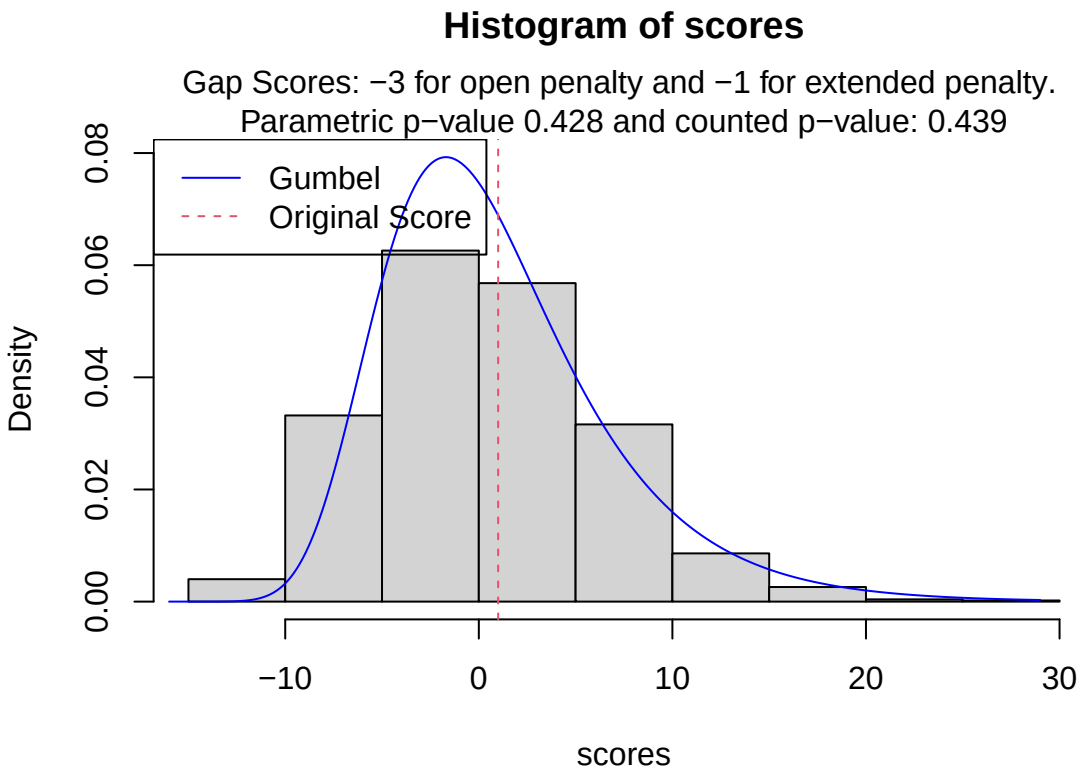
```
## The names of the sequences are 'random sequence 1 consisting of 20 residues' and
## 'random sequence 1 consisting of 20 residues'.
## These sequences are of type 'Protein'.
## The type of alignment that has been done here is 'local'.
## The substitution matrix used is 'PAM30'.
```

```
## The gap scores are -3 for open penalty and -1 for extended penalty.
## The number of shuffles done are 500.
## The shuffling was done on the first sequence mentioned above.
## The original score is 15.
## The estimations of the parameters of the Gumbel distribution are:
##  Scale parameter lambda = 0.398779 and mode u = 14.39277.
## The p-value estimated using the estimated Gumbel distribution is 0.543852.
## The p-value estimated empirically by counting in the histogram is 0.456.
## The estimated K is 0.7772993.
## The standardized score is S' = 0.2421506.
## Summary of the 500 scores calculated after shuffling:
## Min.     1st Qu.  Median   Mean     3rd Qu.  Max
##    10.00    14.00    15.00    15.84    18.00    28.00
```

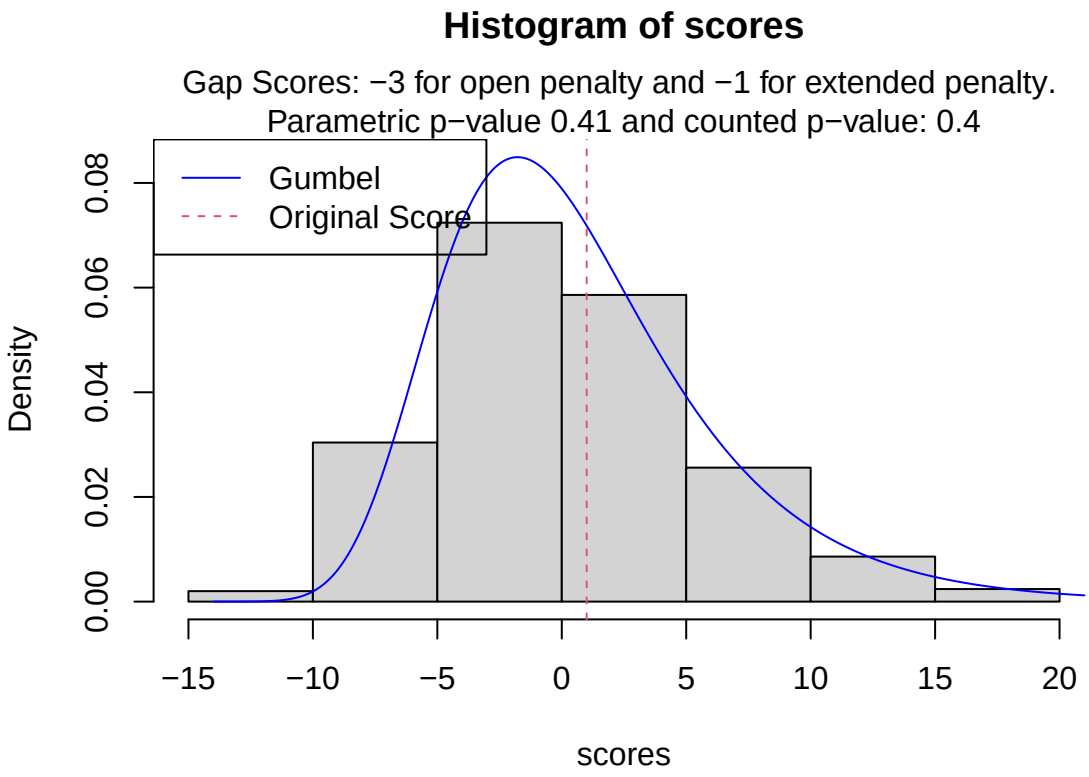
```
# Simple alignment. Output in text and plot.
out <- statSignf(seq1, seq2, plot = T)
```

```
## The names of the sequences are 'random sequence 1 consisting of 20 residues' and
## 'random sequence 1 consisting of 20 residues'.
## These sequences are of type 'Protein'.
## The type of alignment that has been done here is 'global'.
## The substitution matrix used is 'BLOSUM62'.
## The gap scores are -3 for open penalty and -1 for extended penalty.
## The number of shuffles done are 1000.
## The shuffling was done on the first sequence mentioned above.
## The original score is 1.
## The estimations of the parameters of the Gumbel distribution are:
##  Scale parameter lambda = 0.2154623 and mode u = -1.697542.
## The p-value estimated using the estimated Gumbel distribution is 0.4283432.
## The p-value estimated empirically by counting in the histogram is 0.439.
## The estimated K is 0.001734179.
## The standardized score is S' = 0.5812188.
## Summary of the 1000 scores calculated after shuffling:
## Min.     1st Qu.  Median   Mean     3rd Qu.  Max
##  -15.000   -3.000    1.000    0.981    5.000   28.000
```

```
out() # Uses the returned function to plot the results.
```




```
# Simple alignment. Suppress output in text, only plot.
out <- statSignf(seq1, seq2, plot = T, suppress.output = T)
out() # Uses the returned function to plot the results.
```


Histogram of scores
Gap Scores: −3 for open penalty and −1 for extended penalty.
Parametric p−value 0.41 and counted p−value: 0.4
Gumbel
Original Score
Density
0.00
0.02
0.04
0.06
0.08
−15
−10
−5
0
5
10
15
20
scores