

---

# BLAST

## Basic Local Alignment Search Tool

— SCC02713 - Introdução à Bioinformática —

---

# Alinhamento de Sequências

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

\*gaps and mismatches exist\*

Sequence alignment algorithms:



Query: gi|4557757 MutL protein homolog 1; DNA mismatch repair protein Mlh1 [Homo sapiens]  
Matching gi: 463989, 13905126, 27805155, 631299, 730028, 741682, 1079787

COG0323 assigned by Cognitor (36 best hits)

Best hits Common Tree Taxonomy Report 3D structures CDD-Search Gist list

200 BLAST hits to 145 unique species [Sort by taxonomy proximity](#)

4 Archaea 134 Bacteria 22 Metazoa 11 Fungi 4 Plants 0 Viruses 9 Other Eukaryotes

Keep only [ ] Cut-Off 100 Select Reset

	SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
756 aa					
3869	1	AAQ02400	33303773		mutL-like 1, colon cancer, nonpolyposis type
3868	27	AAA17374	466462		human homolog of E. coli mutL gene product,
3833	27	AA085687	604369		hMLH1 gene product
3442	21	XP_346838	34866308		hypothetical protein XP_346837 [Rattus norve
3440	21	AAF64514	7595954		MutL homolog 1 protein [Mus musculus]
3438	21	AAH21815	18255308		MutL protein homolog 1 [Mus musculus]
3380	21	AA038506	1724118		mismatch repair protein [Rattus norvegicus]
2828	21	BAC29954	26332473		unnamed protein product [Mus musculus]
2633	15	AAH57507	34785440		Hypothetical protein MG066301 [Danio rerio]
1757	21	BAB23172	26328358		unnamed protein product [Mus musculus]
1687	8	EAA00135	21287814		ENSANGP00000014016 [Anopheles gambiae str. P
1662	8	AAF59117	7304079		CG11482-PA [Drosophila melanogaster]
1641	8	AAC19117	3192877		mutL homolog [Drosophila melanogaster]

Target  
3LDV A  
3TR2 A  
3TFX A

VTNSP-VVVA LDYHNRDDAL AFVDKI-DPR DCRLKVGKEM FTLFGPQFVR  
AMNDPKVIVA LDYDNLADAL AFVDKI-DPS TCRLKVGKEM FTLFGPDFVR  
---DPKVIVA IDAGTVEQAR AQINPL-TPE LCHLKIGSIL FTRYGPAFVE  
---DRPVIVA LDLDNEEQIN KILSKLGDPH DVFKVVGXEL FYNAGIDVVK

Target  
3LDV A  
3TR2 A  
3TFX A

ELQQRGFIDF LDLKFHDIPN TAAHAVAAAA DLGVWMNVNH ASGGARMMTA  
ELHKRGFSVF LDLKFHDIPN TCSKAVKAAA ELGVWMNVNH ASGGERMMAA  
ELQKQGYRIF LDLKFYDIPQ TVAGACRAVA ELGVWXXNIH ISGGRTXXET  
KLTTQQGYKIF LDLKXHDIPN TVYNGAKALA KLGITFTTVH ALGGSQXIKS

Target  
3LDV A  
3TR2 A  
3TFX A

AREALVPFG-- -KDAPLLIIV TVLTSMEASD LVD-LGMTLS PADYAEERLA  
SREILEPYG-- -KERPLLIGV TVLTSMESAD LQG-IGILSA PQDHVLRRLA  
VVNALQSITL- -KEKPLLIGV TILTSLDGSD LKT-LGIEQK VPDIVCRXA  
AKDGLIAGTPA GHSVPKLLAV TELTSSISDDV LRNEQNCRLP XAEQVLSLA

Target  
3LDV A  
3TR2 A  
3TFX A

ALTQKCGLDG VVCSAQEAVERF KQVFGQEFKL VTPGIRPQGS EAGDQRRIM  
TLTKNAGLDG VVCSAQEASLL KQHLGREFKL VTPGIRPAGS EQGDQRRIM  
TLAKSAGLDG VVCSAQEAALL RKQFDRNFL VTPGIR-----RVX  
KXAKHSGADG VICSPLEVKKL HENIGDDFLY VTPGIRP-----A

Target  
3LDV A  
3TR2 A  
3TFX A

TPEQALSAGV DYMVIGRPVQTQ SVDPAQTLKA INASLQ-----  
TPAQAIASGS DYLVIGRPITQ AAHPEVVEE INSSL-----  
TPRAAIQAGS DYLVIGRPITQ STDPLKALEA IDKDI-----  
TPKAKEWGS SAIVVGRPITL ASDPKAAYEA IKKEFNAENLYFQS

# Alinhamento de Sequências

- BLAST - Basic Local Alignment Search Tool

## Basic Local Alignment Search Tool

**BLAST** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

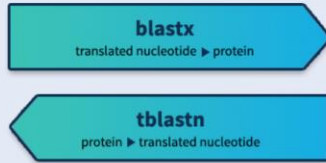
BLAST+ 2.15.0 is here!

We have included two exciting new features in the latest BLAST+ release

Tue, 28 Nov 2023

[More BLAST news...](#)

## Web BLAST



jmb  
Journal of Molecular Biology

Volume 215, Issue 3, 5 October 1990, Pages 403-410



## Basic local alignment search tool

Stephen F. Altschul<sup>1</sup>, Warren Gish<sup>1</sup>, Webb Miller<sup>2</sup>, Eugene W. Myers<sup>3</sup>, David J. Lipman<sup>1</sup>

- <sup>1</sup> National Center for Biotechnology Information National Library of Medicine, National Institutes of Health Bethesda, MD 20894, U.S.A.
- <sup>2</sup> Department of Computer Science The Pennsylvania State University, University Park, PA 16802, U.S.A.
- <sup>3</sup> 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.



# Alinhamento de Sequências

- BLAST - Heurística

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

*k-mer*: substring  
of length *k*

<i>Index of T</i>	
CGTGC:	0, 4
GCGTG:	3
GTGCC:	1
GTGCT:	5
TGCCT:	2
TGCTT:	6

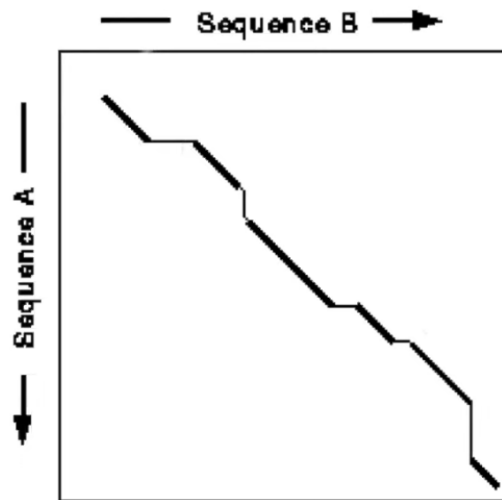
5-mer index

***T*: CGTGCGTGCTT**

- 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

# Alinhamento de Sequências

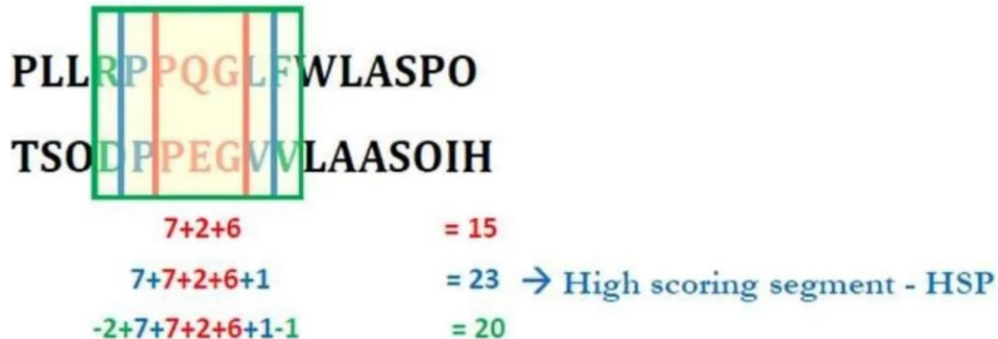
- BLAST - Heurística
  - Para cada trecho de alta pontuação no banco de dados, tenta estender o alinhamento para ambos os lados
  - Forma HSP (High-scoring Segment Pairs)
  - Mantém apenas os HPS estatisticamente significantes
    - Baseado no score de alinhar duas sequências aleatórias
  - Usa o algoritmo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



# 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



# Alinhamento de Sequências

- BLAST – BLOSUM62 – Matriz padrão do BLAST

- Phe → Phe = +6
- Tyr → Phe = +3
- Tyr → Tyr = +7
- Tyr → Asp = -3

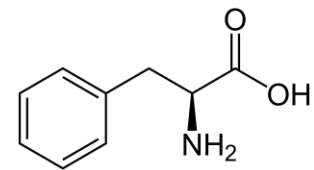
Ala 4  
Arg -1 5  
Asn -2 0 6  
Asp -2 -2 1 6  
Cys 0 -3 -3 -3 9  
Gln -1 1 0 0 -3 5  
Glu -1 0 0 2 -4 2 5  
Gly 0 -2 0 -1 -3 -2 -2 6  
His -2 0 1 -1 -3 0 0 -2 8  
Ile -1 -3 -3 -3 -1 -3 -3 -4 -3 4  
Leu -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4  
Lys -1 2 0 -1 -3 1 1 -2 -1 -3 -2 5  
Met -1 -1 -2 -3 -1 0 -2 -3 -2 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 7  
Val 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4

Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

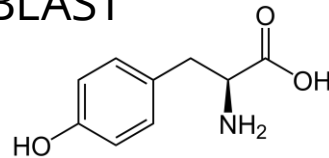
HO c1ccc(N)cc1 NH<sub>2</sub>

Tirosina (TAT ou TAC)

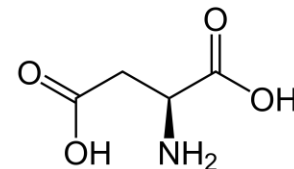
Ácido a



Fenilalanina (TTT ou TTC)



Tirosina (TAT ou TAC)

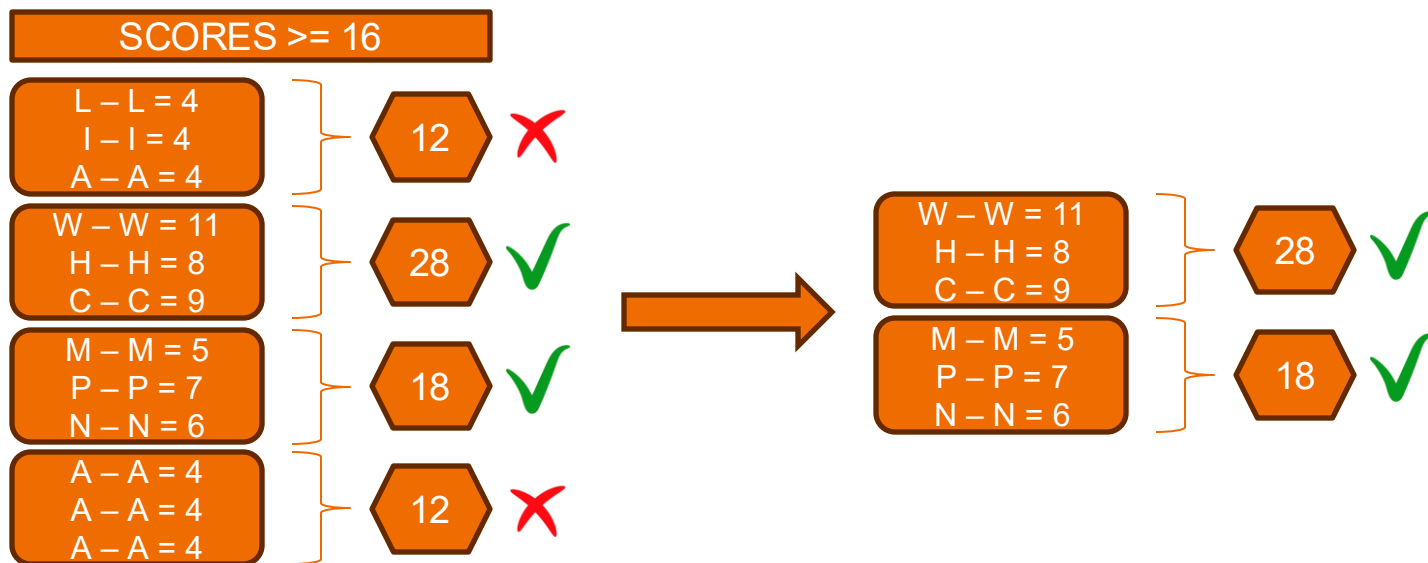


Ácido aspártico (GAT ou GAC)

# Alinhamento de Sequências

- BLAST – Algoritmo

- $W = 3$  (word size)
- Database: LIA WHC MPN AAA

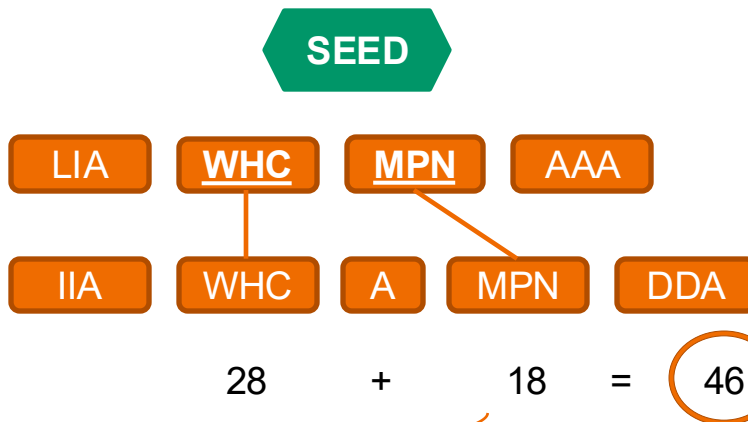




# Alinhamento de Sequências

- BLAST – Algoritmo

- W = 3 (word size)
- Database:
- Query:



Usa um limiar  
pra decidir se  
mantém esse match

## EXTEND

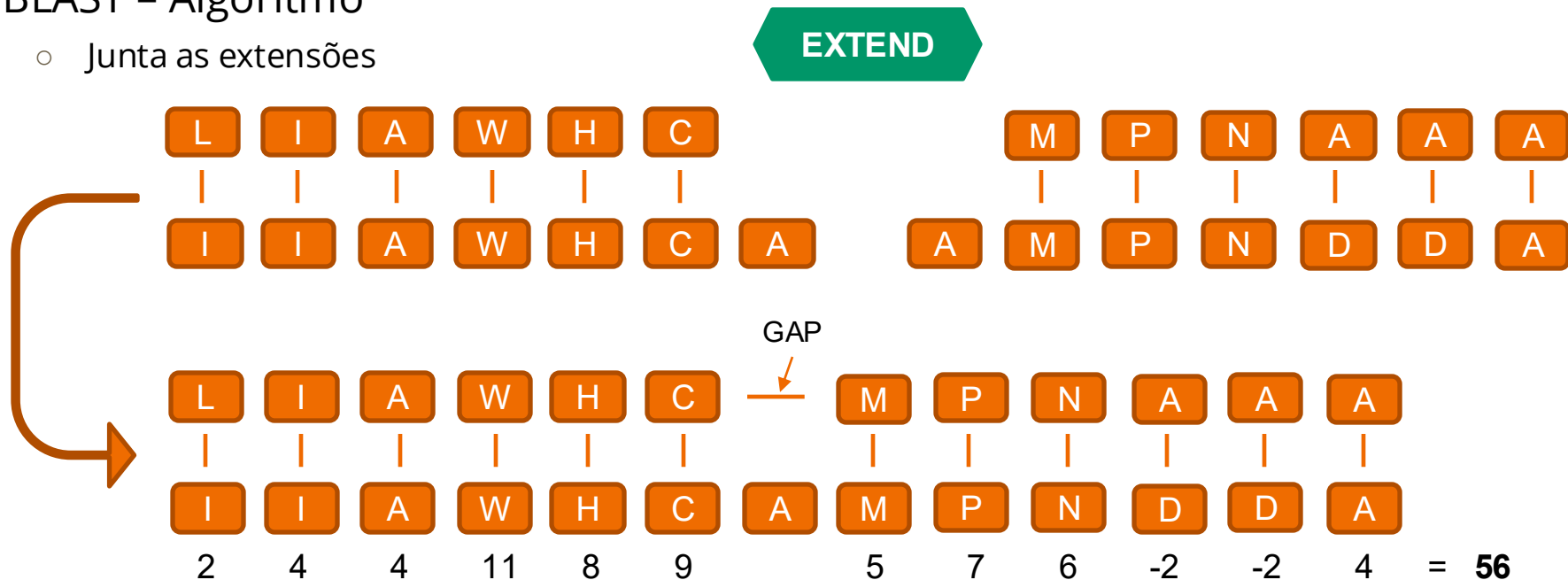
Para sequências longas, estende até o score ser positivo



# Alinhamento de Sequências

- BLAST – Algoritmo

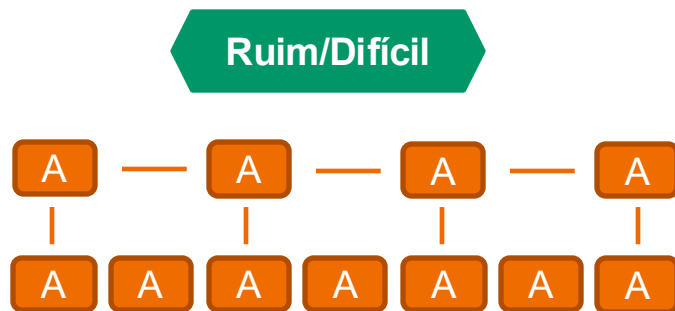
- Junta as extensões



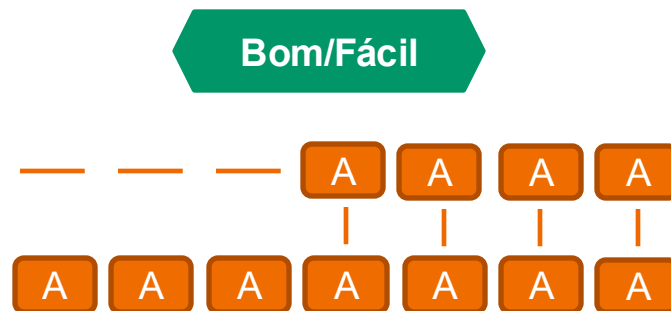
# Alinhamento de Sequências

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

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



3 eventos



1 evento

# Alinhamento de Sequências

- BLAST – Algoritmo

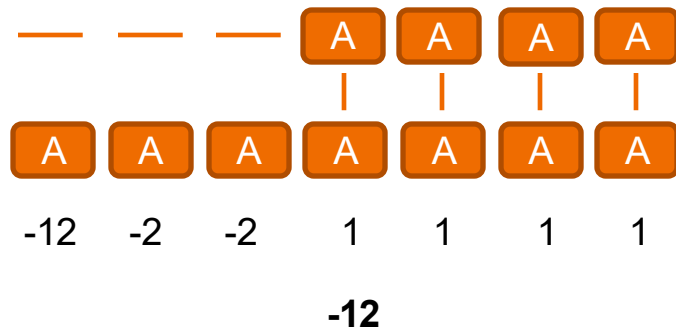
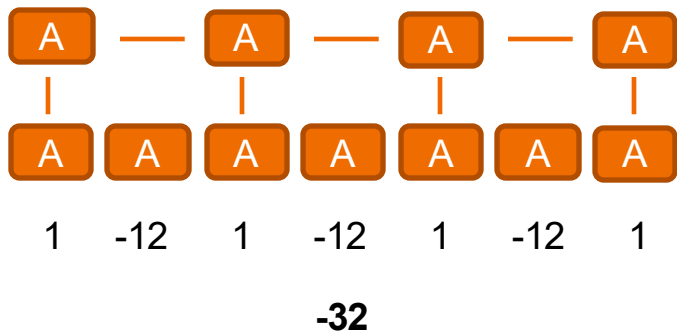
- Mas e os gaps?

- 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?

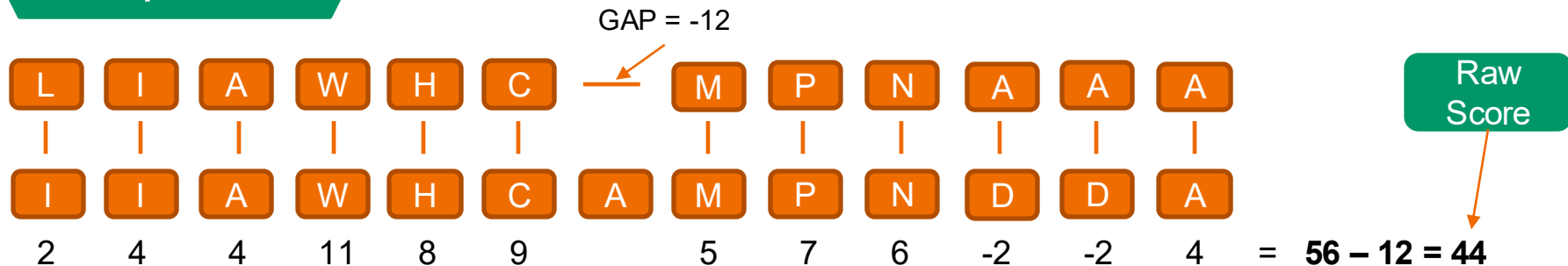
- Duas penalizações: *gap opening* e *gap extension*

Gap opening: inicia o gap = -10

Gap extension: penalização do gap = -2

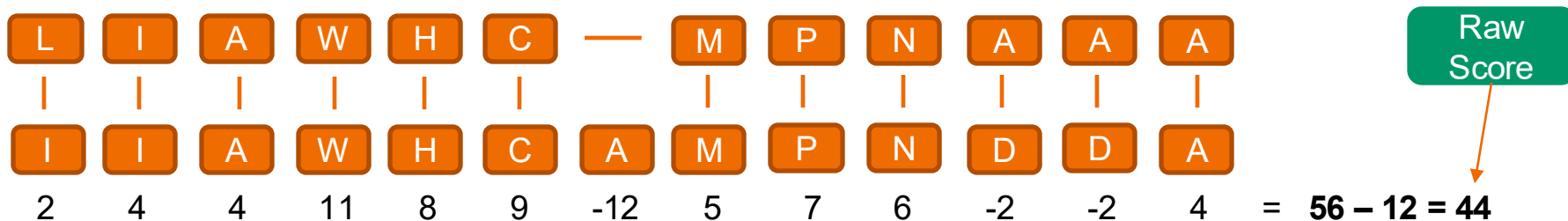
Suponha A:A = 1

## Exemplo inicial



# Alinhamento de Sequências

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



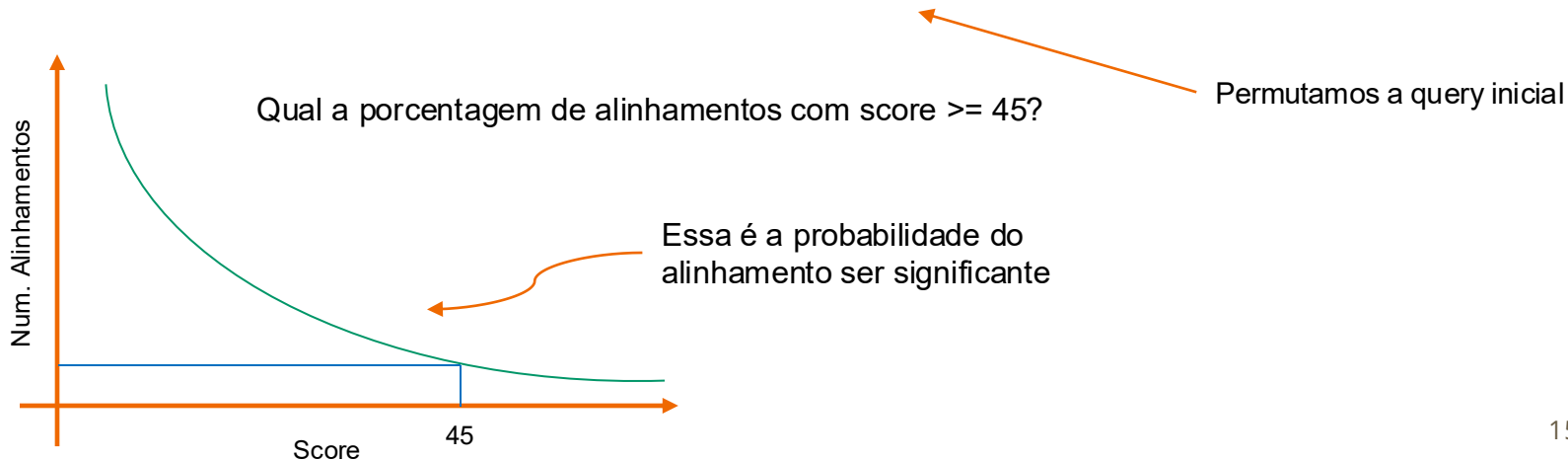
- Qual a chance de termos encontrado nosso alinhamento por pura sorte?
  - Em bancos de dados enormes, essa chance aumenta!
  - E a chance de encontrarmos os matches exatos das palavras (w) aumenta!

# Alinhamento de Sequências

- BLAST – Algoritmo

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

I I A W H C A M P N D D A



# Alinhamento de Sequências

## ● BLAST – Algoritmo

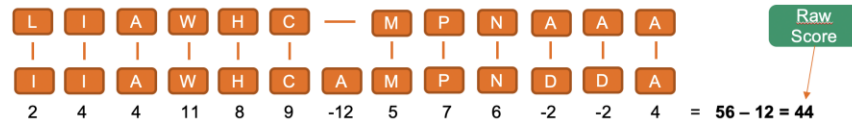
- Como saber se o alinhamento é significativo?
- O raw score (S) é um número que depende apenas do alinhamento
- Devemos converter o score para um bit score:

$$S' = \frac{\lambda S - \ln(k)}{\ln(2)} \quad \blacksquare \quad \lambda \text{ e } k \text{ são dependentes da matriz de pontuação usada e das penalizações de gap usadas}$$

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

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

Tamanho da base de dados  $\rightarrow$   $m$   
 Tamanho da query  $\rightarrow$   $n$





# Alinhamento de Sequências

## ● BLAST – Algoritmo

- Como saber se o alinhamento é significativo?
- 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  $\rightarrow$   
 Tamanho da query  $\rightarrow$   

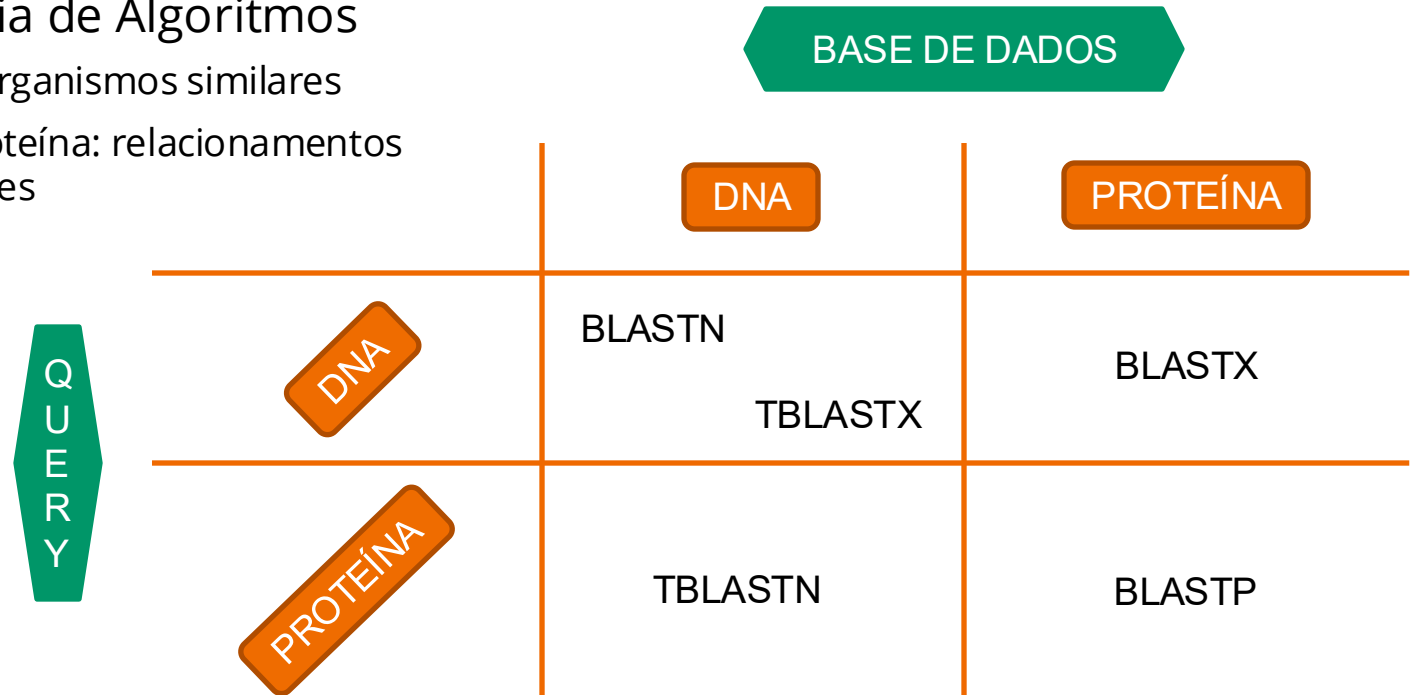
$$E = k m n e^{-\lambda S}$$

	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	$> 10^{-1}$

# Alinhamento de Sequências

- BLAST – Família de Algoritmos

- DNA-DNA: organismos similares
- Proteína-Proteína: relacionamentos mais distantes



É isso aí!

