

# Folien zur Gumble- Extremwertverteilung

17.11.03

Bioinformatik I

Molekulare Biotechnologie

Dr Rainer König

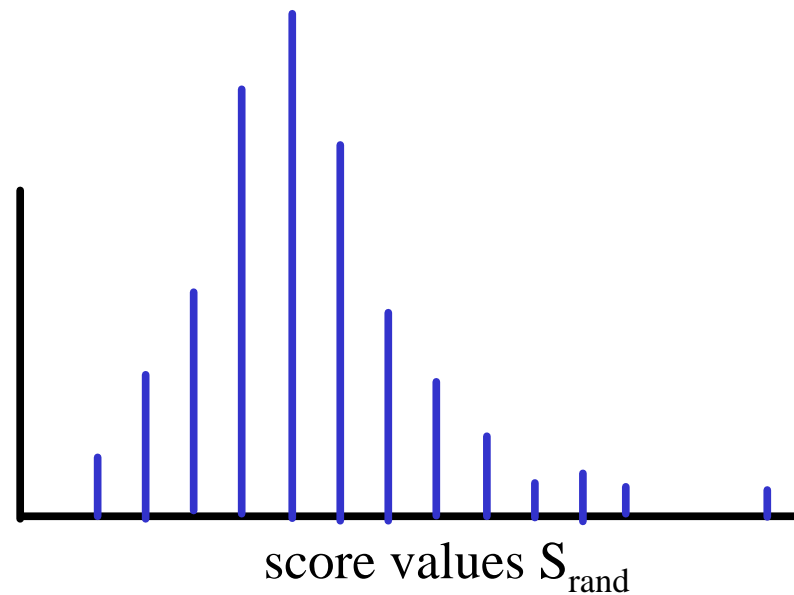
# How to estimate the significance of "my best score" of my 2 sequences?

method:

1. take the 2 given sequences
2. shuffle them
3. obtain the best score
4. repeat 1.-3.  $\approx 1000$  times
5. calculation:

plot of the distribution:

number of  
scores



This extreme value distribution fits to the  
"Gumble extreme value distribution"

mathematically spoken:  $p(S_{\text{rand}}) \sim \exp(-S_{\text{rand}} - \exp(-S_{\text{rand}}))$

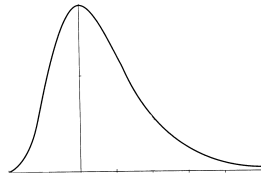
but: we are interested in the probability that a score from shuffled/random sequences is as good or better than the score  $S_{\text{real}}$  we yielded with the real sequences.

=> evaluation of the probability  $p(S_{\text{rand}} \geq S_{\text{real}})$

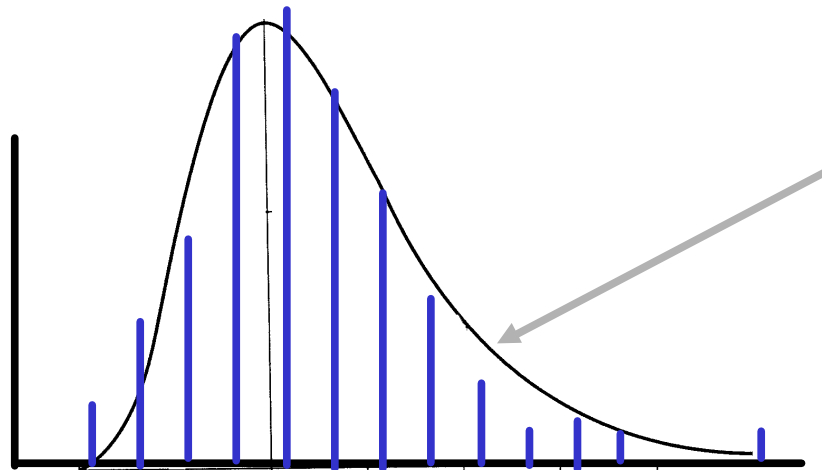
$$p(S_{\text{rand}} \geq S_{\text{real}}) = \int_{S_{\text{real}}}^{\infty} p(S) dS$$
$$= 1 - \exp(-\exp(-S_{\text{real}}))$$

if  $S_{\text{real}}$  high:  $\approx \exp(-S_{\text{real}})$  (Taylor)

for our purpose we have to fit the gumble curve with the outcome of the shuffled sequences, introduce  $K$  and  $\lambda$ :



not fitted gumble curve,  $K=1$ ,  $\lambda=1$



fitted gumble curve with optimal  $K$  and  $\lambda$

we yield the probability for a random sequence to get at least the score value of "my best score":

$$\text{(appr.) } p\text{-value} = p(S_{\text{rand}} \geq S_{\text{real}}) \approx K m n \exp(-\lambda S_{\text{real}})$$

$m, n$ : sequences' length  
 $K, \lambda$ : fitting parameters

remark. - the exact p-value =  $1 - \exp(-Kmn \exp(-\lambda S_{\text{real}}))$   
- the approximated p-value = e-value,  
for database search is m the sequence length of the query,  
n the effective length of sum of all sequences of the database

**practically:**

- 2 sequences



- LAlign



finds the best local alignment by the Smith Waterman method  
and a score (example, next slides: score=401)

- statistic program PRSS

needs sequences and score of LAlign, shuffles the sequences  $\approx 1000$  times  
gives out the extreme value distribution for the shuffled sequences,  
K,  $\lambda$  for the gumble distribution  
and the p-value for my best score

# Output of LAlign

LALIGN finds the best local alignments between two sequences  
version 2.0u64 March 1998

Please cite:

X. Huang and W. Miller (1991) Adv. Appl. Math. 12:373-381

Comparison of:

(A) lamc1.pro LAMC1 REFORMAT of: cipro.pro check: -1 from: 1 - 237 aa  
(B) p22c2.pro P22C2 REFORMAT of: p22 check: 4729 from: 1 to - 216 aa  
using matrix file: blosum50.mat, gap penalties: -12/-2

36.1% identity in 208 aa overlap; score: 401 [1/2 bits]

```

      30          40          50          60          70          80
LAMC1 KKNELGLSQESVADKMGMGQSGVGALFNGINALNAYNAALLAKILKVSVEEFSPSIAREI
      .....: . : ... :..... : . . : : : : . . . . .
P22C2 RRRKCLKIRQAALGKMVGVSNVAISQWERSETEPNGENLLALSALQCSPPDYLLKGDLSQT
      20          30          40          50          60          70

      90          100         110         120         130         140
LAMC1 YEMYEAVSMQPSLRSEYEYPVFSHVQAGMFSPELRTFTKGDARWVSTTKKASDSAFWLE
      :. . : : . : : : : : : : : : : : : : : : : : : : : : : : : :
P22C2 NVAYHS-RHEP--RGSY--PLISWVSAGQWMEAVEPYHKRAIENWHDTTVDCSEDSFWLD
      80          90         100         110         120

      150         160         170         180         190         200
LAMC1 VEGNSMTAPTGSKPSFPDGMILVDPEQAVEP--GDFCIARLGGD-EFTFKKLIRDSGQV
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
P22C2 VQGDSMTAPAGL--SIPEGMIILVDPE--VEPRNGKLVVAKLEGENEATFKKLVMDAGRK
      130         140         150         160         170         180

      210         220         230
LAMC1 FLQPLNPQYPMIPCNESSVVGKVIASQ
      : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
P22C2 FLKPLNPQYPMIEINGNCKIIGVVVDAK
      190         200         210
```

# Output of PRSS

lamcl.pro, 237 aa vs p22c2.pro

```
      s-w  est
24      0    0:
26      0    0:
28      3    1:*==
30     13    6:====*====
32     27   21:====*====
34     68   50:====*====*
36     98   84:====*====*
38    128  111:====*====*
40    129  123:====*====*
42    105  121:====*====*
44    110  108:====*====*
46     63   91:====*====*
48     75   72:====*====*
50     35   56:====*====*
52     48   42:====*====*
54     30   32:====*====*
56     19   23:====*====*
58     17   16:====*====*
60      6   13:====*====*
62      7    9:====*====*
64      7    6:====*====*
66      2    5:====*====*
68      4    3:====*====*
70      0    2:====*====*
72      1    2:====*====*
74      0    1:====*====*
76      1    1:====*====*
78      2    1:====*====*
80      0    0:
82      0    0:
84      0    0:
86      1    0:
88      1    0:
90      0    0:
92      0    0:
94      0    0:
96      0    0: O
```

216000 residues in 1000 sequences,  
BLOSUM50 matrix, gap penalties: -12,-2  
unshuffled s-w score: 401; shuffled score range: 30 - 89  
Lambda: 0.16931 K: 0.020441; P(401)= 3.7198e-27  
For 1000 sequences, a score >=401 is expected 3.72e-24 times