

# Análise de significância de alinhamentos

## Análise de significância de um alinhamento

Tão importante como escolher o método de *scoring* ou encontrar o alinhamento que maximiza o score é saber avaliar a *significância estatística* do alinhamento obtido.

Como se compara o score obtido no alinhamento com o score obtido alinhando duas sequências não-relacionadas ?

OU

Qual a probabilidade de obter um score idêntico ao obtido ao alinhar duas sequências aleatórias ?

# Análise de significância de um alinhamento



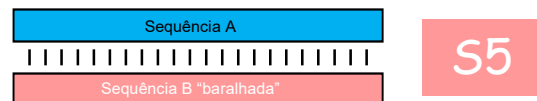
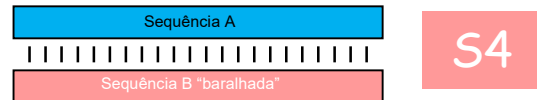
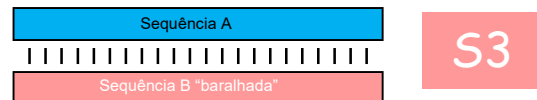
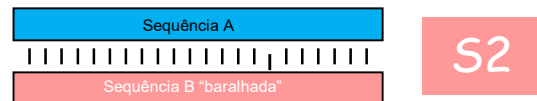
Para duas sequências relacionadas, esperamos que o score do alinhamento original seja superior ao score do alinhamento “baralhado”, ou seja  $X > Y$

Ao “baralhar” a sequência permutando os aminoácidos, mantemos a percentagem de composição.

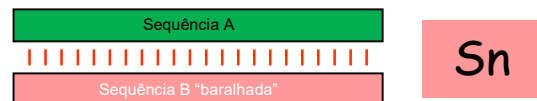
A sequência deve ser baralhada muitas vezes e o alinhamento repetido, para obter uma *distribuição de scores*.

# Sequências relacionadas

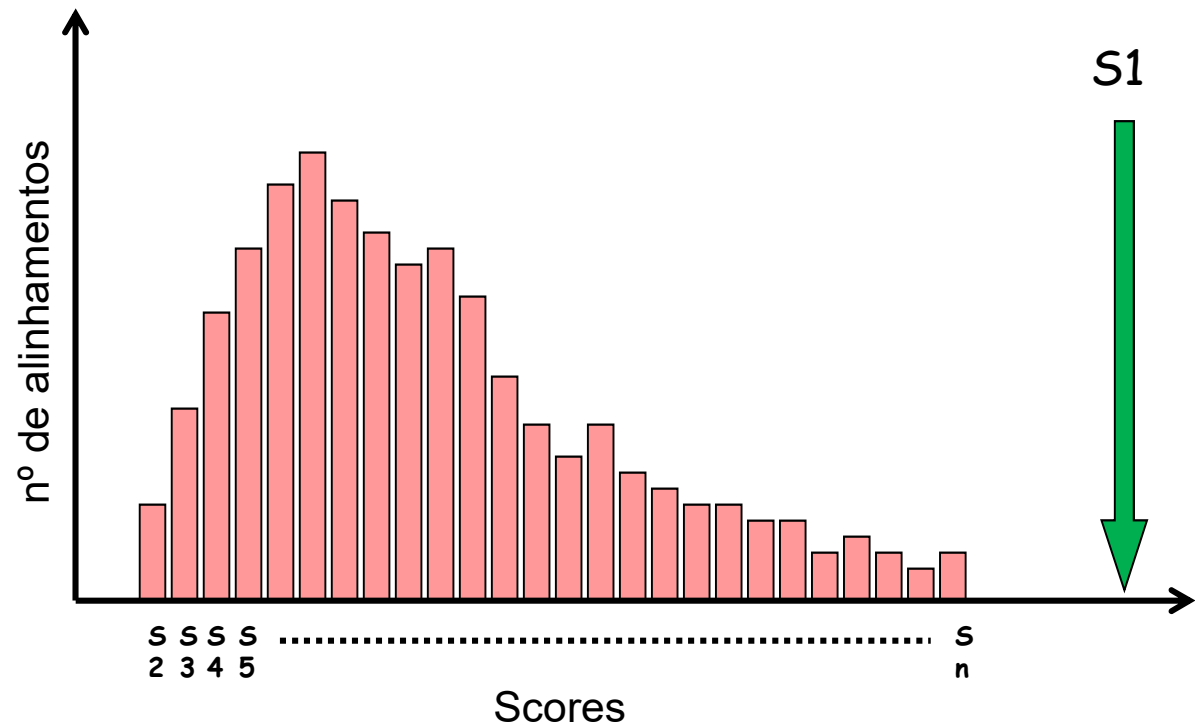
Scores:



⋮



O score S1 para o alinhamento das duas sequências encontra-se distante do intervalo de valores prováveis para um alinhamento aleatório



SeqA - MV-LSPADKTNVKAAWGKVGHAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSHGSAQVKGHG  
 SeqB - MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK

Score = 129, id = 40%

SeqA - MVLSPADKTNVKAAWGKVGHAHAGEYGAEALERMFLSF-PTTKTYFPH-FDLSHGSAQVK----GHG  
 SeqB - RLL--VSKNSV----NLVGTETEVGGWSPVEQLVTAFRPKTMHFEggwLAVAAGDDKVELMDPFYP

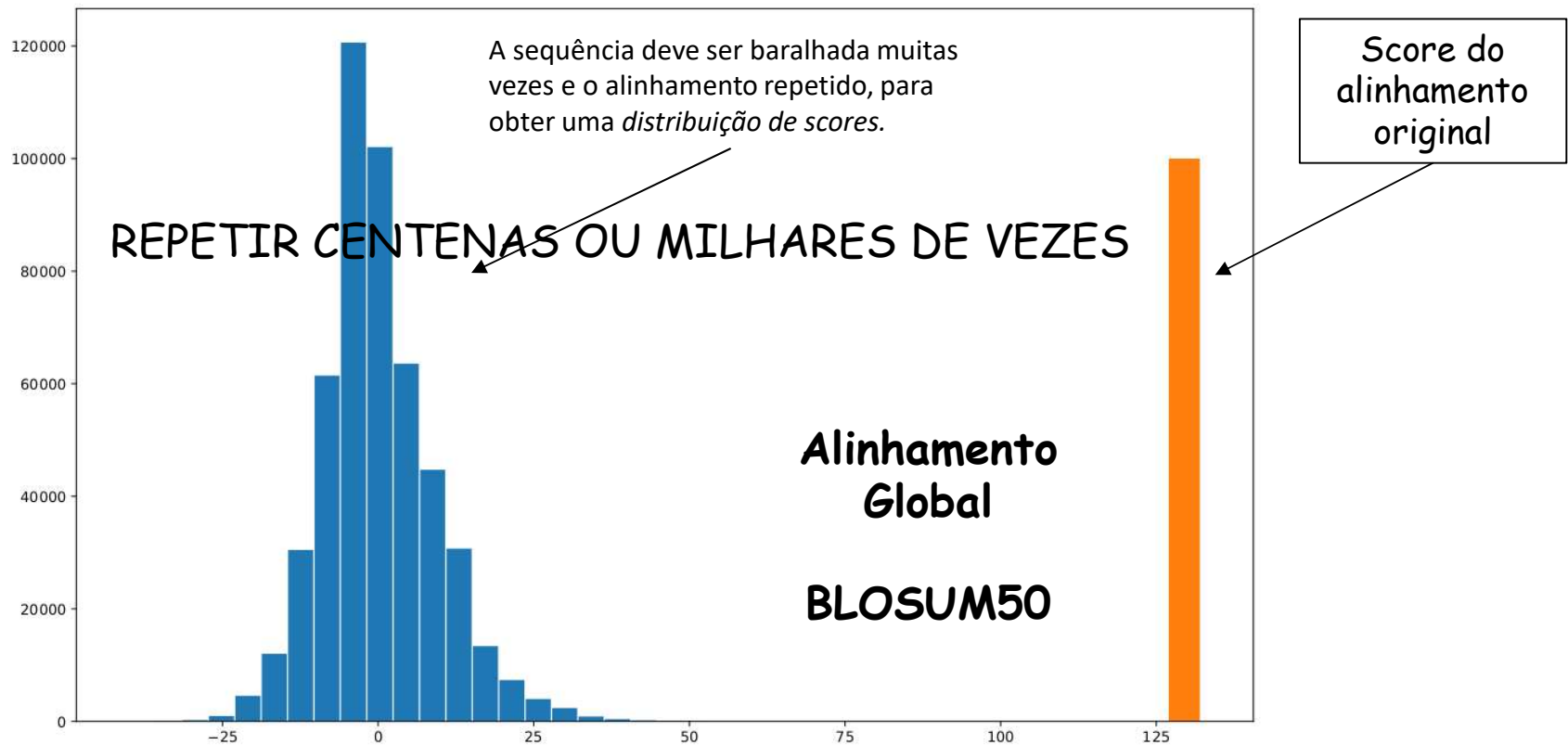
Score = 1

SeqA - MVLSPADKTNVKAAWGKVGHAHAGE-YGAEALERMFLSFPTTKT-----YFPHFDLSHGSAQVKGH-G  
 SeqB - ---GPWDAFTLADVITYFHGEKLGENVGLEPMSKMNPg-PVVQTADETWSLELVGAVVF-----LRSKVR

Score = -12

SeqA - MVLSPADKTNVKAAWGKVGHAH---AGEYGAEALERMFLSFPTTKTYF---PHFDLSHGSAQVKGHG-----  
 SeqB - ---LPSYRS--FRMWGSLEPNDVVAGT--PTGLE-----WHELKTVFATGP--DLFDAKQNVGEGVGVVKLEAVM

Score = 8



SeqA - LSPADKTNVKAANGKVGAGHAGEYGAEALERMFLSFPTTKTYFPHF-DLS  
 SeqB - LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLS

Score = 95, id = 44%

SeqA - AEALERMFLS-FPTTKTYFP-HFDLSHGSAQVK  
 SeqB - AAPLDVAFLGPVPVSPLFMDGHLEVSTEEWDVK

Score = 18

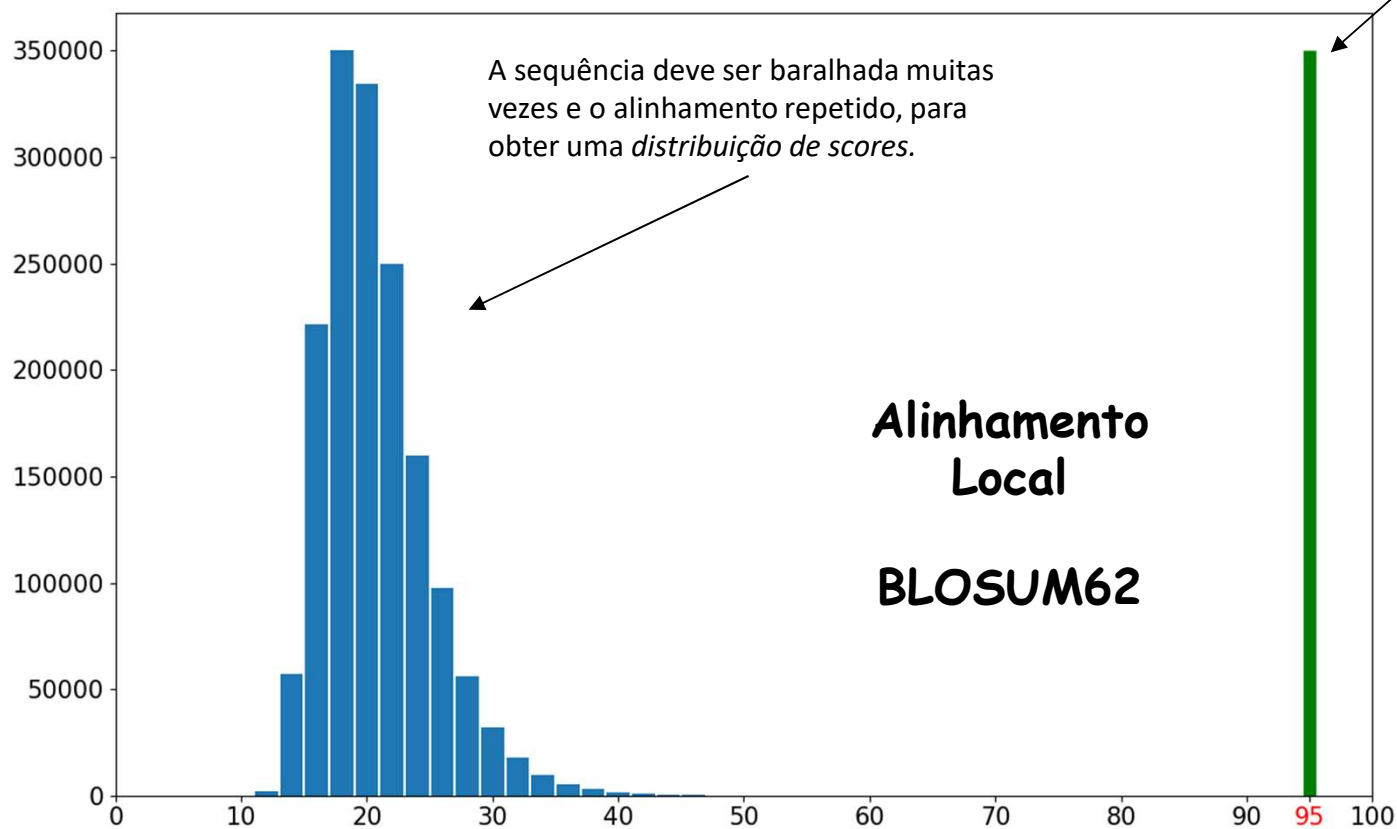
SeqA - ERMFLSFPTTKTYFP-HFDLSHGSAQVKGH  
 SeqB - KRQNSSVDTATTGFPVLPWFVKEGLAEVVGH

Score = 34

SeqA - GAHAGEYG-AEALERMFLSFPTTKTYFPHFDLSHGSAQ  
 SeqB - GFTEGKTGMATWMSPVVVELAVTKLPFLFGDLVRGELE

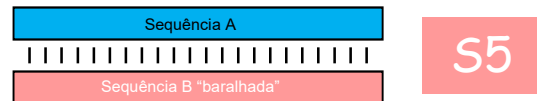
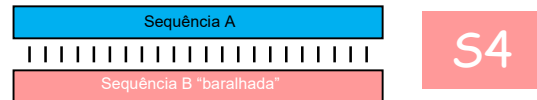
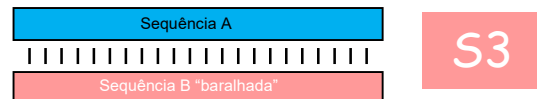
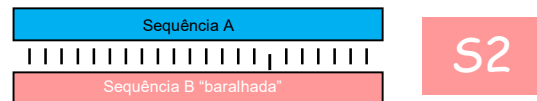
Score = 22

Score do  
alinhamento  
original

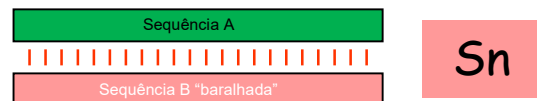


# Sequências não-relacionadas

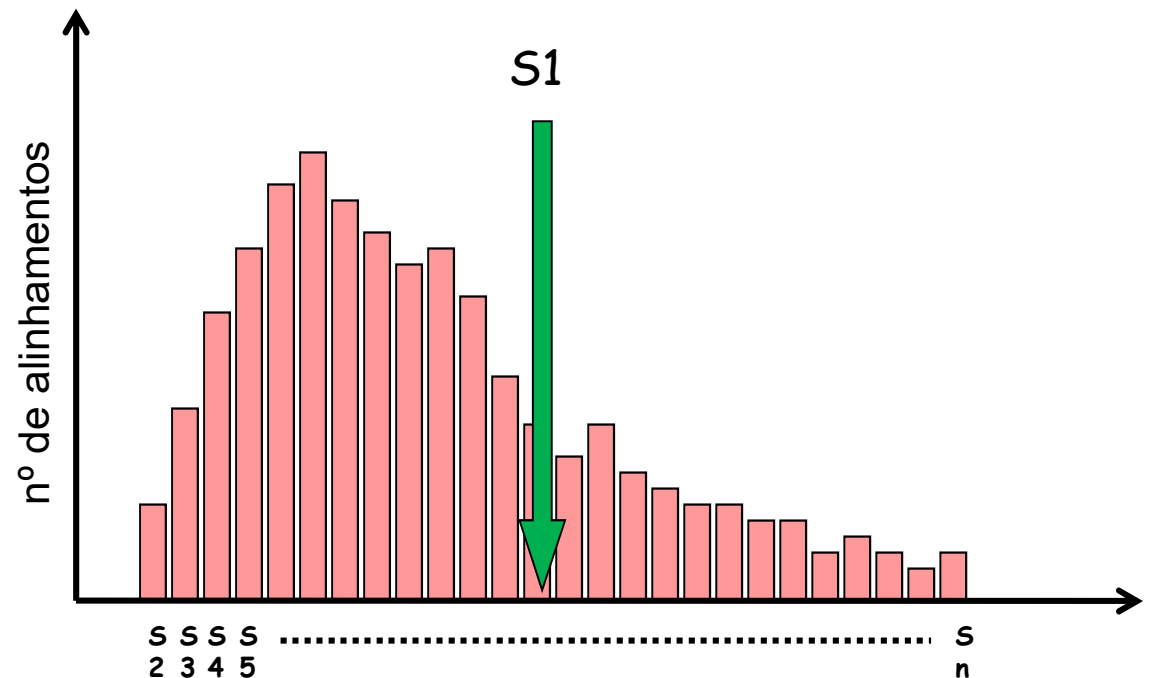
Scores:



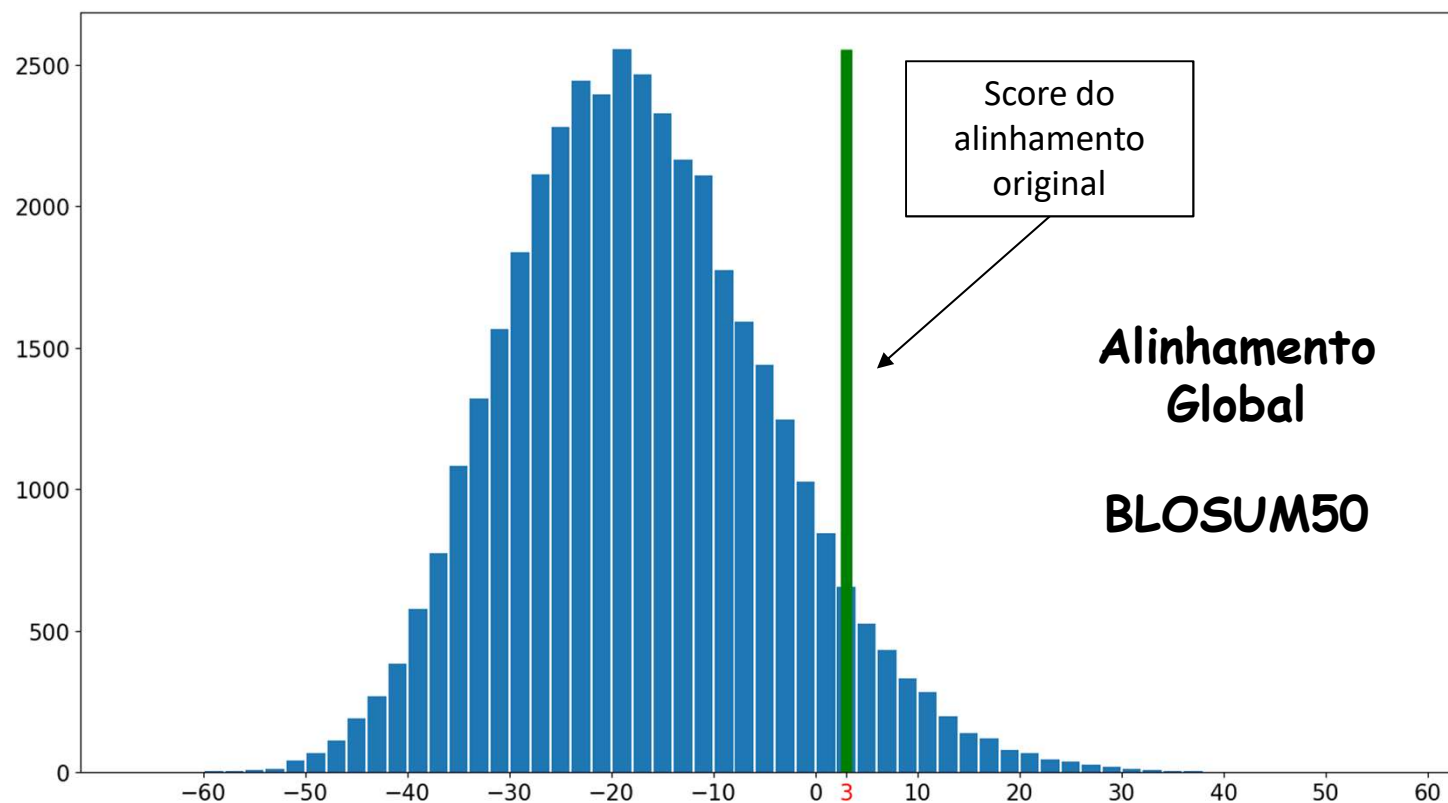
...



O score S1 para o alinhamento das duas sequências encontra-se dentro do intervalo de valores prováveis para um alinhamento aleatório



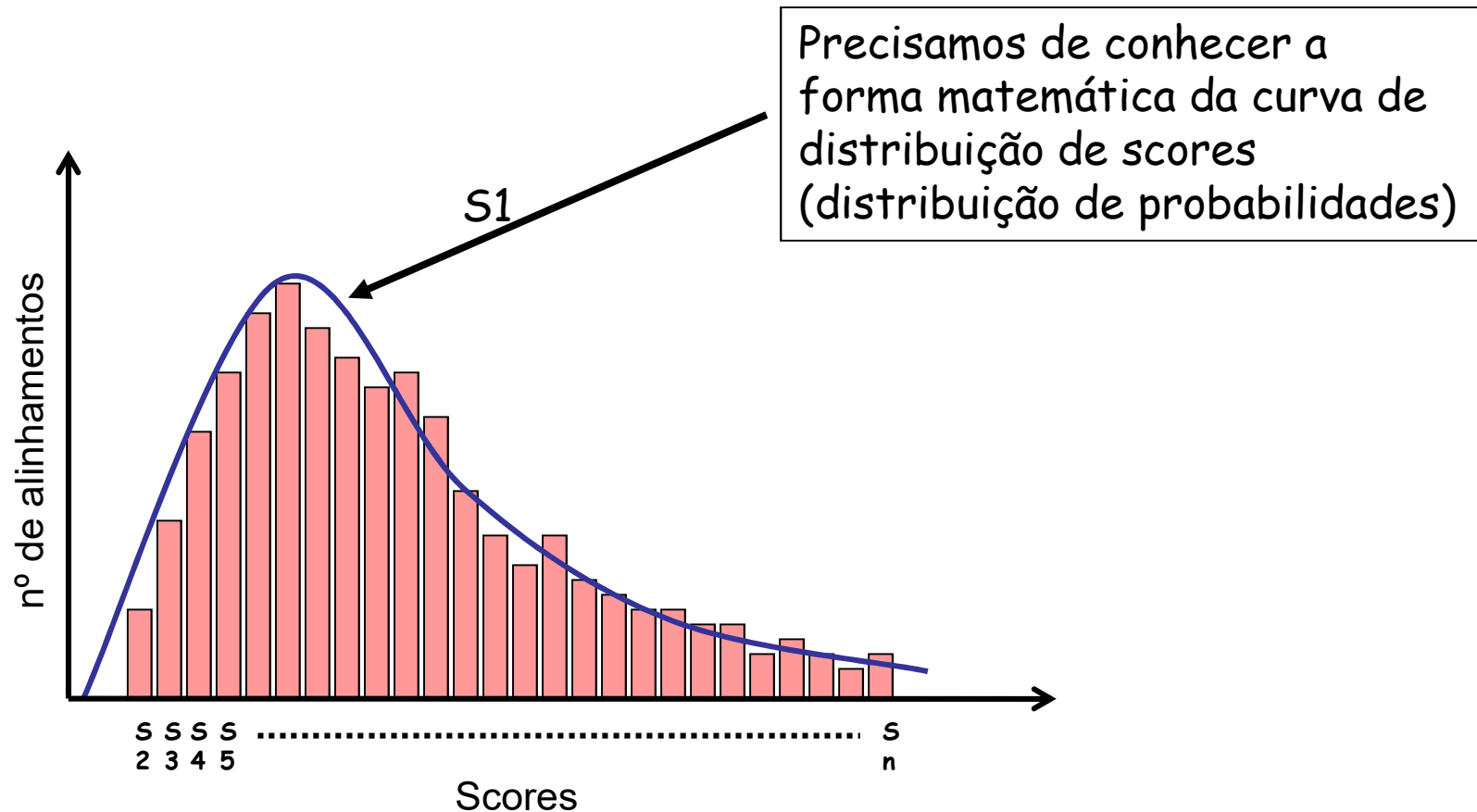
|        |   |             |
|--------|---|-------------|
| SeqA - | MVHLTPEEKSAVTALWGKVVNDEVG-GEALGRLLVV-----YPWTQRRFFESF-----GDLST-PDAVMGNPK | Score = 3   |
| SeqB - | MVH-----YKLMCFDVRGLGEVIRQLFYLGDVSFEDFRVSREEFKSLKSNLPSGQLPVLEIDGVM---      | id = 19%    |
| SeqA - | MVHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTQRF--FESFGDLSTPDAV--MGNP--K        | Score = -38 |
| SeqB - | -----DSKGSEFCLYSVMGRLEFFEQVLMDDLILMKSGDRFRVGVSNELIPPHQVRLEGVIFLV          |             |
| SeqA - | MVHLTPEEKSAVTALWGKVVNDEVGGE-----ALGRLLVVYPWTQRRFFESFGDLSTPDAV-MGNPK       | Score = 13  |
| SeqB - | M-DLSGQEPN-MCILYGMV-LSFVGSELVKLIFFGEKLRVVP---RLHEGERSFVQKDRLDVYSDF        |             |
| SeqA - | -MVHLTPEEKSAV-----TALWGK-VNDEVG-----GEALGRL-LVVYPWTQRRFFESFGDLSTPDAVMGNPK | Score = -14 |
| SeqB - | PMEEDSGQRRSGMFFPDGYLSSCIFKKSVDVLEQGVDMFVENLGFLELVVY-----KLIRLHLR          |             |





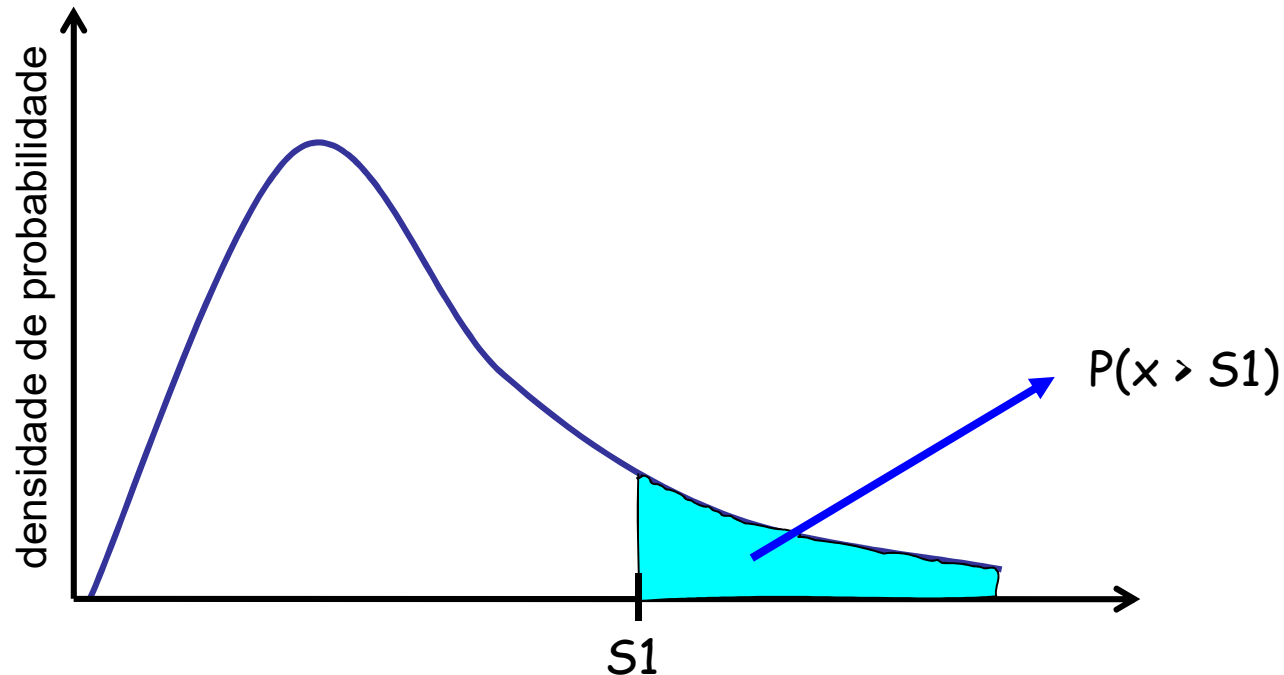
# Como calcular a probabilidade obter um dado score ?

- Para poder quantificar a significância estatística de um dado alinhamento, precisamos de calcular a probabilidade de obter um determinado *score* num alinhamento aleatório.



## A probabilidade é obtida a partir da curva de distribuição

- A probabilidade de obter um score  $x$  igual ou superior a  $S1$  será dada pela área debaixo da curva de densidade de probabilidade entre  $S1$  e  $+\infty$
- O histograma com a distribuição de scores tem que ser normalizado para que lhe possa ser ajustada uma densidade de probabilidade (a probabilidade de obter um score  $s$  tal que  $-\infty < s < +\infty$  tem que ser  $=1$ )



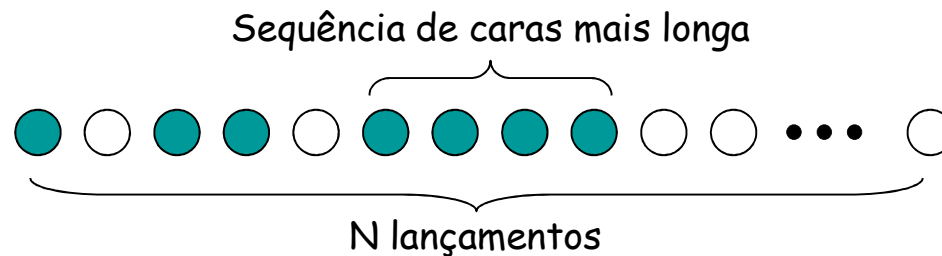
# Qual a curva de distribuição ?

- Ainda não existe um tratamento completamente geral para o problema estatístico do alinhamento de sequências
- O problema pode ser formulado de forma rigorosa para o caso do alinhamento local sem gaps
- As curvas de distribuição obtidas neste caso são aplicadas de forma empírica a situações mais complexas, como seja o alinhamento local com gaps ou o alinhamento global
- A distribuição de scores não é dada por uma distribuição normal (Gaussiana), mas sim por uma distribuição de valor extremo (Gumbel)

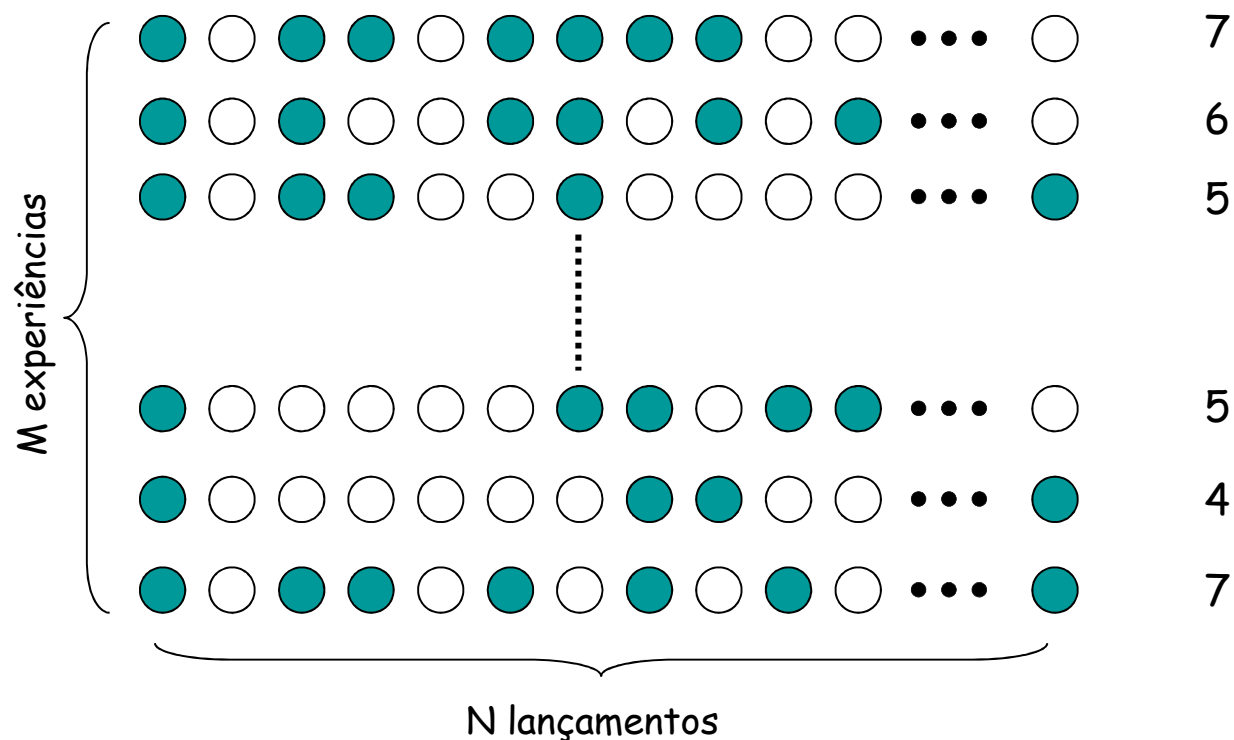
O alinhamento local sem gaps pode considerar-se como a busca da subsequência mais longa entre duas sequências:

→ AAVHVVRTVHAKLVGGARKHHIVT →  
 ← GAVHVTSKVHAKLIVGARLHTIRG ←  
 máxima sobreposição

Para sequências aleatórias, pode mostrar-se que este problema segue uma distribuição de probabilidade semelhante à do seguinte problema: sequência de caras *mais longa* num conjunto de  $n$  lançamentos de uma moeda.

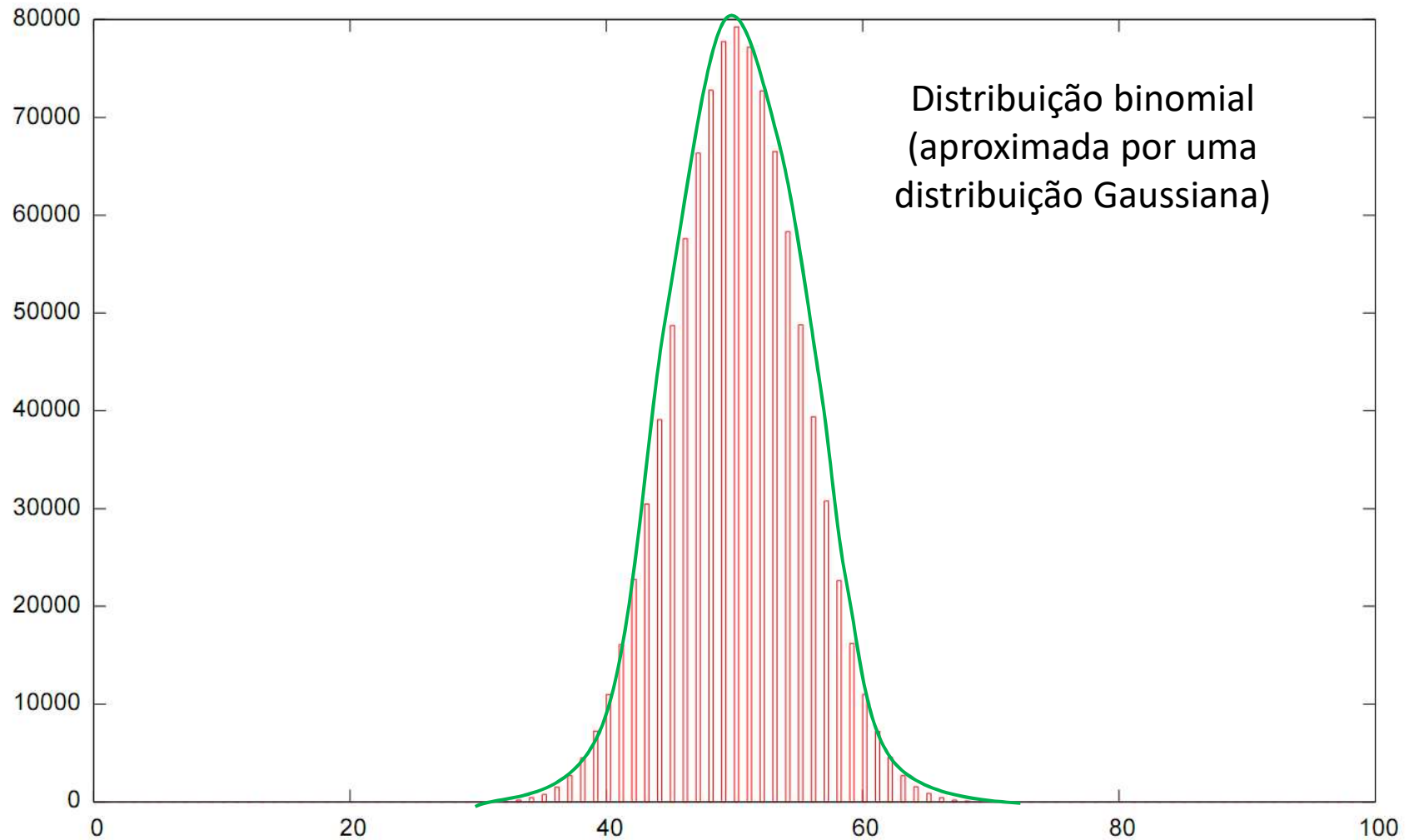


# Distribuição do número total de caras

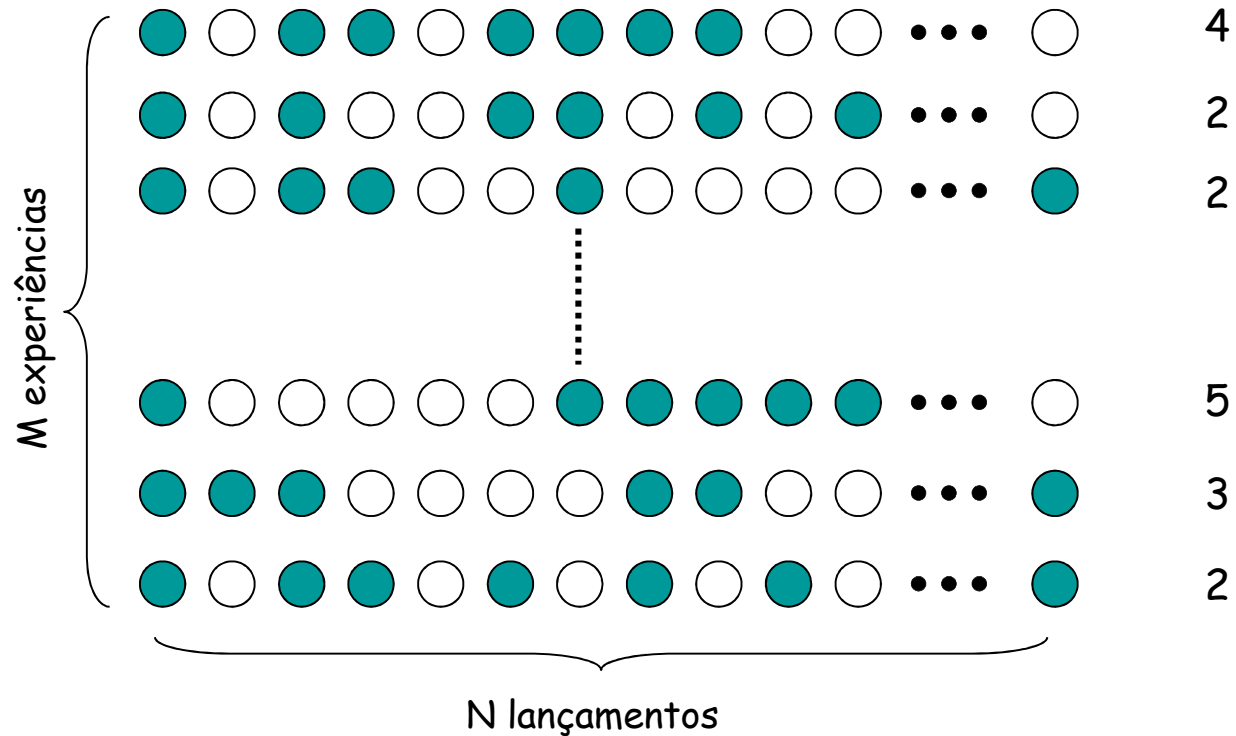


Consideremos  $M$  repetições dos  $N$  lançamentos de uma moeda, contando em cada uma das vezes o número de caras obtido.

# Distribuição do número total de caras

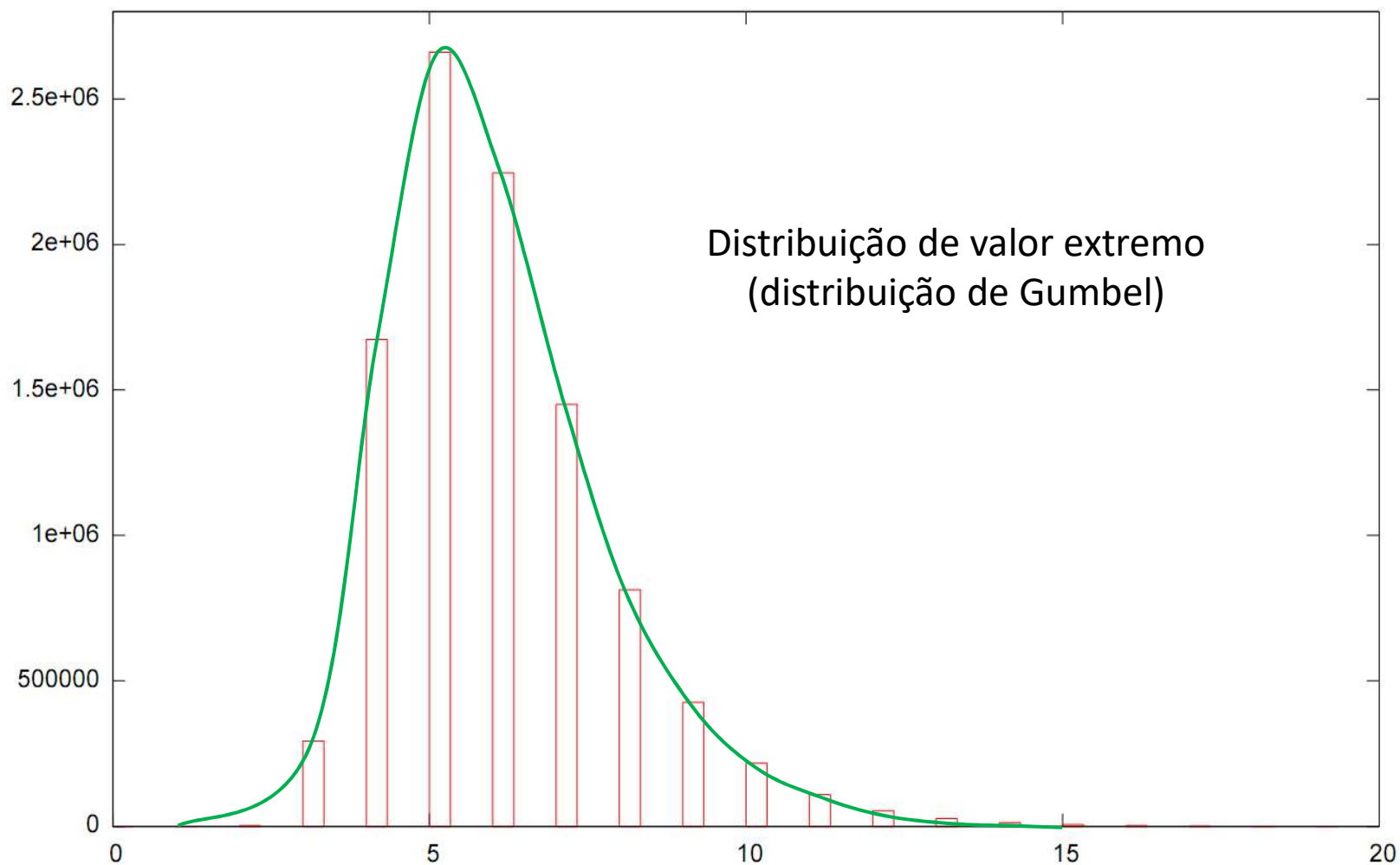


# Distribuição da sequência mais longa de caras



Consideremos  $M$  repetições dos  $N$  lançamentos de uma moeda, contando em cada uma das vezes o número de caras obtido.

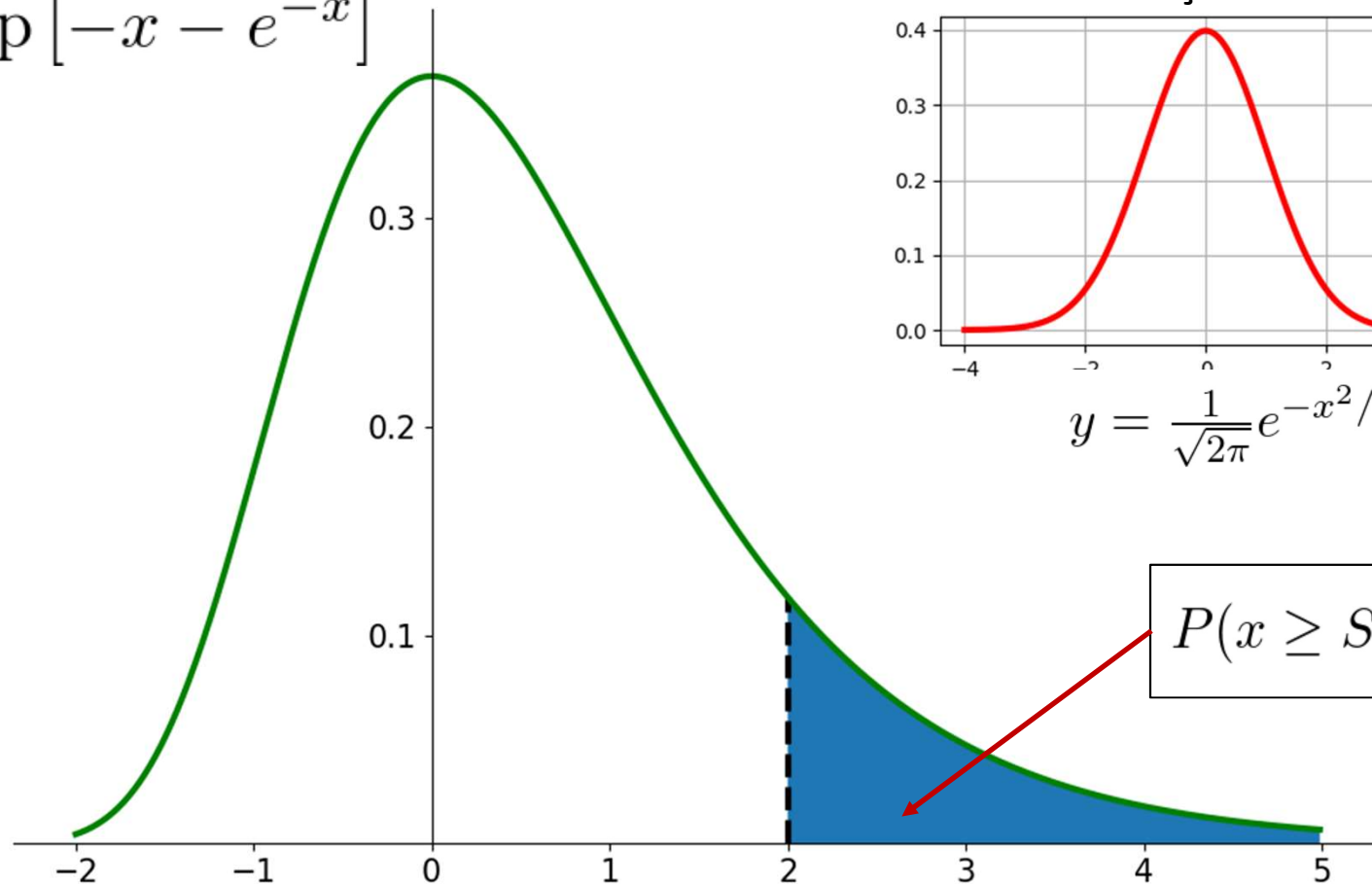
# Distribuição da sequência mais longa de caras





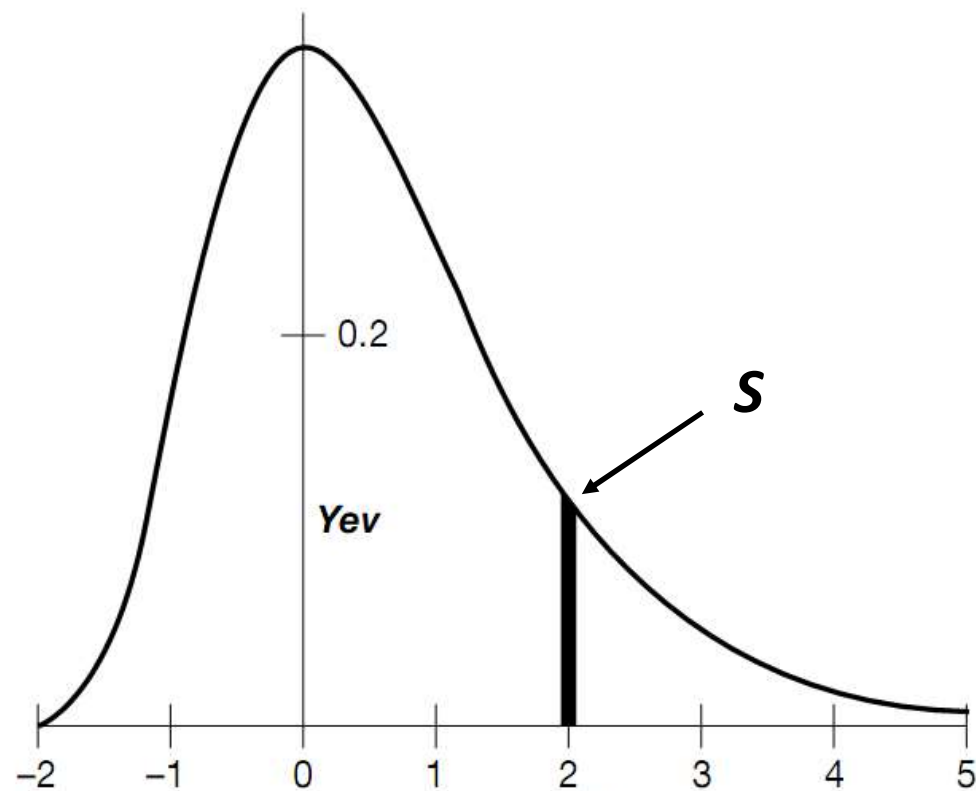
# Distribuição de valor extremo (Gumbel)

$$y = \exp[-x - e^{-x}]$$



$$P(x \geq S)$$

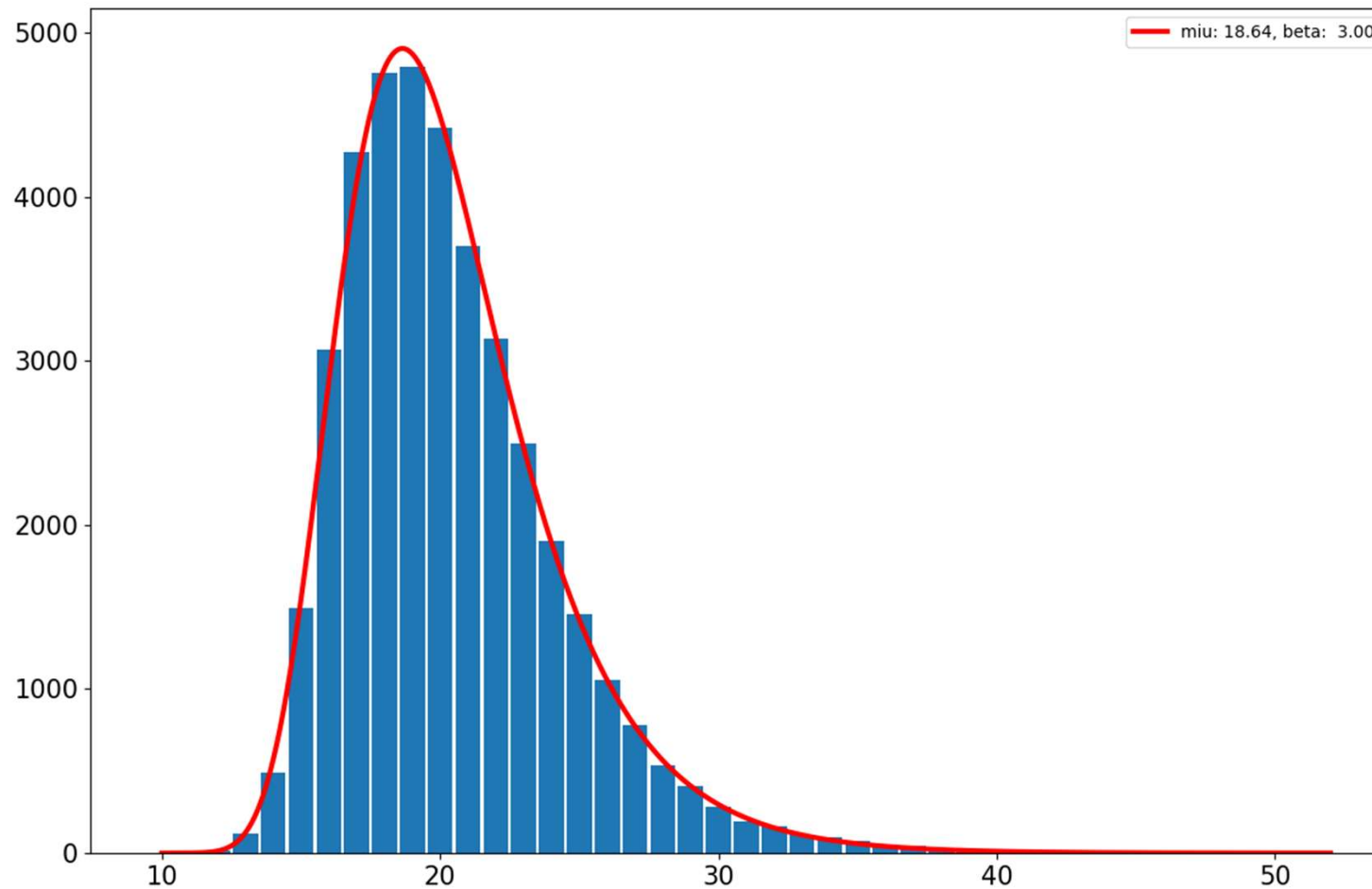
$$P(x \geq S) = 1 - \exp \left[ -e^{\frac{S-\mu}{\beta}} \right]$$



$$P(x \geq S) = 1 - \exp(-K m n e^{-\lambda S})$$

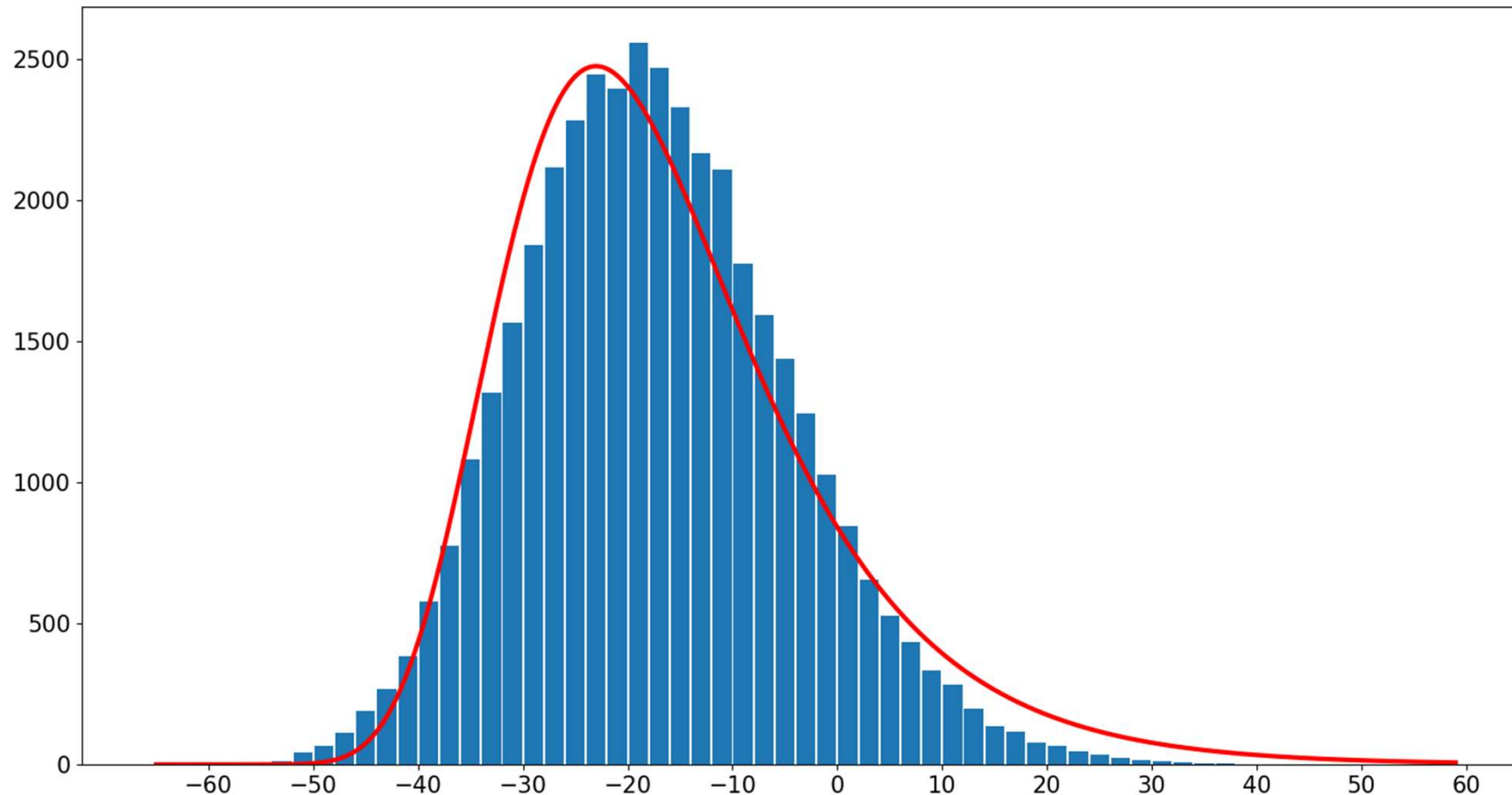
Em que  $m$  e  $n$  são os comprimentos das sequências, e  $K$  e  $\lambda$  são parâmetros que descrevem a distribuição e variam consoante a matriz e o esquema de "gap penalty" usados e também com o comprimento das duas sequências.

# Distribuição de valor extremo: alinhamentos locais com gaps



A distribuição de scores aleatórios ajusta-se perfeitamente a uma distribuição de valor extremo com os valores **miu** e **beta** indicados.

# Distribuição de valor extremo: alinhamentos globais com gaps



A distribuição de scores aleatórios não segue uma curva de valor extremo...

- **P-value**

É a probabilidade de encontrar ao menos um alinhamento aleatório com score  $\geq S$ , dada por:

$$P(s \geq S) = 1 - \exp(-K m n e^{-\lambda S})$$

- **E-value**

corresponde ao número esperado de alinhamentos aleatórios capazes de produzir um score *pelo menos* igual ao score  $S$  do alinhamento original:

$$E = L \times P(s \geq S)$$

em que  $L$  é o número total de alinhamentos gerados.

- **Bit scores**

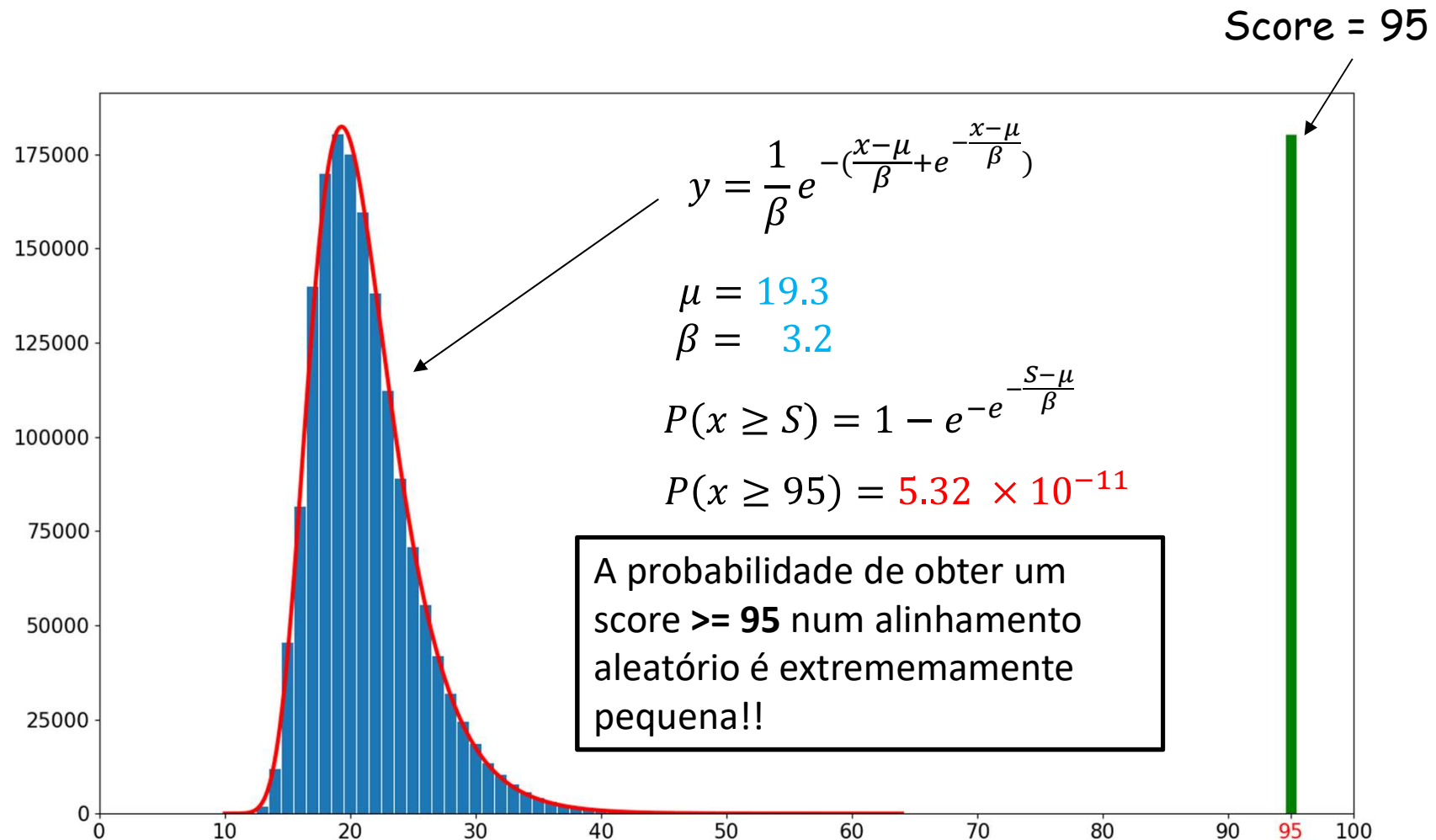
Os bit scores são obtidos normalizando os valores de  $S$  de forma a torná-los independentes dos valores de  $K$  e  $\lambda$ ,

$$S' = (\lambda S - \ln K) / \ln 2$$

ficando os P values, para valores pequenos, simplesmente dados por:

$$P = m n 2^{-S'}$$

# Exemplo: probabilidade de score num alinhamento aleatório



Ajuste de uma curva de distribuição de valor extremo (linha vermelha) a um conjunto de scores de alinhamentos aleatórios (histograma, a azul), sendo o score do alinhamento não-aleatório **igual 95**

# Z-scores

Define-se como **z score** a distância de um determinado valor relativamente à **média** da distribuição, expressa em unidades de **desvio padrão**.

Para o caso da distribuição de valor extremo, a probabilidade de obter um valor  $Z$  superior a um determinado valor  $z$  é dada por:

$$P(Z > z) = 1 - \exp(-e^{-1.285z - 0.5772})$$

# Parâmetros da distribuição de valor extremo

| Scoring matrix | Gap opening penalty <sup>b</sup> | Gap extension penalty <sup>b</sup> | $K$   | $\lambda$ | $H^c$ |
|----------------|----------------------------------|------------------------------------|-------|-----------|-------|
| BLOSUM50       | $\infty^a$                       | 0- $\infty$                        | 0.232 | 0.11      | 0.34  |
| BLOSUM50       | 15                               | 8-15                               | 0.09  | 0.222     | 0.31  |
| BLOSUM50       | 11                               | 8-11                               | 0.05  | 0.197     | 0.21  |
| BLOSUM50       | 11                               | 1                                  | —     | —         | —     |
| BLOSUM62       | $\infty^a$                       | 0- $\infty$                        | 0.318 | 0.13      | 0.40  |
| BLOSUM62       | 12                               | 3-12                               | 0.1   | 0.305     | 0.38  |
| BLOSUM62       | 8                                | 7-8                                | 0.06  | 0.270     | 0.25  |
| BLOSUM62       | 7                                | 1                                  | —     | —         | —     |
| PAM250         | $\infty^a$                       | 0- $\infty$                        | 0.229 | 0.09      | 0.23  |
| PAM250         | 15                               | 5-15                               | 0.06  | 0.215     | 0.20  |
| PAM250         | 10                               | 8-10                               | 0.031 | 0.175     | 0.11  |
| PAM250         | 11                               | 1                                  | —     | —         | —     |



## PRSS3 - evaluates the significance of a protein sequence alignment

|  |   |                  |                                 |
|--|---|------------------|---------------------------------|
| Number of shuffles :   | <input type="text" value="200"/>        | window size:     | <input type="text" value="10"/> |
| Scoring matrix :   | <input type="text" value="default"/>    |                  |                                 |
| gap opening penalty:   | <input type="text" value="2"/>          | tension penalty: | <input type="text" value="2"/>  |
| First sequence title (optional):   | <input type="text"/>                    |                  |                                 |
| Input sequence format  | <input type="text" value="identity"/>   |                  |                                 |
| 1st Query sequence: or ID or AC or GI (see above for valid formats)                | <input type="text"/>                    |                  |                                 |
| Second sequence title (optional):  | <input type="text"/>                    |                  |                                 |
| Input sequence format  | <input type="text" value="Plain Text"/> |                  |                                 |
| 2nd Query sequence: or ID or AC or GI (see above for valid formats)                | <input type="text"/>                    |                  |                                 |
| <input type="button" value="Run PRSS"/> <input type="button" value="Clear Input"/> |   |                  |                                 |

PRSS output for pig trypsin vs. pig elastase - Mozilla Firefox

File Edit View History Bookmarks ScrapBook Tools Help del.icio.us

http://www.ch.embnnet.org/cgi-bin/PRSS3\_form\_parser

Google Calendar Gmail Wiley InterScience: J... Tony Schreiner's We... Prediction of Second... Gmail - Inbox Marés - Portos Princi... Index of /cd oranger View On Black EmoRate Photo Com...

C++ Cprogramming.com Tutorial: Func... elastase in UniProtKB PAM 500 Matrix LALIGN Server PRSS output for pig trypsin v... Gmail - bioinfo

```
# /usr/molbio/bin/prss3 -s /usr/molbio/share/fasta3/default.mat -f -12 -g -2 -v 10 wwwtmp/.PRSS.5883.1.seq wwwtmp/.PRSS.5883.2.seq 200
PRSS evaluates statistical significance using Smith-Waterman
version 3.4t25 Sept 2, 2005
Please cite:
W.R. Pearson (1996) Meth. Enzymol. 266:227-258

wwwtmp/.PRSS.5883.1.seq ->sp|P00761|TRYP_PIG Trypsin precursor (EC 3.4.21.4).[Sus scrofa] 231 aa
vs wwwtmp/.PRSS.5883.2.seq ->sp|P00772|ELA1_PIG (ELA1)Elastase-1 precursor (EC 3.4.21.36).[Sus scrofa] shuffled sequence

    opt      E()
< 20      0      0:
22      0      0:          one = represents 1 library sequences
24      0      0:
26      0      0:
28      0      0:
30      0      0:
32      1      1:*
34      6      3:====
36      7      6:=====
38      9      10:=====
40     16     14:=====
42     15     17:===== *
44     19     18:=====
46     13     19:===== *
48     19     18:=====
50     19     16:=====
52     12     14:===== *
54     11     12:=====
56     14     10:=====
58      8      8:=====
60     11     7:=====
62      4      5:=====
64      4      4:=====
66      1      3:=====
68      2      3:=====
70      1      2:=====
72      2      2:=====
74      3      1:=====
76      1      1:*
78      1      1:*
80      1      1:*
82      0      0:
84      0      0:
86      0      0:
88      0      0:
90      0      0:
92      0      0:
94      0      0:
96      0      0:
98      0      0:
100     0      0:
102     0      0:
```

Downloads postal\_nata\_20... CV-IBrito.pdf Bioinformatica.rar

Done FoxyProxy: Default

PRSS output for pig trypsin vs. pig elastase - Mozilla Firefox

File Edit View History Bookmarks ScrapBook Tools Help deljdo.us

http://www.ch.embnnet.org/cgi-bin/PRSS3\_form\_parser

Google Calendar Gmail Wiley InterScience: J... Tony Schreiner's We... Prediction of Second... Gmail - Inbox Marés - Portos Prind... Index of /cd oranger View On Black EmoRate Photo Com...

C++ Programming.com Tutorial: Func... elastase in UniProtKB PAM 500 Matrix LALIGN Server PRSS output for pig trypsin v... Gmail - bioinfo

```
52 12 11:=====*
54 11 12:=====*
56 14 10:=====*
58 8 8:=====*
60 11 7:=====*
62 4 5:=====*
64 4 4:=====*
66 1 3:=====*
68 2 3:=====*
70 1 2:=====*
72 2 2:=====*
74 3 1:=====*
76 1 1:=====*
78 1 1:=====*
80 1 1:=====*
82 0 0:
84 0 0:
86 0 0:
88 0 0:
90 0 0:
92 0 0:
94 0 0:
96 0 0:
98 0 0:
100 0 0:
102 0 0:
104 0 0:
106 0 0:
108 0 0:
110 0 0:
112 0 0:
114 0 0:
116 0 0:
118 0 0:
>120 0 0:

53200 residues in 200 sequences
(shuffled, win: 10) MLE statistics: Lambda= 0.0690; K=0.001178
Kolmogorov-Smirnov statistic: 0.0280 (N=25) at 60

Smith-Waterman (3.40 March 2004) function [/usr/molbio/share/fastas/default.mat matrix (15:-5)], open/ext: -12/-2
Scan time: 0.290
The best scores are: s-w bits E(200)
sp|P00772|ELA1_PIG (ELA1)Elastase-1 precursor (EC ( 266) 535 63.0 1.4e-12

231 residues in 1 query sequences
53200 residues in 200 library sequences
Scomplib [34t25]
start: Tue Jan 8 18:43:22 2008 done: Tue Jan 8 18:43:23 2008
Scan time: 0.290 Display time: 0.000

Function used was PRSS [version 3.4t25 Sept 2, 2005]
```

Downloads postal\_nata\_20... CV-IBrito.pdf Bioinformatica.rar Clear

Done FoxyProxy: Default