BIOBANCOS

2º Encontro

- *BLAST - Parte 2*

Mulheres em
Bioinformática
& Data Science LA
Promovendo a colaboração entre mulheres

PERGUNTA: BLAST

Qual a diferença entre o Max score, o Total score, o query cover, o E-value, a Identidade?

O que significa cada parâmetro?

PERGUNTA # 3: BLAST

O que precisamos olhar no resultado?

Qual característica é a mais relevante?

O Max score? O Total score? O query cover?

O E-value? A Identidade? Ou todos eles?

Sequences producing significant alignments

Download | New Select columns | Show 100

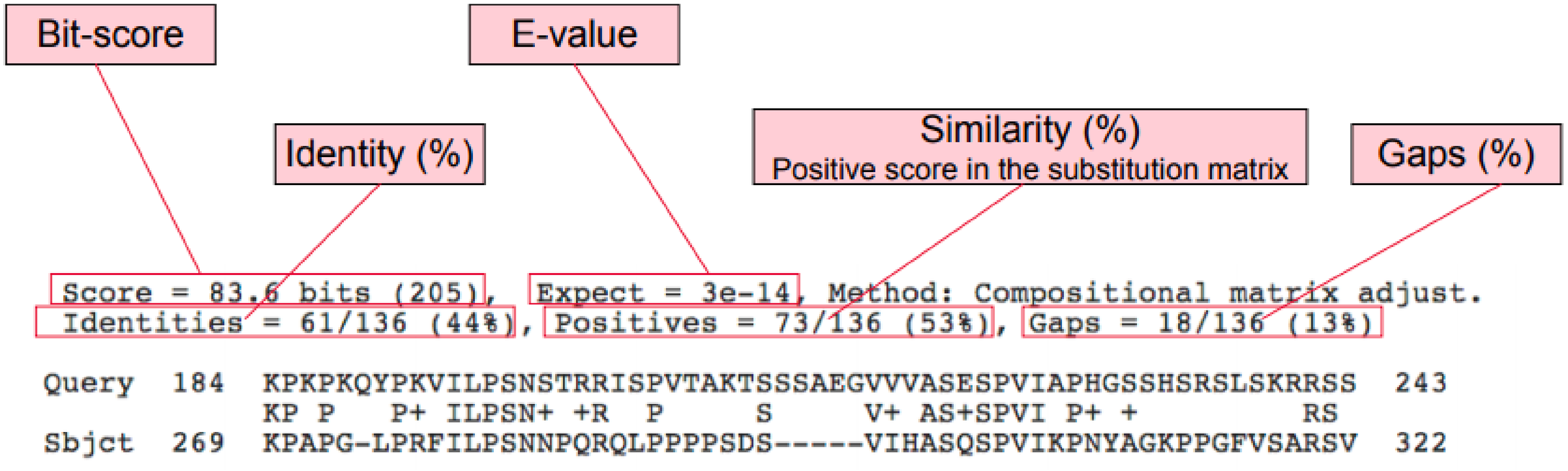
select all *100 sequences selected*

GenBank | Graphics | Distance tree of results | New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Shigella dysenteriae strain ATCC 13313 16S ribosomal RNA, partial sequence	Shigella dysente…	2747	2747	100%	0.0	100.00%	1487	NR_026332.1
Shigella flexneri strain ATCC 29903 16S ribosomal RNA, partial sequence	Shigella flexneri	2669	2669	100%	0.0	99.06%	1488	NR_026331.1
Escherichia fergusonii ATCC 35469 16S ribosomal RNA, complete sequence	Escherichia ferg…	2658	2658	100%	0.0	98.92%	1542	NR_074902.1
Shigella sonnei strain CECT 4887 16S ribosomal RNA, partial sequence	Shigella sonnei	2656	2656	99%	0.0	98.92%	1530	NR_104826.1
Escherichia marmotae strain HT073016 16S ribosomal RNA, partial sequence	Escherichia mar…	2652	2652	100%	0.0	98.86%	1504	NR_136472.1
Escherichia fergusonii ATCC 35469 16S ribosomal RNA, partial sequence	Escherichia ferg…	2627	2627	98%	0.0	98.85%	1473	NR_027549.1
Shigella boydii strain P288 16S ribosomal RNA, partial sequence	Shigella boydii	2615	2615	99%	0.0	98.58%	1515	NR_104901.1
Escherichia coli strain NBRC 102203 16S ribosomal RNA, partial sequence	Escherichia coli	2614	2614	98%	0.0	98.77%	1467	NR_114042.1
Escherichia fergusonii strain NBRC 102419 16S ribosomal RNA, partial sequence	Escherichia ferg…	2612	2612	98%	0.0	98.70%	1467	NR_114079.1
Escherichia albertii strain Albert 19982 16S ribosomal RNA, partial sequence	Escherichia albertii	2593	2593	99%	0.0	98.31%	1494	NR_025569.1
Citrobacter amalonaticus strain CECT 863 16S ribosomal RNA, partial sequence	Citrobacter amal…	2542	2542	100%	0.0	97.52%	1504	NR_104823.1
Escherichia coli strain U 5/41 16S ribosomal RNA, partial sequence	Escherichia coli	2538	2538	96%	0.0	98.40%	1450	NR_024570.1
Kosakonia quasisacchari strain WCHEs120001 16S ribosomal RNA, partial sequence	Kosakonia quasi…	2534	2534	100%	0.0	97.44%	1536	NR_169476.1
Kosakonia sacchari strain SP1 16S ribosomal RNA, partial sequence	Kosakonia sacch…	2529	2529	100%	0.0	97.38%	1500	NR_118333.1
Citrobacter koseri strain LMG 5519 16S ribosomal RNA, partial sequence	Citrobacter koseri	2525	2525	99%	0.0	97.37%	1494	NR_118105.1
Metakosakonia massiliensis JC163 16S ribosomal RNA, partial sequence	Metakosakonia …	2525	2525	100%	0.0	97.31%	1499	NR_125600.1

Example: BLAST - Pho4p (*S. cerevisiae*)

Results (output) of BLAST



Mulheres em Bioinformática & Data Science LA
Promovendo a colaboração entre mulheres

Max Score x Total Score:

- Max Score: Pontuação de alinhamento mais alta calculada a partir da soma dos matchs por nucleotídeos ou aminoácidos correspondentes e penalidades por incompatibilidades e gaps.

- Total Score: A soma das pontuações de alinhamento de todos os segmentos da mesma sequência.

Query Cover:

- A cobertura da consulta é um número que descreve quanto da sequência de consulta é coberta pela sequência de destino;
- Se a sequência de destino no banco de dados abranger toda a sequência de consulta, a cobertura da consulta será 100%;
- Nos diz quão iguais as sequências são em relação umas às outras.

Identity:

- A porcentagem de identidade é um número que descreve a semelhança da sequência de consulta com a sequência de destino (quantos caracteres em cada sequência são idênticos);
- Quanto maior a porcentagem de identidade, mais significativa é a correspondência.

E-Value (Expeted Value = Valor Esperado):

- É um número que descreve quantas vezes esperaríamos um match por acaso em um banco de dados desse tamanho;
- Quanto menor o E-value,
mais significativa é a correspondência.

PERGUNTA: BLAST

Qual a diferença entre o Max score, o Total score, o query cover, o E-value, a Identidade?

O que significa cada parâmetro?

E-value x P-Value:

$$P(S > s) = 1 - \exp[-KMNe^{-\lambda s}]$$
$$E(S > s) = KMNe^{-\lambda s}$$

M: Tamanho da query

N: Tamanho do subject

x: Score

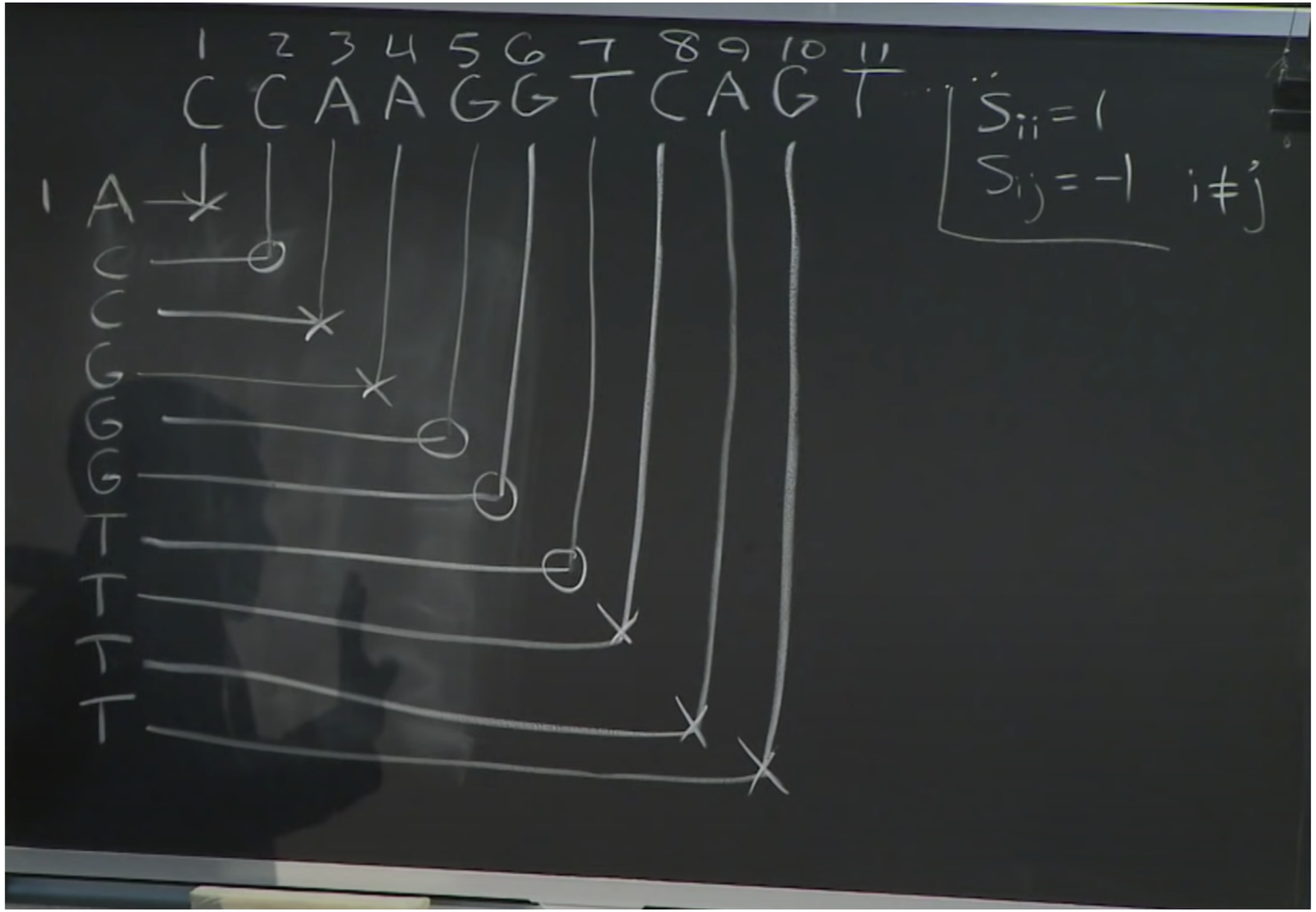
K e Lambda: Parâmetros positivos que dependem da composição da matriz e das sequências

E-value x P-Value:

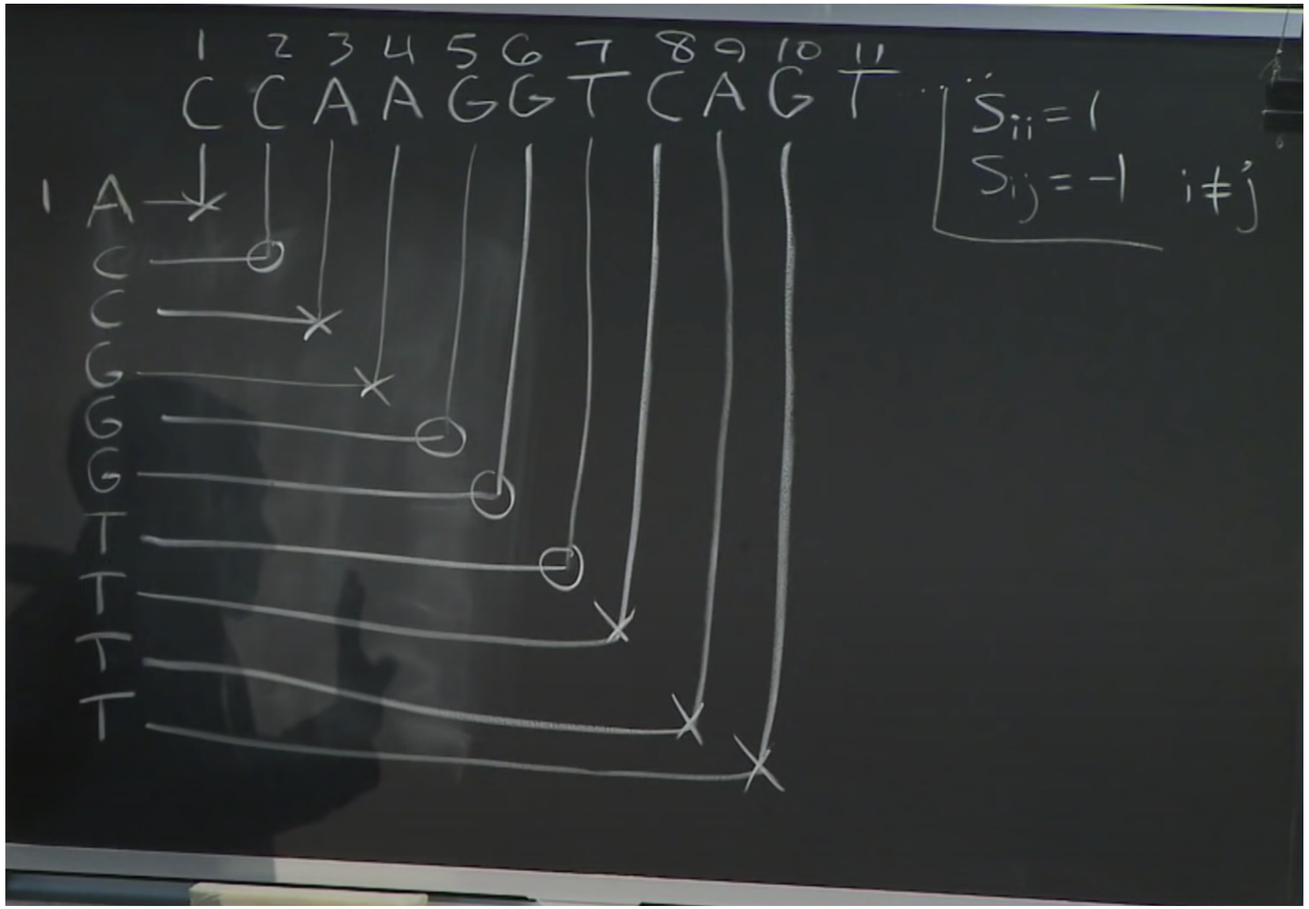
$$P(S > s) = 1 - \exp[-KMNe^{-\lambda s}]$$
$$E(S > s) = KMNe^{-\lambda s}$$

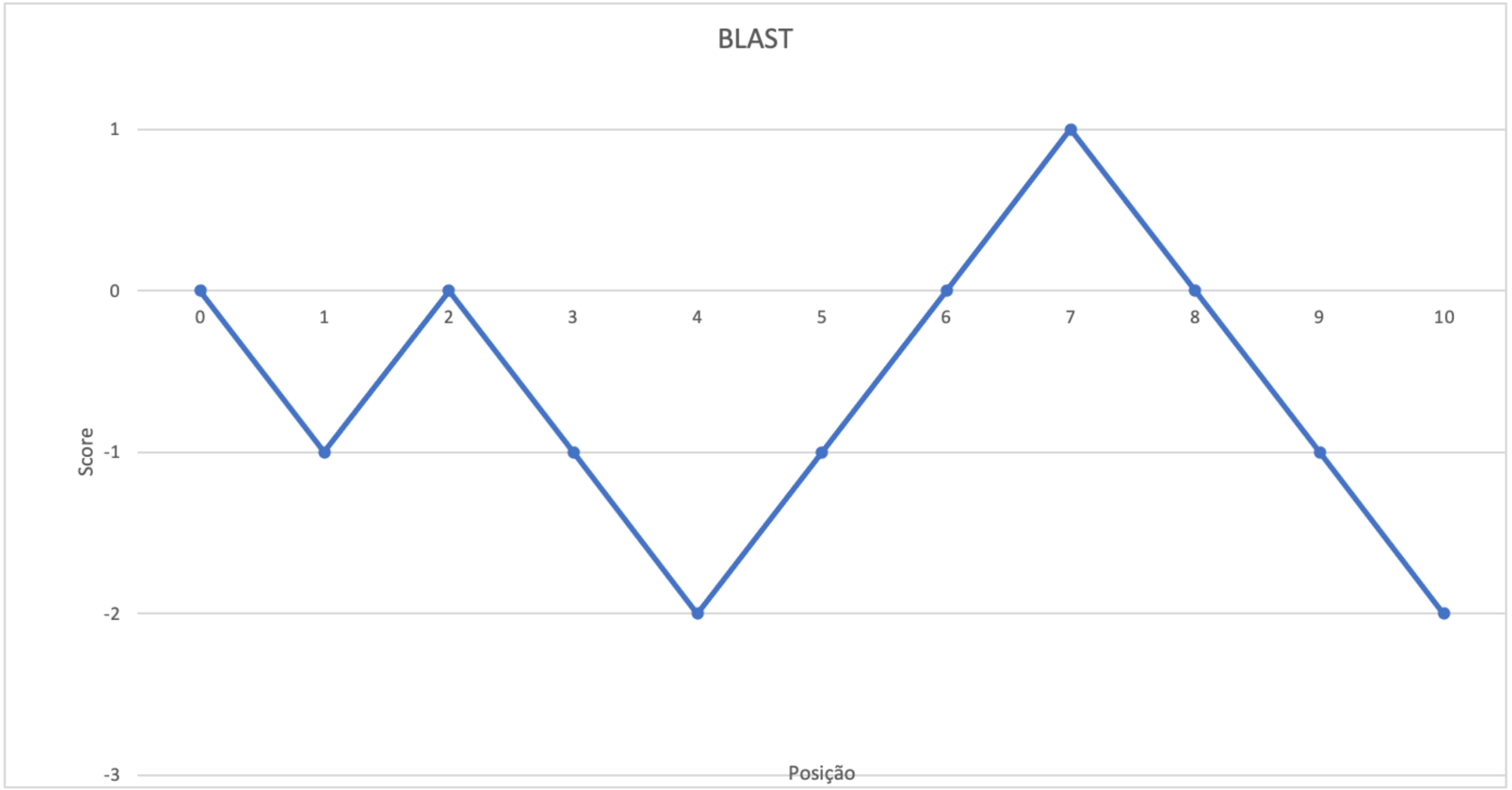
O BLAST mostra o valor E em vez de P porque é mais fácil entender a diferença entre, por exemplo, um valor E de 5 e 10 do que valores de P de 0,993 e 0,99995.

No entanto, quando E < 0,01, os valores P e o valor E são quase idênticos.

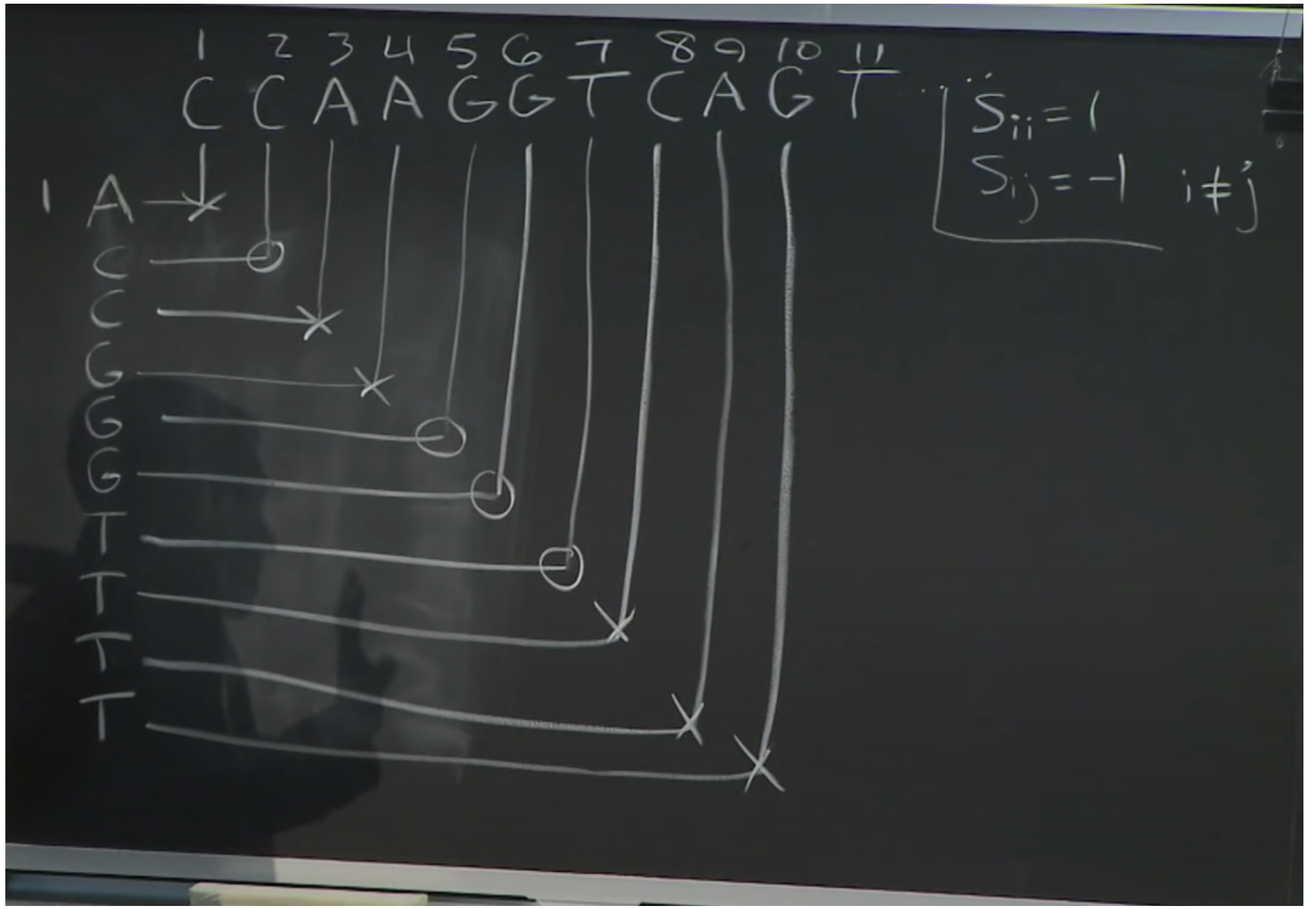


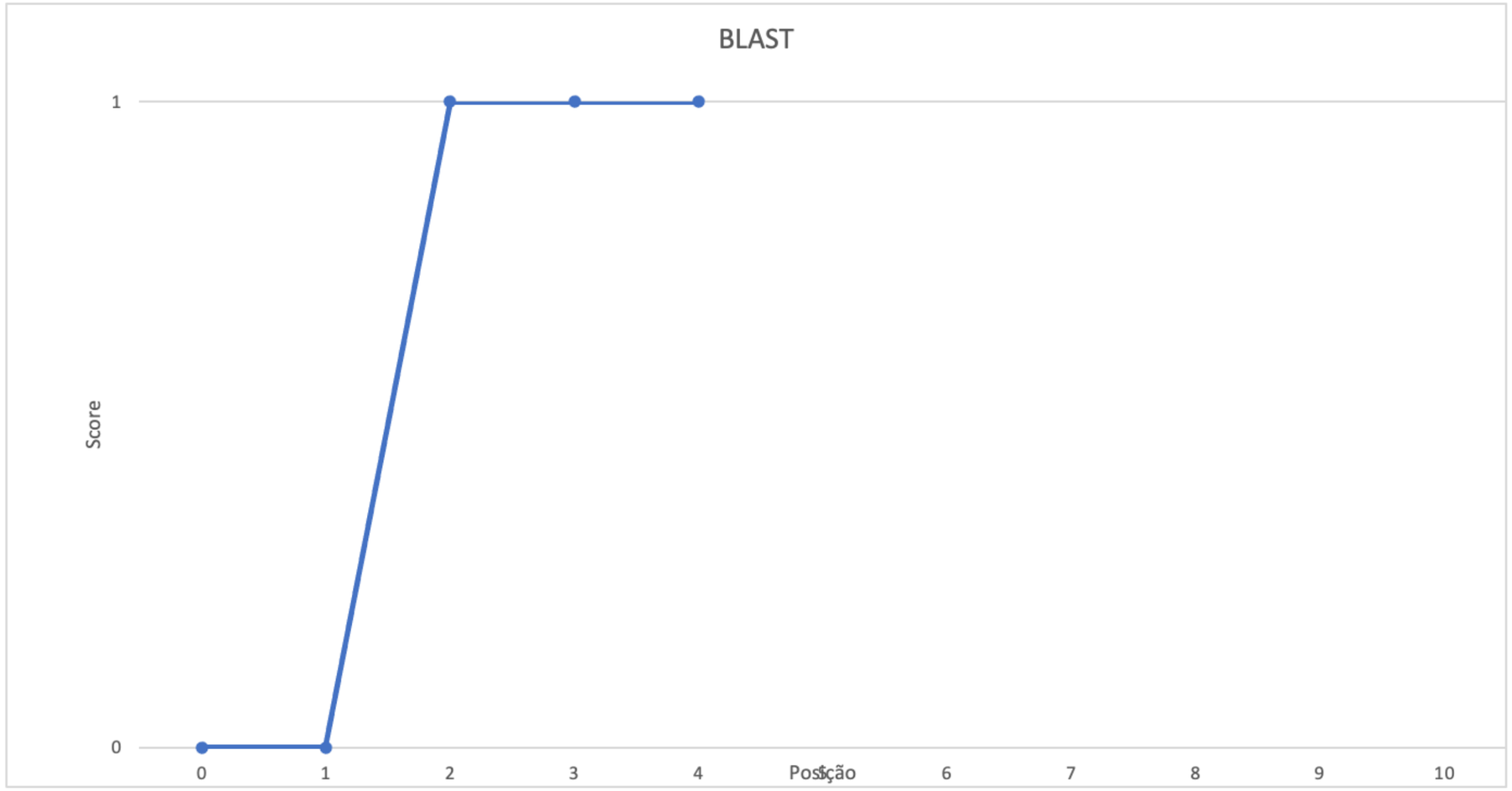
Posição	Classificação	Score	Score acumulado	
0	Nenhuma	0	0	
1	Mismatch	-1	-1	
2	Match	1	0	
3	Mismatch	-1	-1	
4	Mismatch	-1	-2	Score_min
5	Match	1	-1	
6	Match	1	0	
7	Match	1	1	Score_max
8	Mismatch	-1	0	
9	Mismatch	-1	-1	
10	Mismatch	-1	-2	





Posição	Classificação	Score	Score acumulado	
0	Nenhuma	0	0	
1	Mismatch	0	0	
2	Match	1	1	
3	Mismatch	0	1	
4	Mismatch	0	1	Max_inf
5	Match	1	2	
6	Match	1	3	
7	Match	1	4	Max_sup
8	Mismatch	0	4	
9	Mismatch	0	4	
10	Mismatch	0	4	





Score, Bit-score, P-value, E-value

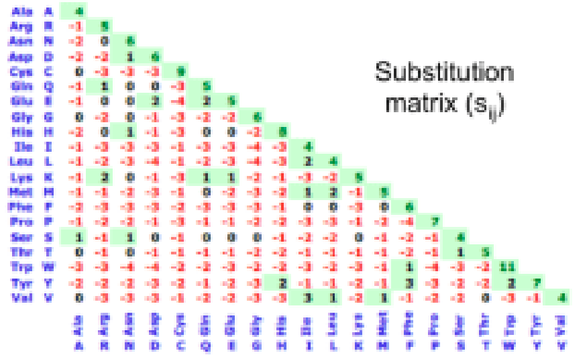
Score: A number used to assess the biological relevance of a finding.

In the context of sequence alignments, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used (substitution matrix, gap penalty).

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

Example:

```
R  L  A  S  V  -  E  T  D  M  W  T  P  L  T  L  R  Q  H
.  |  .  |  :     :  |  .  :        .  |  .  .  |
T  L  T  S  L  A  Q  T  T  L  -  -  K  A  H  L  G  T  H
-1 +4 +0 +4 +1 -4 +2 +5 -1 +2 -4 -1 -1 -1 -2 +4 -2 -1 +8 = 12
```



Substitution matrix (s_{ij})

gap penalty (s_{i-})

gap opening	-4
gap extension	-1
end gap	0