BLAST

Basic Local Alignment Search Tool

SCC02713 - Introdução à Bioinformática

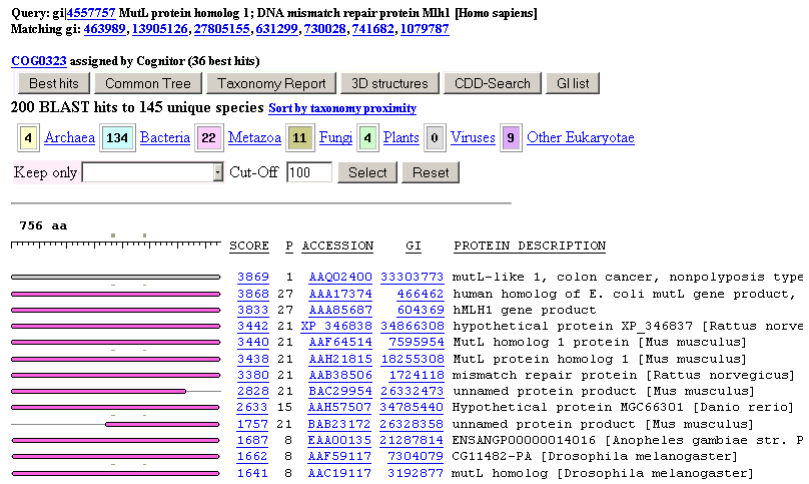
https://www.youtube.com/watch?v=8A-msg23u0w&list=PLpPXw4zFa0uLMHwSZ7DMeLGjIUgo1IBbn&index=30
https://www.youtube.com/watch?v=Ge00nXqIIow
https://www.youtube.com/watch?v=uGhZygAMQik

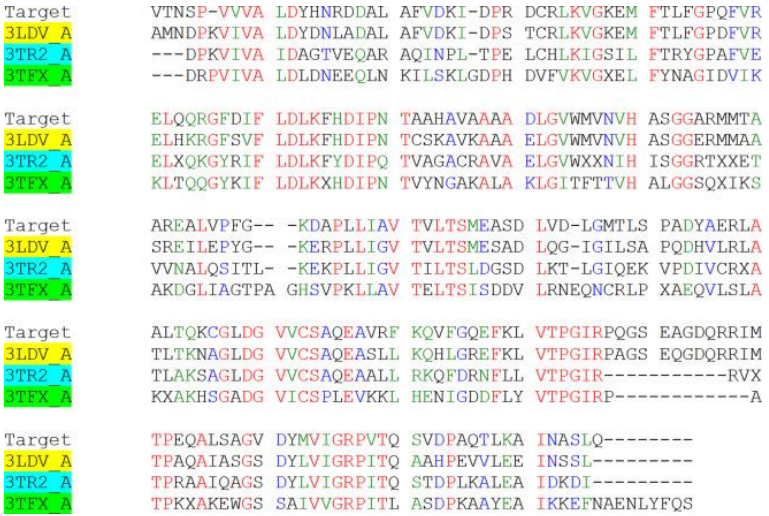
Alinhamento de Sequências

gaps and mismatches exist

Sequence alignment algorithms:

- E se quisermos alinhar várias sequências?
  - Uma contra muitas
  - Muitas contra muitas

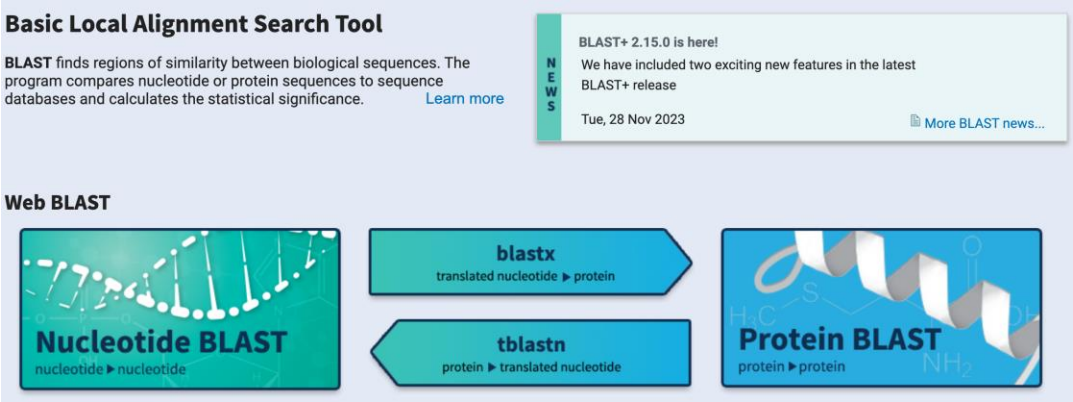




https://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blink.html
https://www.sciencedirect.com/science/article/pii/B9780128096338204846

Alinhamento de Sequências

- BLAST - Basic Local Alignment Search Tool



jmb Journal of Molecular Biology
Volume 215, Issue 3, 5 October 1990, Pages 403-410

Basic local alignment search tool

Stephen F. Altschul [1], Warren Gish [1], Webb Miller [2], Eugene W. Myers [3], David J. Lipman [1]

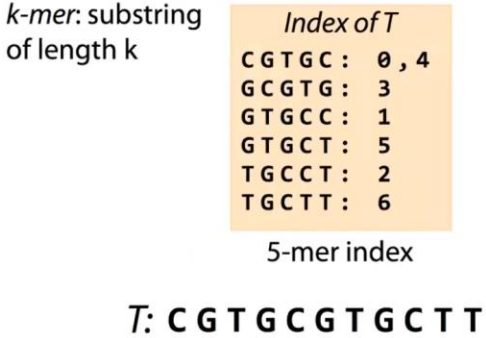
[1] National Center for Biotechnology Information National Library of Medicine, National Institutes of Health Bethesda, MD 20894, U.S.A.
[2] Department of Computer Science The Pennsylvania State University, University Park, PA 16802, U.S.A.
[3] Department of Computer Science University of Arizona, Tucson, AZ 85721, U.S.A.

Received 26 February 1990, Accepted 15 May 1990, Available online 6 February 2007.

https://blast.ncbi.nlm.nih.gov/Blast.cgi
https://www.sciencedirect.com/science/article/pii/S0022283605803602?via%3Dihub

Alinhamento de Sequências

- BLAST - Heurística
  - Quebra as sequências em palavras (k-mers) e faz hash de suas localizações para acelerar pesquisas posteriores

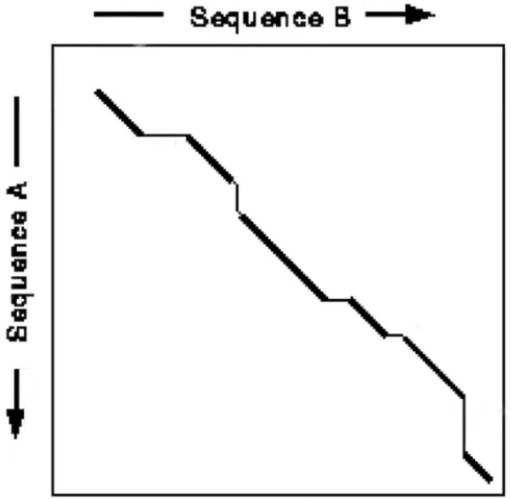


  - Para cada k-mer na query, encontre os k-mers no banco de dados que têm bom match
  - Somente k-mers com uma pontuação de alinhamento maior que um limiar são mantidas

https://www.youtube.com/watch?v=Ge00nXqIlow

Alinhamento de Sequências

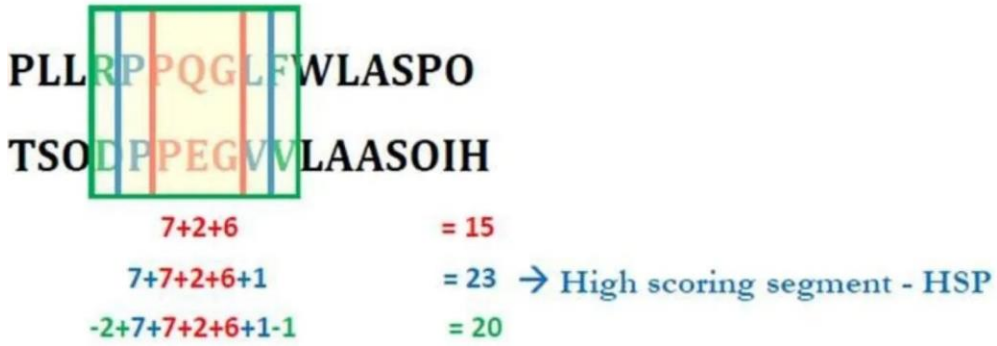
- BLAST - Heurística
  - Para cada trecho de alta pontuação no banco de dados, tenta estender o alinhamento para ambos os lados
  - Forma HSP (High-scoring Segment Pairs)
  - Mantém apenas os HPS estatisticamente significantes
    - Baseado no score de alinhar duas sequências aleatórias
  - Usa o algoritmo Smith-Waterman para unir os HSP e obter o alinhamento final ótimo



https://www.youtube.com/watch?v=Ge00nXqIlow

Alinhamento de Sequências

- BLAST – Seed and Extend
  - BLAST 1: Estende os k-mers para a esquerda e para a direita usando alinhamentos sem gaps
  - A extensão continua enquanto a pontuação não cair abaixo de um limiar determinado
  - BLAST 2: Estende os HSPs usando alinhamento com gaps

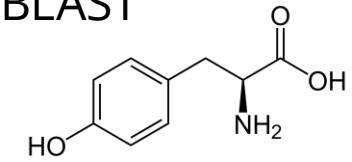


https://www.youtube.com/watch?v=Ge00nXqIlow
https://medium.com/@gokcenazakyol/biological-sequence-analysis-ii-multiple-sequence-alignment-alignment-algorithms-tools-6dac6d87119a

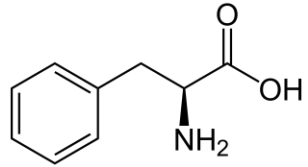
Alinhamento de Sequências

- BLAST – BLOSUM62 – Matriz padrão do BLAST
  - Phe -> Phe = +6
  - Tyr -> Phe = +3
  - Tyr -> Tyr = +7
  - Tyr -> Asp = -3

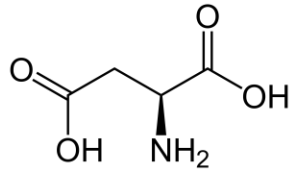
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4



Tirosina (TAT ou TAC)



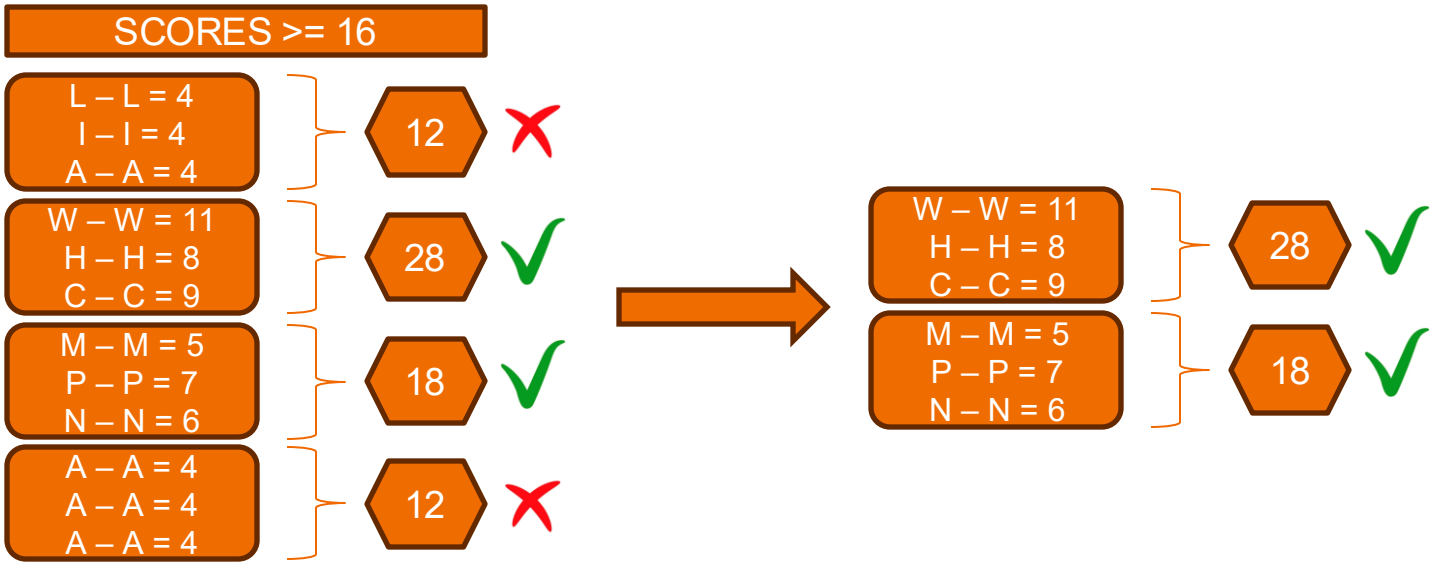
Fenilalanina (TTT ou TTC)



Ácido aspártico (GAT ou GAC)

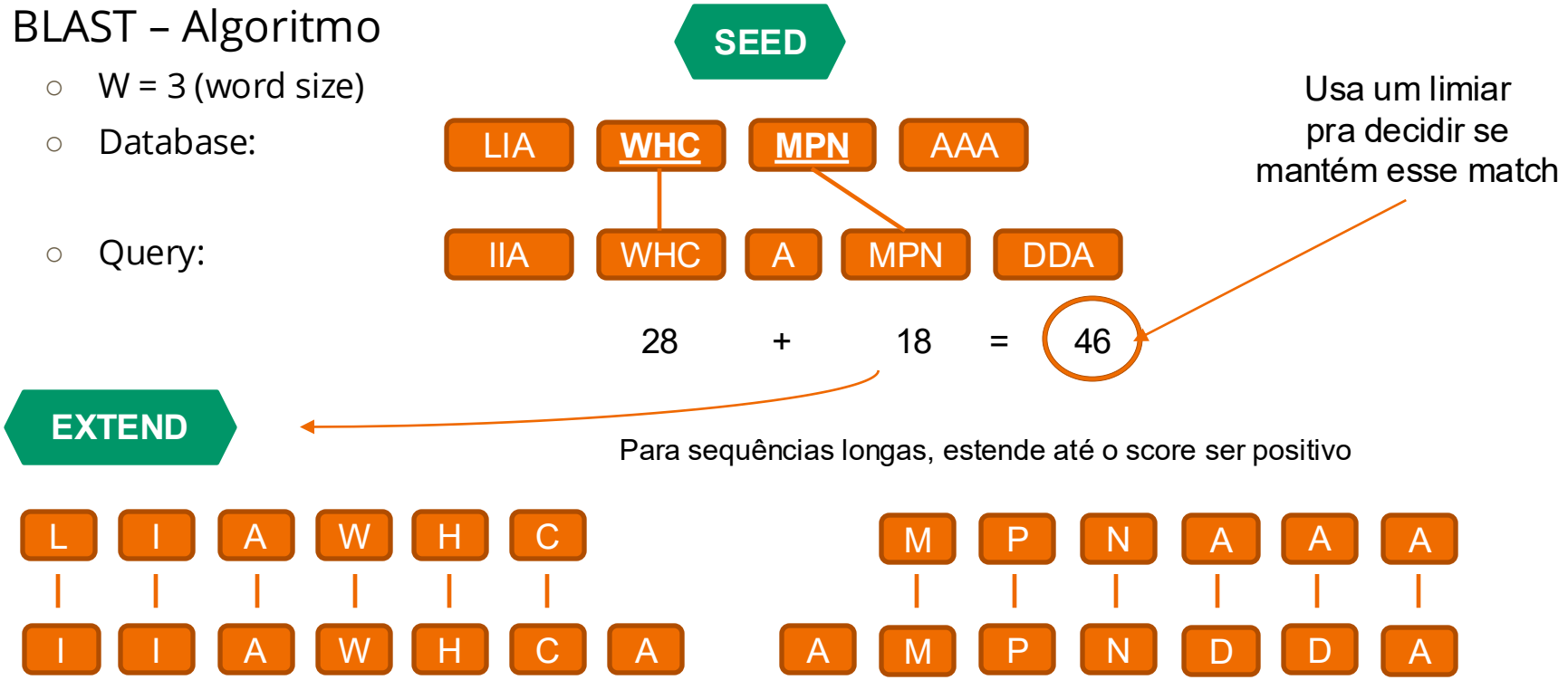
Alinhamento de Sequências

- BLAST – Algoritmo
  - W = 3 (word size)
  - Database: LIA WHC MPN AAA



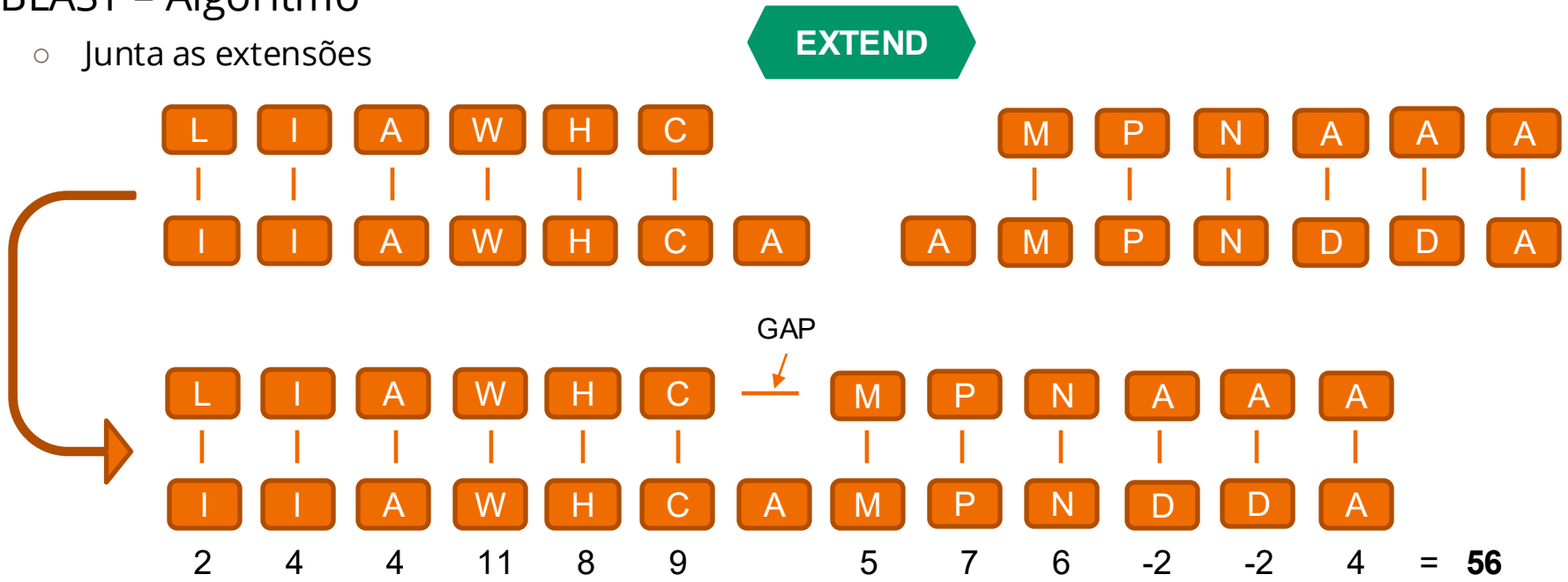
Alinhamento de Sequências

- BLAST – Algoritmo
  - W = 3 (word size)
  - Database:
  - Query:



Alinhamento de Sequências

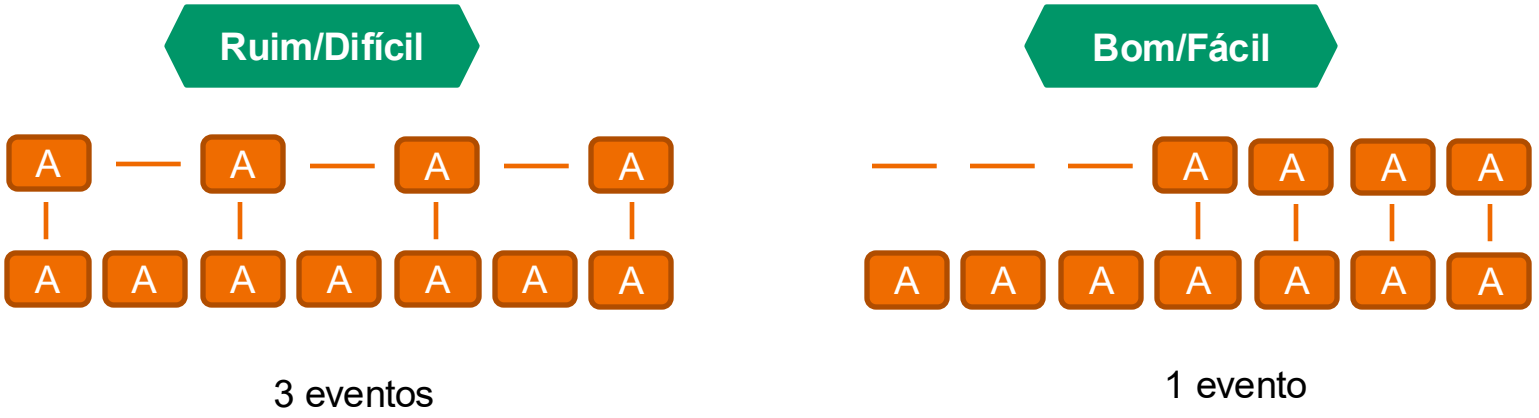
- BLAST – Algoritmo
  - Junta as extensões



Alinhamento de Sequências

- BLAST – Algoritmo
  - Mas e os gaps?
    - Duas penalizações: *gap opening* e *gap extension*

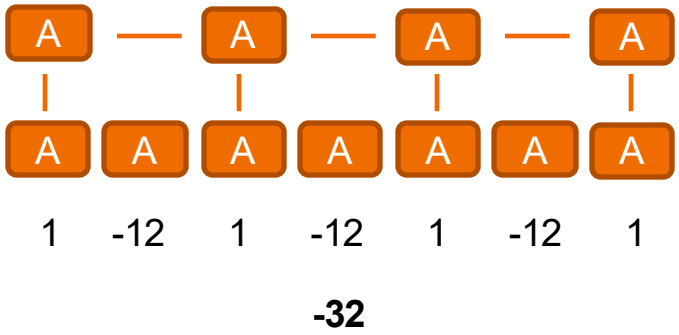
Gap opening: inicia o gap
Gap extension: penalização do gap

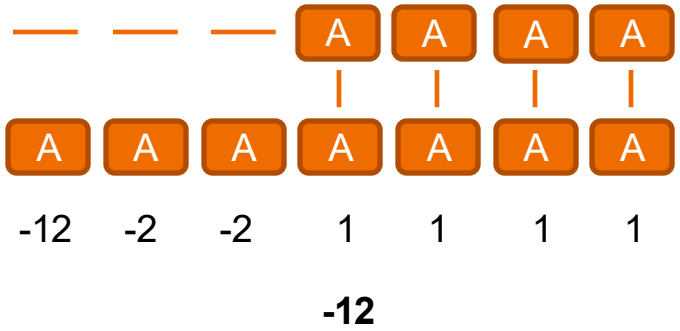


Alinhamento de Sequências

- BLAST – Algoritmo
  - Mas e os gaps?
    - Duas penalizações: *gap opening* e *gap extension*

Gap opening: inicia o gap = -10
Gap extension: penalização do gap = -2
Suponha A:A = 1



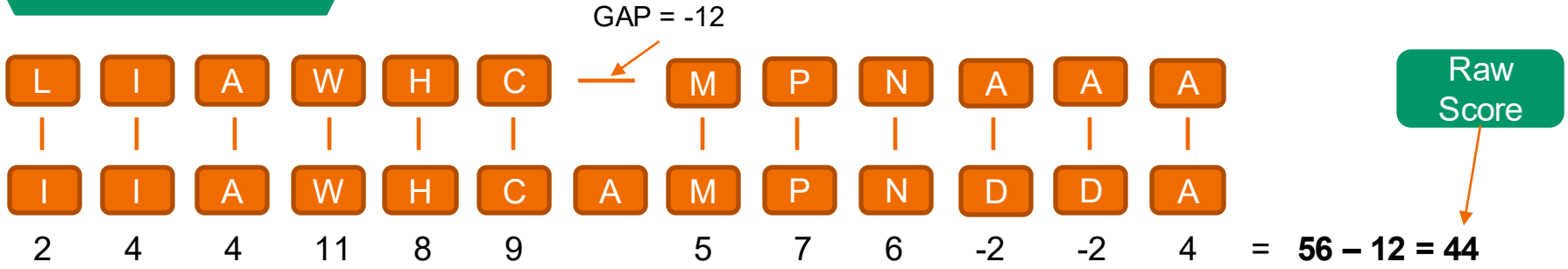


Alinhamento de Sequências

- BLAST – Algoritmo
  - Mas e os gaps?
    - Duas penalizações: *gap opening* e *gap extension*

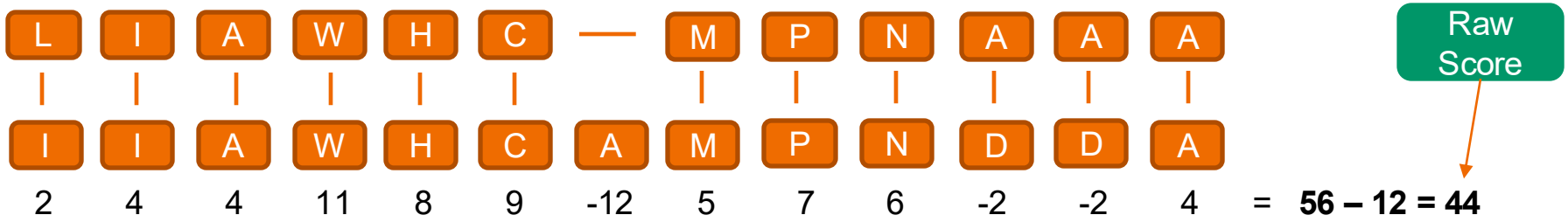
Gap opening: inicia o gap = -10
Gap extension: penalização do gap = -2
Suponha A:A = 1

Exemplo inicial



Alinhamento de Sequências

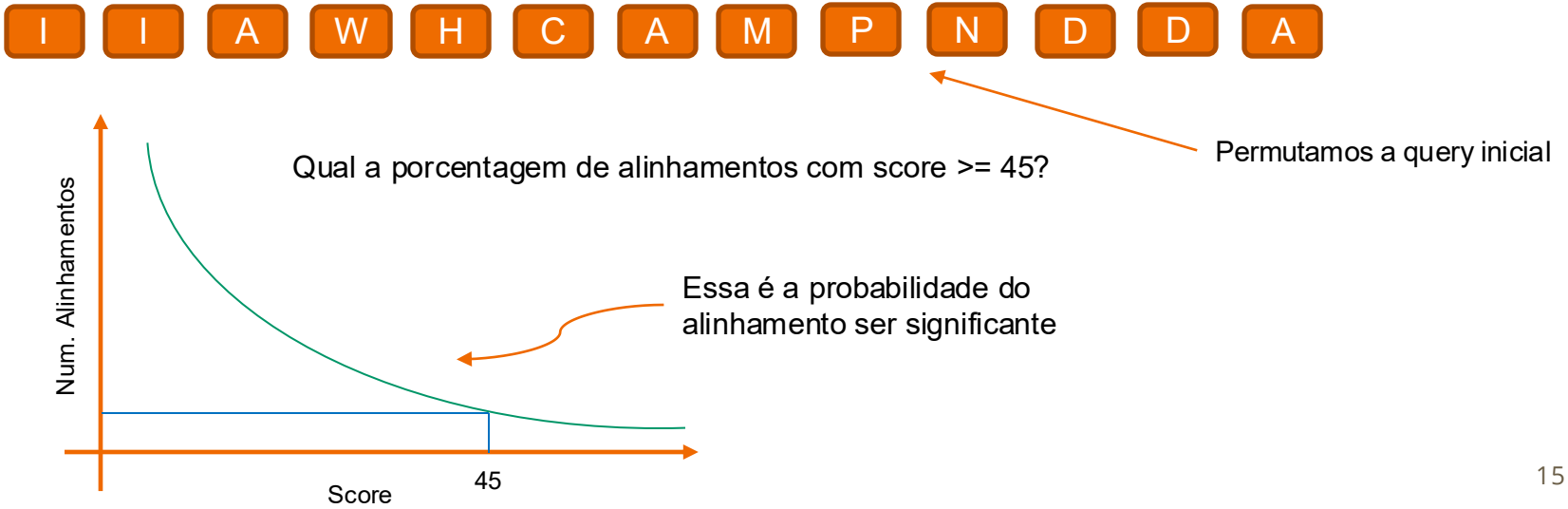
- BLAST – Algoritmo
  - Como saber se o alinhamento é significante?



  - Qual a chance de termos encontrado nosso alinhamento por pura sorte?
    - Em bancos de dados enormes, essa chance aumenta!
    - E a chance de encontrarmos os matches exatos das palavras (w) aumenta!

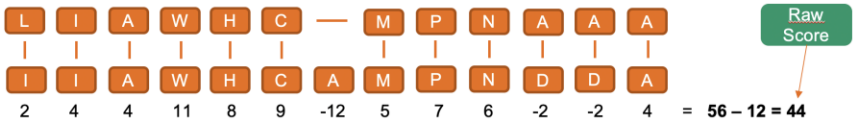
Alinhamento de Sequências

- BLAST – Algoritmo
  - Como saber se o alinhamento é significante?
  - O BLAST é altamente paralelizável. Portanto podemos fazer permutações da nossa query e fazer buscas usando essas permutações



Alinhamento de Sequências

- BLAST – Algoritmo
  - Como saber se o alinhamento é significante?



  - O raw score (S) é um número que depende apenas do alinhamento
  - Devemos converter o score para um bit score:

$$S' = \frac{\lambda S - \ln(k)}{\ln(2)}$$

    - λ e k são dependentes da matriz de pontuação usada e das penalizações de gap usadas
  - Queremos normalizar o score S baseado também no tamanho da sequência query e no tamanho da base de dados

Tamanho da base de dados

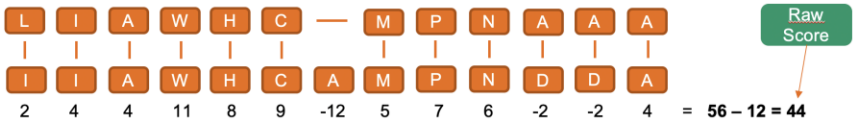
Tamanho da query

  - Expectation value = $E = k\,m\,n\,e^{-\lambda S}$

https://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html

Alinhamento de Sequências

- BLAST – Algoritmo
  - Como saber se o alinhamento é significante?



  - Queremos normalizar o score S baseado também no tamanho da sequência query e no tamanho da base de dados
  - Expectation value =

Tamanho da base de dados

Tamanho da query

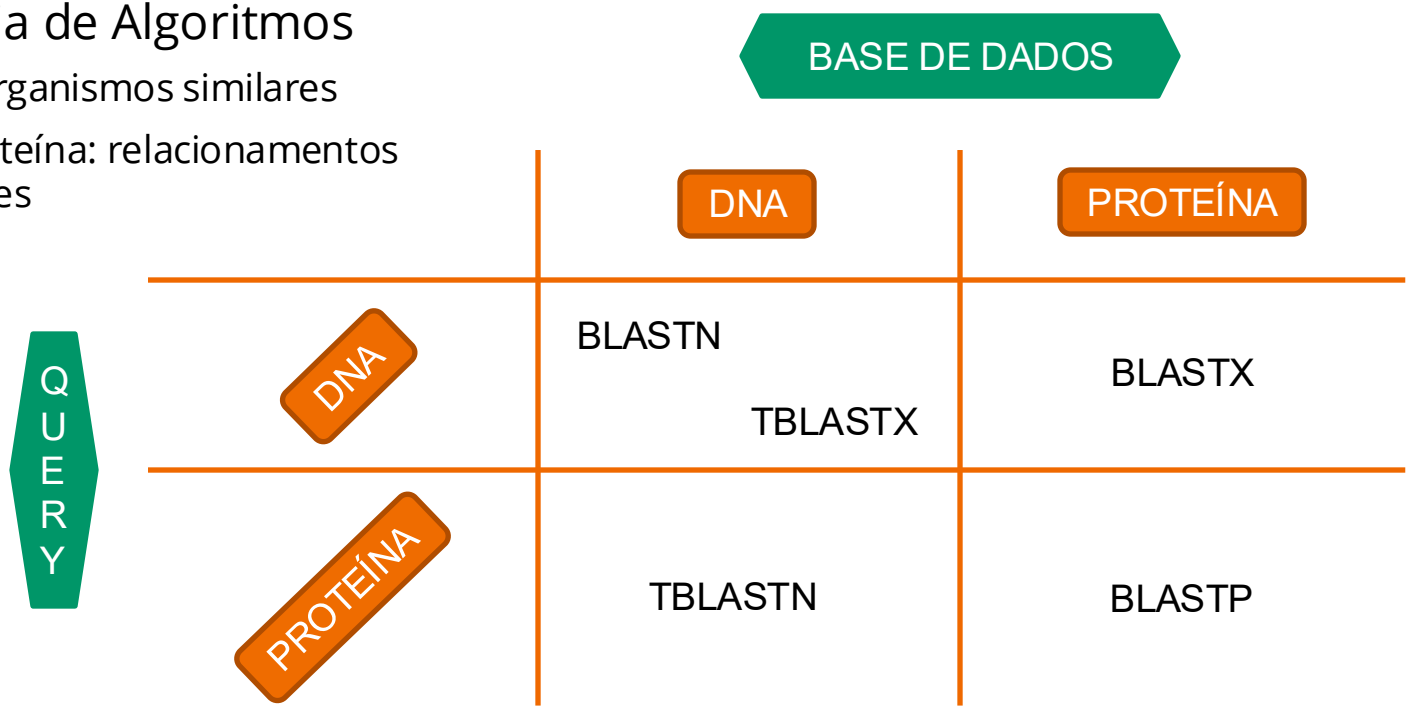
$$E = k\,m\,n\,e^{-\lambda S}$$

	E-value
Proteínas idênticas	$< 10^{-100}$
Quase idênticas	10^{-50} - 10^{-100}
Proteínas homólogas	10^{-5} - 10^{-50}
Homólogos distantes	10^{-1} - 10^{-5}
Talvez aleatória	$> 10^{-1}$

https://www.youtube.com/watch?v=1l2p_Q7QyB4&list=PLpPXw4zFa0uLMHwSZ7DMeLGjlUgo1lBbn&index=39

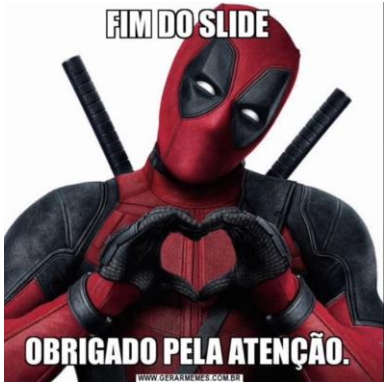
Alinhamento de Sequências

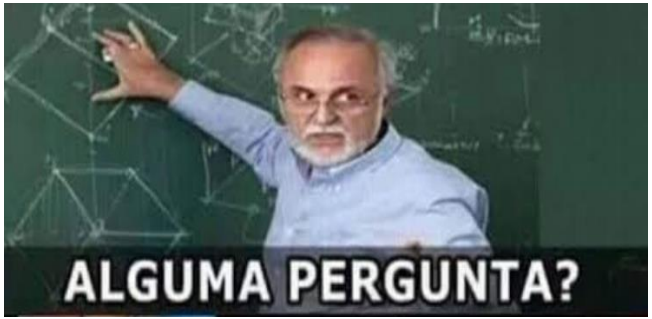
- BLAST – Família de Algoritmos
  - DNA-DNA: organismos similares
  - Proteína-Proteína: relacionamentos mais distantes



https://www.youtube.com/watch?v=1I2p_Q7QyB4&list=PLpPXw4zFa0uLMHwSZ7DMeLGjIUgo1IBbn&index=39

É isso aí!


FIM DO SLIDE
OBRIGADO PELA ATENÇÃO.
WWW.GERARMEMES.COM.BR


ALGUMA PERGUNTA?