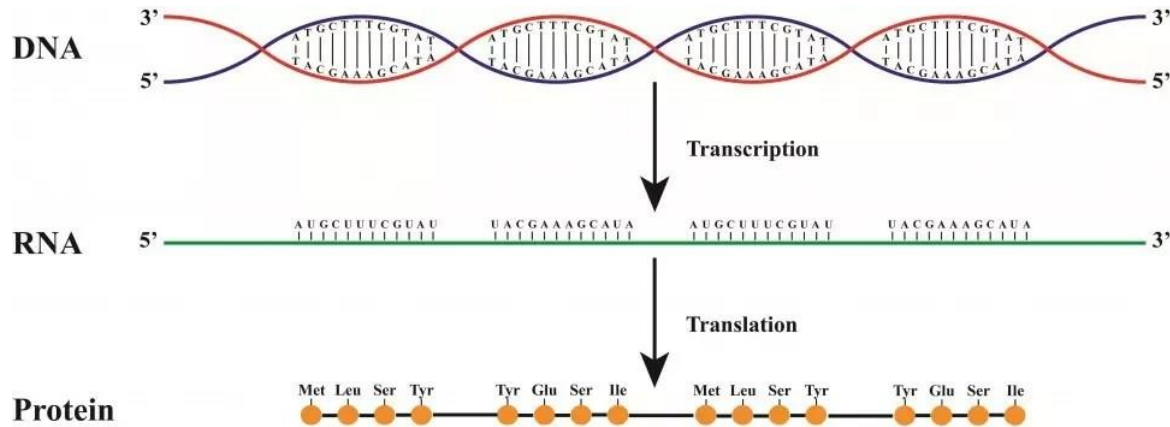

Alinhamento de Sequências

— SCC02713 - Introdução à Bioinformática —

<https://www.youtube.com/watch?v=wqsR8qOptto>
<https://www.youtube.com/watch?v=d9zprAGoCXY>
<https://www.youtube.com/watch?v=of3B02hZGS0>
<https://www.youtube.com/watch?v=ipp-pNRlp4g>
<https://www.youtube.com/watch?v=sSJYxzeFWU>
<https://www.youtube.com/watch?v=lu9ScxSejSE>

Alinhamento de Sequências

- O que é uma sequência?
 - Arranjo de duas ou mais coisas relacionadas em uma ordem sucessiva
 - Coisas relacionadas: DNA (ACGT), RNA (ACUG) e proteínas (aminoácidos)



Alinhamento de Sequências

- O que é alinhamento de sequências?
 - Maneira de arranjar as sequências de DNA, RNA ou proteínas de maneira a identificar regiões de similaridade e identidade, que podem ser consequência de relações estruturais, funcionais ou evolutivas



- Objetivo: determinar a similaridade entre diferentes sequências

Alinhamento de Sequências

- Similaridade x Identidade

- Para sequências de nucleotídeos (DNA ou RNA): têm o mesmo significado

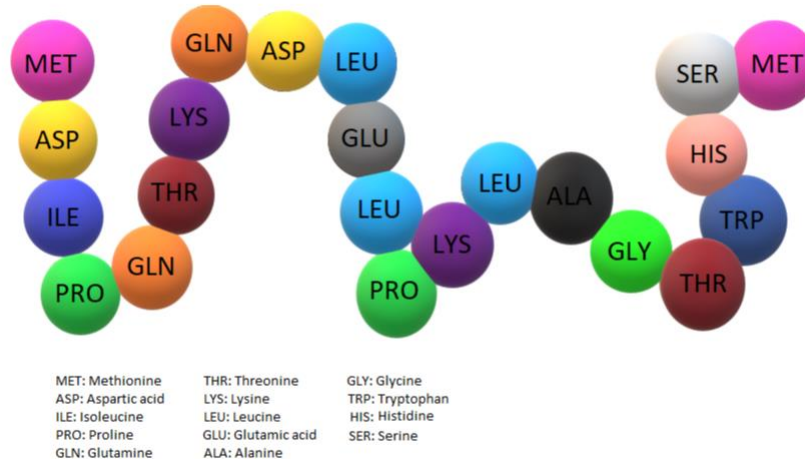


Duas sequências de DNA podem ter um alto grau de similaridade (ou identidade) - mesmo significado

Alinhamento de Sequências

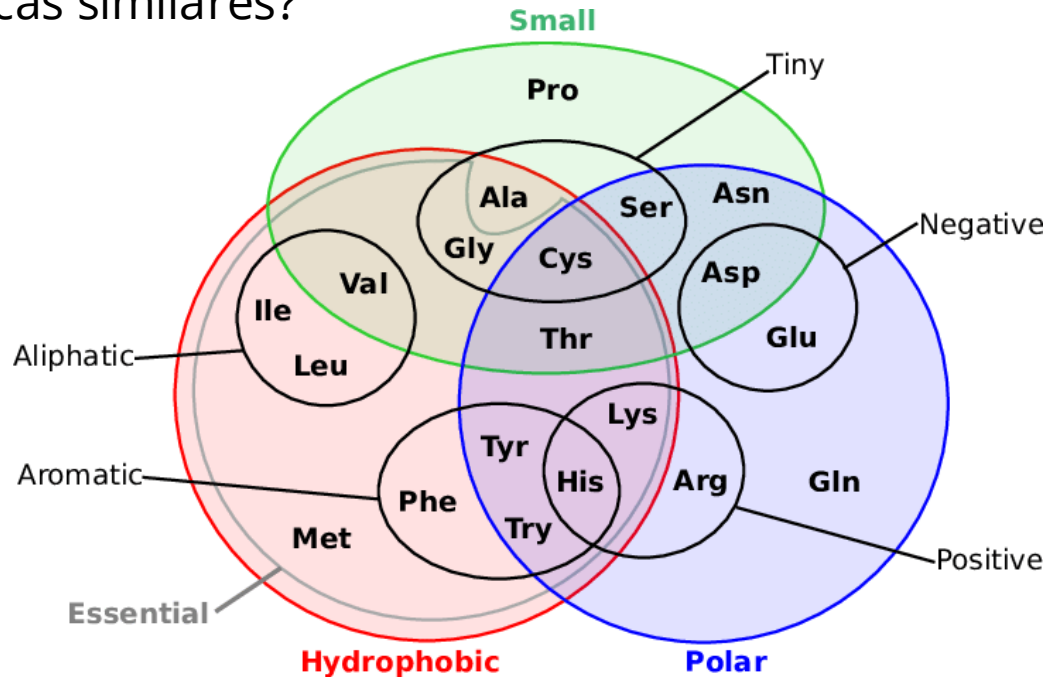
- Similaridade x Identidade

- Para sequências de proteínas: identidade e similaridade têm significados diferentes
- Identidade: % de correspondências exatas entre duas sequências alinhadas
- Similaridade: % de resíduos alinhados que compartilham características semelhantes



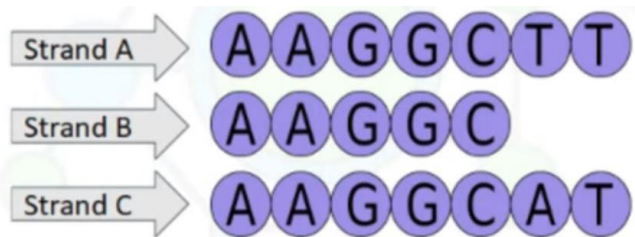
Alinhamento de Sequências

- Características similares?



Alinhamento de Sequências

- Se uma sequência $A = B$ e $B = C$:
 - A é igual a C em termos de identidade?



- $\text{Identidade}(A,B) = 100\%$ (5 nucleotídeos idênticos / $\min(\text{comprimento}(A), \text{comprimento}(B))$)
- $\text{Identidade}(B,C) = 100\%$
- $\text{Identidade}(A,C) = 85\%$ (6 nucleotídeos idênticos / 7)
- Portanto, 100% de identidade não significa que as sequências são iguais

Alinhamento de Sequências

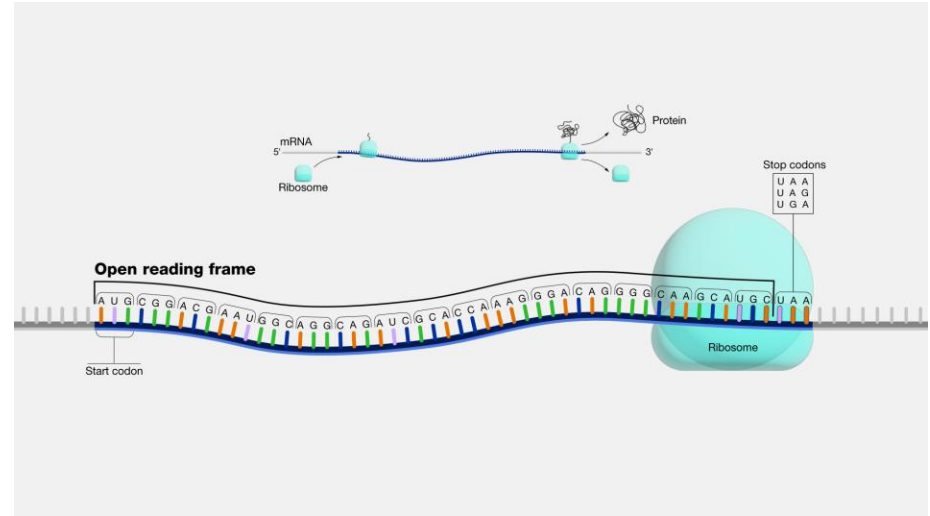
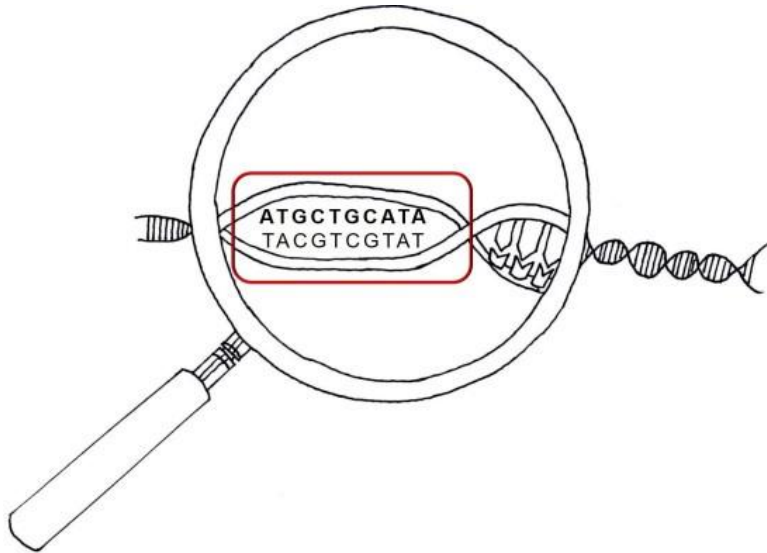
- Podemos alinhar duas ou mais sequências
 - Alinhamento em par e alinhamento múltiplo

Histone H1 (residues 120-180)

HUMAN	KKASKPKKAASKAPT	KKPKATPVKKAKKKL	AATPKKAKKPKTV	KA	PKV	KASKPKKAK	PVK
CHIMP	KKASKPKKAASKAPT	KKPKATPVKKAKKKL	AATPKKAKKPKTV	KA	PKV	KASKPKKAK	PVK
MOUSE	KKAAPKKAASKAP	SKKPKATPVKKAKKKP	AATPKKAKKPKV	V	KPKV	KASKPKKAK	TVK
RAT	KKAAPKKAASKAP	SKKPKATPVKKAKKKP	AATPKKAKKPKI	V	KPKV	KASKPKKAK	PVK
COW	KKAAPKKAASKAP	SKKPKATPVKKAKKKP	AATPKTKKPKTV	KA	PKV	KASKPKKTK	PVK
	:	**:	*****:	****	**	*****:	**
NON-CONSERVED AMINO ACIDS	Conservative	Conservative	Non-conservative	Conservative	Non-conservative	Semi-conservative	Non-conservative

Alinhamento de Sequências

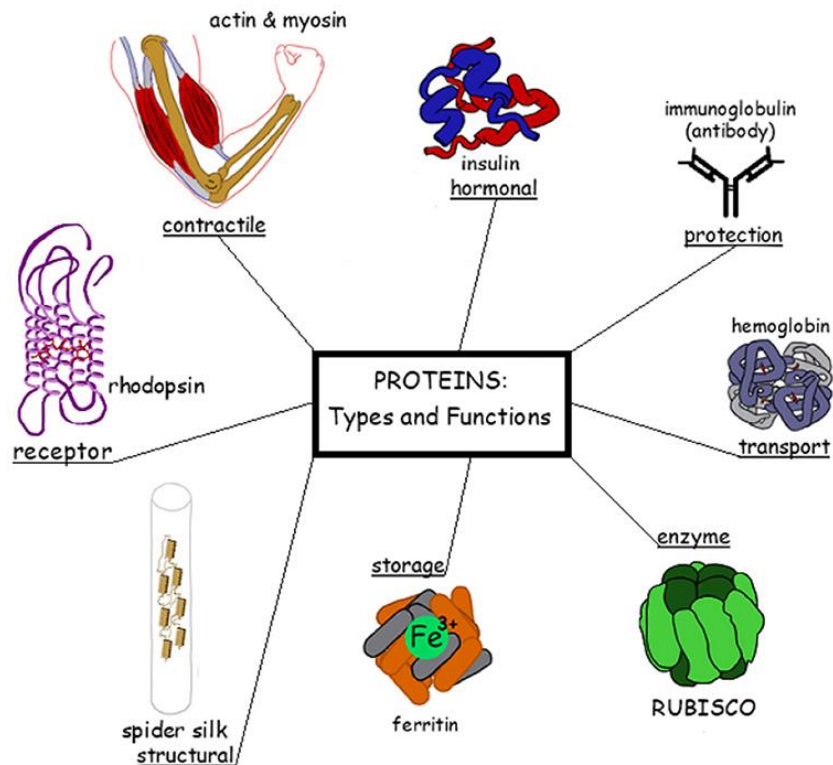
- Por que alinhar sequências?
 - Encontrar genes



Alinhamento dos ORFs

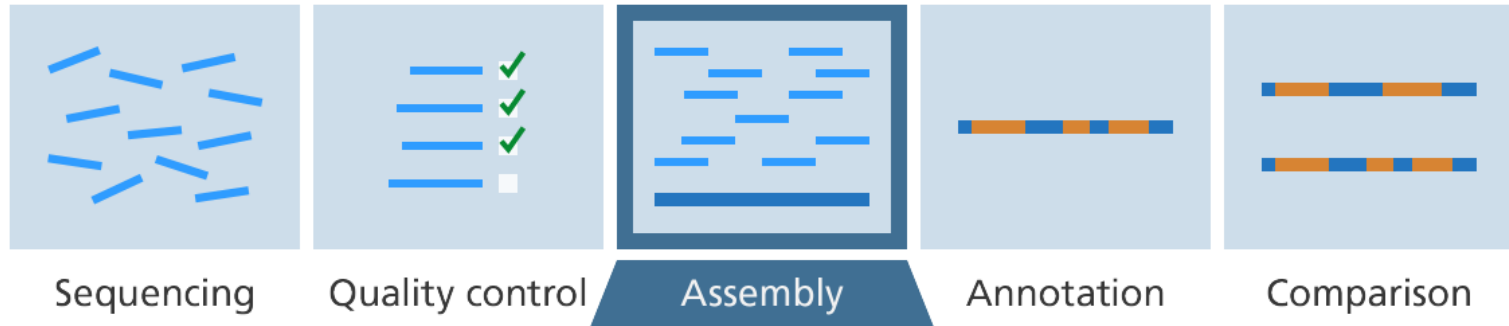
Alinhamento de Sequências

- Por que alinhar sequências?
 - Predição de funções
 - Proteínas de sequências similares podem ter funções similares



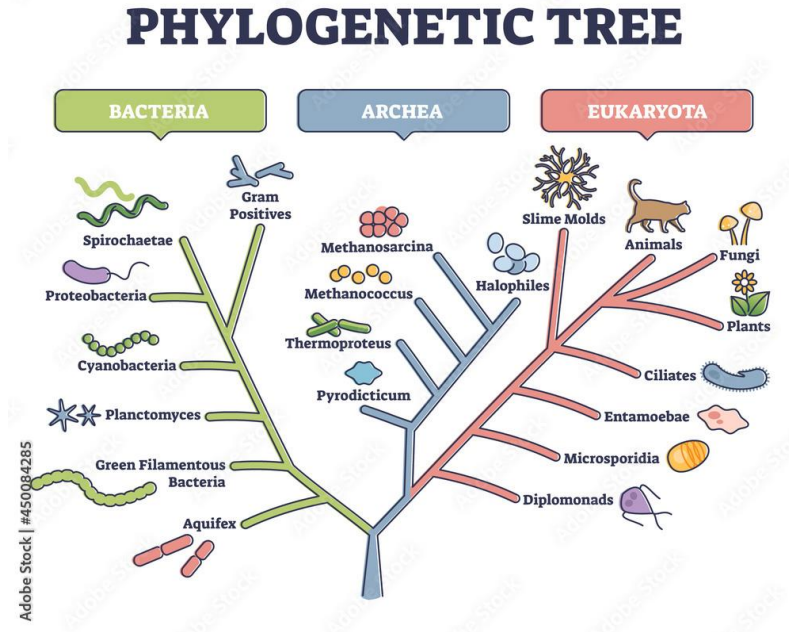
Alinhamento de Sequências

- Por que alinhar sequências?
 - Montagem de genomas
 - Um genoma é montado baseado no alinhamento entre fragmentos gerados por NGS



Alinhamento de Sequências

- Por que alinhar sequências?
 - Identificação de genes homólogos: que tem um ancestral evolutivo comum
 - Homologia: o estudo de similaridades entre organismos para determinar ancestrais comuns



Alinhamento de Sequências

- Por que alinhar sequências?
 - Gêmeos idênticos. Eles têm DNA idêntico? Pesquisem!



Alinhamento de Sequências

- O conceito de homologia

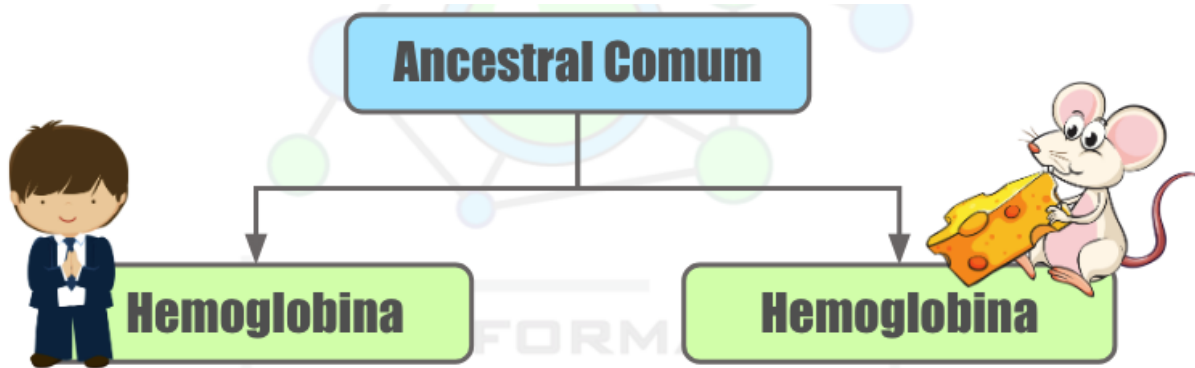


Homólogos são duas ou mais sequências que descendem de uma sequência ANCESTRAL COMUM.
Homólogos são resultados de evolução divergente.

Alinhamento de Sequências

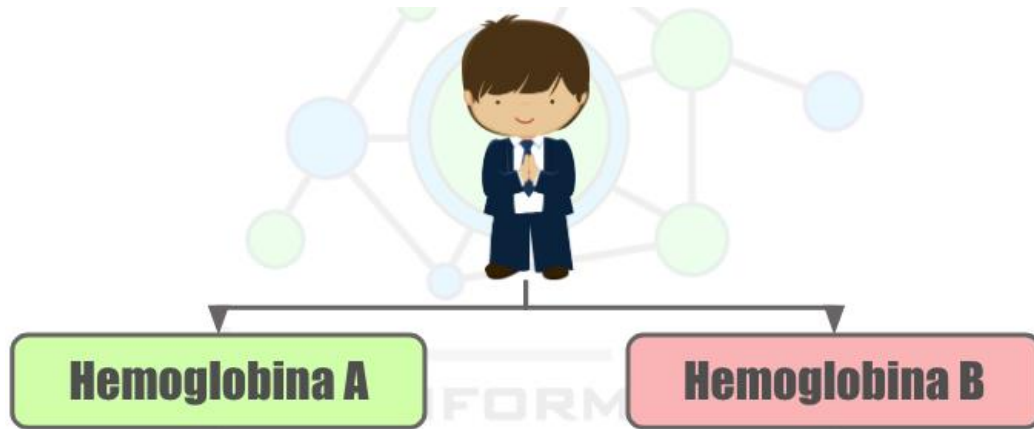
- Sequências ortólogas

- Sequências homólogas cujo último evento evolutivo foi uma especiação (geração de uma nova espécie)
- Sequências semelhantes ou idênticas em espécies DIFERENTES



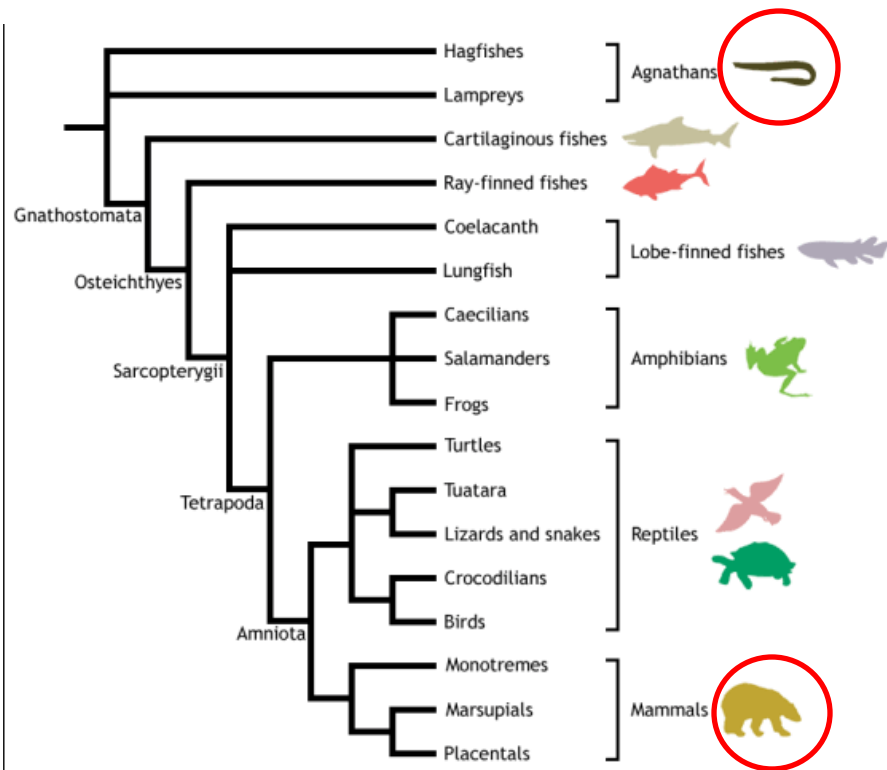
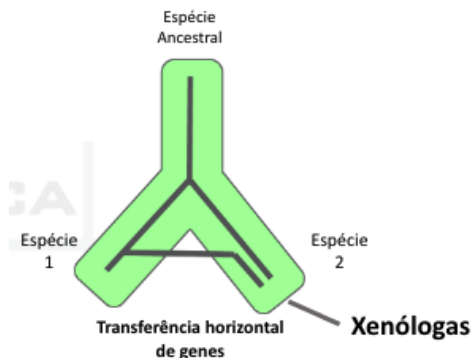
Alinhamento de Sequências

- Sequências parálogas
 - Sequências homólogas que divergiram dentro de uma espécie, ou seja, surgem durante o processo de duplicação genética.
 - Sequências semelhantes dentro das MESMAS espécies



Alinhamento de Sequências

- Sequências xenólogas
 - Sequências semelhantes entre organismos DISTANTEMENTE RELACIONADOS na história evolutiva
 - Transferência horizontal de genes



Alinhamento de Sequências

- Genes análogos

- Genes que têm FUNÇÕES IDÊNTICAS ou SEMELHANTES, mas NÃO COMPARTILHAM UM ANCESTRAL COMUM
- Os análogos têm atividade homóloga, mas origem heteróloga

EVOLUÇÃO CONVERGENTE



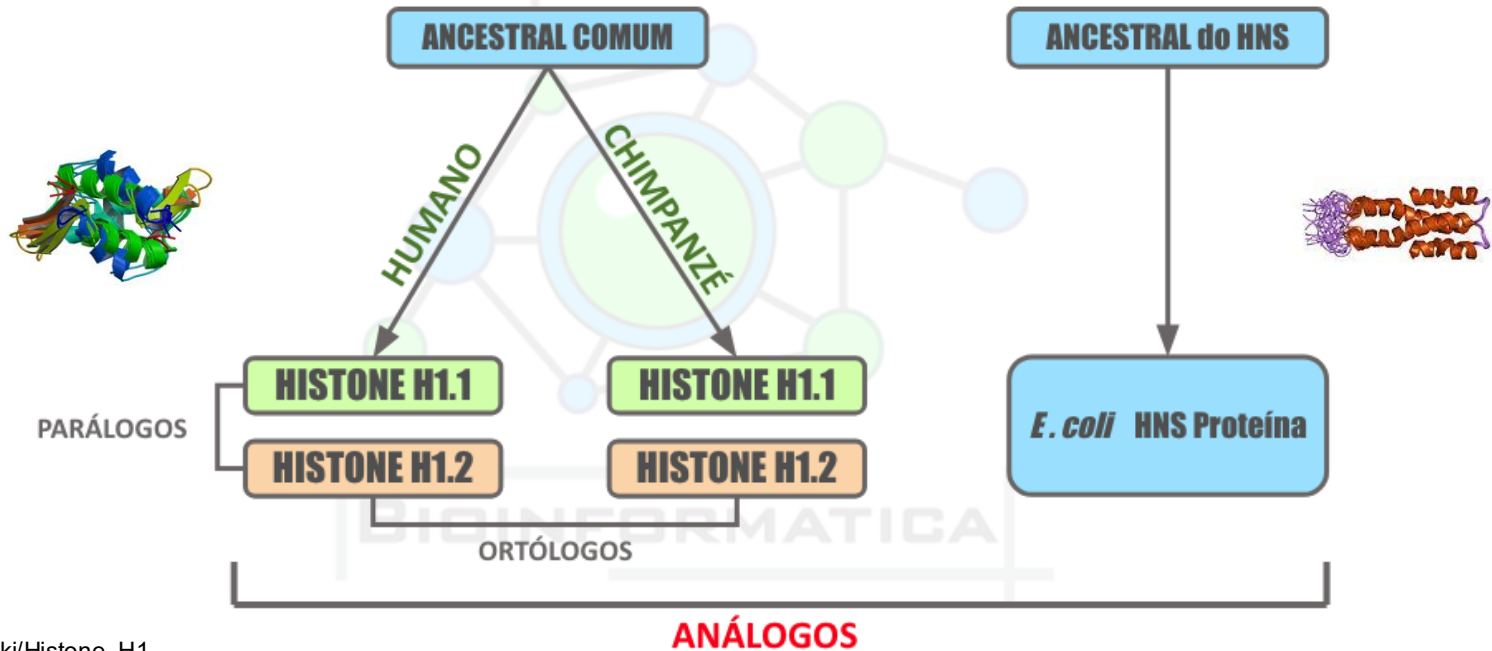
Olho da lula



Olho humano

Alinhamento de Sequências

- Homólogos x análogos



Alinhamento de Sequências

- Homólogas
 - Duas ou mais sequências que descendem de uma sequência **ancestral comum**
- Ortólogas
 - Sequências que são resultados do **processo de especiação**
- Parálogas
 - Sequências que são resultados do **processo de duplicação de genes**
- Xenólogas
 - Sequências semelhantes entre organismos **distantemente relacionados** na história evolutiva
- Análogas
 - Sequências que apresentam estrutura ou função semelhante, **mas não compartilham nenhuma sequência ancestral comum**

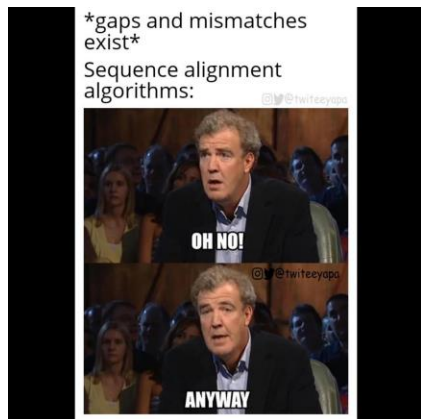
Alinhamento de Sequências



- Por que alinhar sequências?
 - Maneira eficiente e simples de determinar os relacionamentos
 - Funcionais
 - Estruturais
 - Evolutivos

Alinhamento de Sequências

- Componentes de um alinhamento
 - Matches
 - Mismatches
 - Gaps



String1: WEAREHUMANS

String2: WEARENOTHUMANZ

WEAREHUMANS
|||||
WEARENOTHUMANZ

WEARE---HUMANS
||||| |||||
WEARENOTHUMANZ

Alinhamento de Sequências

- Como identificar qual alinhamento é melhor?

- Deve haver uma pontuação para matches
- Deve haver uma penalização para mismatches
- Deve haver uma penalização para gaps

A1:	A—TGAG
Query:	ATGGCG
A2:	ATG—AG

A pontuação total é a soma de todas pontuações e penalizações

A pontuação total reflete a qualidade do alinhamento

Alinhamento de Sequências

- Dado o esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- 0 para gaps

A1: A-TGAG
Query: ATGGCG
A2: ATG-AG



A1: A-TGAG
 | | |
Query: ATGGCG
 | | | |
A2: ATG-AG
 | | | |
 $+1+0-1+1-1+1 = 1$
 $+1+1+1+0-1+1 = 3$

Alinhamento de Sequências

Alinhamento Global

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'

|||||

5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'

Query Sequence

- Tenta alinhar uma sequência inteira
- Alinha todas as letras da query e do target
- Melhor para sequências relacionadas
- Algoritmo comum: Needleman-Wunsch

Alinhamento Local

Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'

Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'

- Alinha regiões de alta similaridade
- Alinha substrings da query com substrings do target
- Melhor para sequências mais divergentes
- Algoritmo comum: Smith-Waterman

Alinhamento de Sequências

- O início (1 com 1)
 - Na década de 1970 os cientistas não estavam preocupados em ter que alinhar muitas sequências. Eles queriam encontrar o melhor alinhamento entre duas sequências
 - O algoritmo Needleman-Wunsch é o primeiro algoritmo para encontrar o alinhamento entre duas sequências e pontuar suas similaridades

	A	C	T	G
A	1	-1	-1	-1
C	-1	1	-1	-1
T	-1	-1	1	-1
G	-1	-1	-1	1

value of matching G with A

value of matching C with gap

value of matching G with G

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U	
		0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5	
A	-2	0	0	1	0	-1	-2	-3	
T	-3	-1	-1	0	2	1	0	-1	
T	-4	-2	-2	-1	1	1	0	-1	
A	-5	-3	-3	-1	0	0	0	-1	
C	-6	-4	-2	-2	-1	-1	1	0	
A	-7	-5	-3	-1	-2	-2	0	0	

Alinhamento de Sequências

- O algoritmo Needleman-Wunsch
 - Inicialize uma matriz $N \times M$
 - Preencha a matriz do canto superior esquerdo até o canto inferior direito de maneira recursiva, usando um esquema de pontuação
 - Faça um traceback

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U	
		0	-1	-2	-3	-4	-5	-6	-7
G		-1	1	0	-1	-2	-3	-4	-5
A		-2	0	0	1	-1	-2	-3	
T		-3	-1	-1	0	2	1	0	-1
T		-4	-2	-2	-1	1	1	0	-1
A		-5	-3	-3	-1	0	0	0	-1
C		-6	-4	-2	-2	-1	-1	1	0
A		-7	-5	-3	-1	-2	-2	0	0

jmb
Journal of Molecular Biology

Volume 48, Issue 3, 28 March 1970, Pages 443-453



A general method applicable to the search for similarities in the amino acid sequence of two proteins ☆

Saul B. Needleman, Christian D. Wunsch

Show more ▾

+ Add to Mendeley 🔗 Share 🗒 Cite

[https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4)

[Get rights and content ↗](#)

Alinhamento de Sequências

- O algoritmo Needleman-Wunsch

- Seq1: TGGTG

- Seq2: ATCGT

- Seq1 = m

- Seq2 = n

Passo 1: Inicializar a matriz T

		$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
	m	T	G	G	T	G	
$j=0$	n						
$j=1$	A						
$j=2$	T						
$j=3$	C						
$j=4$	G						
$j=5$	T						

$T_{(i,j)}$ is the cell at the intersection of i & j

$T_{(3,2)}$
 $T_{(4,3)}$
 Which cell is $T_{(i-1, j-1)}$
 $4-1=3$
 $3-1=2$
 Which cell is $T_{(i, j-1)}$
 Which cell is $T_{(i-1, j)}$
 $i=4$
 $j=3$ $3-1=2$ $T_{(4,2)}$

Alinhamento de Sequências

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps

● O algoritmo Needleman-Wunsch

- Passo 1: Inicializar a matriz T

		$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
		<i>m</i>	T	G	G	T	G
$J=0$	<i>n</i>	0					
$J=1$	A						
$J=2$	T						
$J=3$	C						
$J=4$	G						
$J=5$	T						

$$T_{(i,j)} = \max \begin{cases} T_{(i-1,j-1)} + \sigma(S1_{(i)}, S2_{(j)}) \\ T_{(i-1,j)} + \text{gap penalty} \\ T_{(i,j-1)} + \text{gap penalty} \end{cases}$$

Alinhamento de Sequências

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps

- O algoritmo Needleman-Wunsch

- Passo 1: Inicializar a matriz T

	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
	m	T	G	G	T	G
$J=0$	n	0				
$J=1$	A					
$J=2$	T					
$J=3$	C					
$J=4$	G					
$J=5$	T					

$$T_{(i,j)} = \max \begin{cases} T_{(i-1,j-1)} + \sigma(S1_{(i)}, S2_{(j)}) = \times \\ T_{(i-1,j)} + \text{gap penalty} = 0 + -2 = -2 \\ T_{(i,j-1)} + \text{gap penalty} = \times \end{cases}$$

$T_{(1,0)}$

Alinhamento de Sequências

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps

• O algoritmo Needleman-Wunsch

- Passo 1: Inicializar a matriz T

	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
	m	T	G	G	T	G
$J=0$	n	0	-2	-4		
$J=1$	A					
$J=2$	T					
$J=3$	C					
$J=4$	G					
$J=5$	T					

$$T_{(i,j)} = \max \begin{cases} T_{(i-1,j-1)} + \sigma(S1_{(i)}, S2_{(j)}) \\ T_{(i-1,j)} + \text{gap penalty} = -2 + (-2) = -4 \\ T_{(i,j-1)} + \text{gap penalty} \end{cases}$$

Handwritten notes: $T_{(1,0)}$, $T_{(2,0)}$

Alinhamento de Sequências

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps

• O algoritmo Needleman-Wunsch

- Passo 1: Inicializar a matriz T

		$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
		m	T	G	G	T	G
$J=0$	n	0	-2	-4	-6	-8	-10
$J=1$	A						
$J=2$	T						
$J=3$	C						
$J=4$	G						
$J=5$	T						

$$T_{(i,j)} = \max \begin{cases} T_{(i-1,j-1)} + \sigma(S1_{(i)}, S2_{(j)}) \\ T_{(i-1,j)} + \text{gap penalty} = -2 + (-2) = -4 \\ T_{(i,j-1)} + \text{gap penalty} \end{cases}$$

Handwritten notes: $T_{(1,0)}$, $T_{(2,0)}$

Alinhamento de Sequências

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps

• O algoritmo Needleman-Wunsch

- Passo 1: Inicializar a matriz T

	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	
	m	T	G	G	T	G	
$J=0$	n	0	-2	-4	-6	-8	-10
$J=1$	A	-2					
$J=2$	T						
$J=3$	C						
$J=4$	G						
$J=5$	T						

$$T_{(i,j)} = \max \begin{cases} T_{(i-1,j-1)} + \sigma(S1_{(i)}, S2_{(j)}) \rightarrow X \\ T_{(i-1,j)} + \text{gap penalty} \rightarrow X \\ T_{(i,j-1)} + \text{gap penalty} \rightarrow 0 + (-2) = -2 \end{cases}$$

$$\begin{aligned} &T_{(0,1)} \\ &T_{(-1,1)} \\ &T_{(0,0)} \end{aligned}$$

Alinhamento de Sequências

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps

• O algoritmo Needleman-Wunsch

- Passo 1: Inicializar a matriz T

		$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
		m	T	G	G	T	G
$J=0$	n	0	-2	-4	-6	-8	-10
$J=1$	A	-2	-1				
$J=2$	T						
$J=3$	C						
$J=4$	G						
$J=5$	T						

$$T_{(i,j)} = \max \begin{cases} T_{(i-1,j-1)} + \sigma(S1_{(i)}, S2_{(j)}) \rightarrow 0 + (-1) = -1 \checkmark \\ T_{(i-1,j)} + \text{gap penalty} \rightarrow -2 + (-2) = -4 \checkmark \\ T_{(i,j-1)} + \text{gap penalty} \rightarrow -2 + (-2) = -4 \checkmark \end{cases}$$

$$\begin{aligned} &T_{(1,1)} \\ &T_{(0,1)} \\ &T_{(1,0)} \end{aligned}$$

Alinhamento de Sequências

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps

• O algoritmo Needleman-Wunsch

- Passo 1: Inicializar a matriz T

	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	
	m	T	G	G	T	G	
$J=0$	n	0	-2	-4	-6	-8	-10
$J=1$	A	-2	-1	-3			
$J=2$	T						
$J=3$	C						
$J=4$	G						
$J=5$	T						

$$T_{(i,j)} = \max \begin{cases} T_{(i-1,j-1)} + \sigma(S1_{(i)}, S2_{(j)}) & -2 + (-1) = -3 \checkmark \\ T_{(i-1,j)} + \text{gap penalty} & -1 + (-2) = -3 \checkmark \\ T_{(i,j-1)} + \text{gap penalty} & -4 + (-2) = -6 \checkmark \end{cases}$$

$$T_{(2,1)}$$

$$T_{(1,1)}$$

$$T_{(2,0)}$$

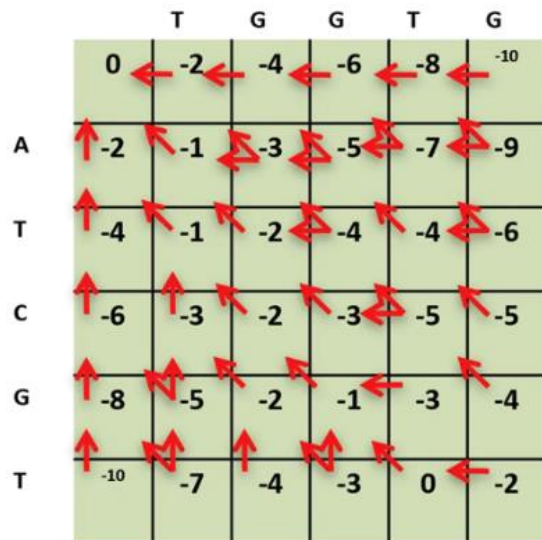


Alinhamento de Sequências

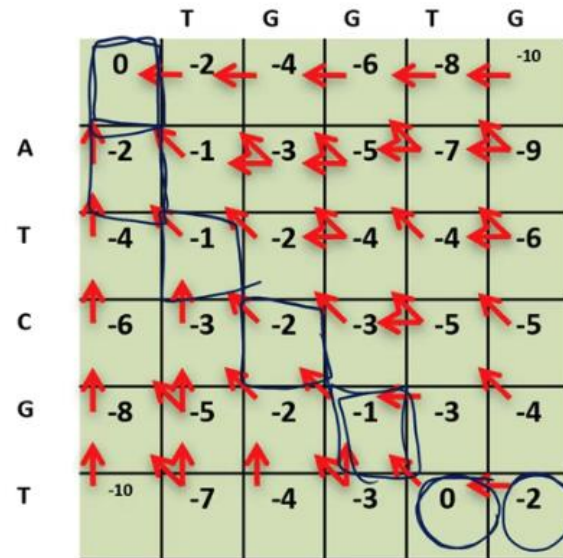
- O algoritmo Needleman-Wunsch
 - Passo 2: Traceback

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps



Volte seguindo as setas



Alinhamento de Sequências

Esquema de pontuação:

- +1 para todo match
- -1 para mismatches
- -2 para gaps

- O algoritmo Needleman-Wunsch

- Passo 2: Traceback

- O caminho através da matriz T é o traceback (em rosa aqui)
 - Para encontrar o melhor alinhamento, siga o traceback do canto superior esquerdo até o canto inferior direito e observe as letras alinhadas em cada célula



		sequence S_1					
		T	G	G	T	G	
sequence S_2	A	0	-2	-4	-6	-8	-10
	T	-2	-1	-3	-5	-7	-9
	C	-4	-1	-2	-4	-4	-6
	G	-6	-3	-2	-3	-5	-5
	G	-8	-5	-2	-1	-3	-4
	T	-10	-7	-4	-3	0	-2

Alinhamento de Sequências

- O algoritmo Needleman-Wunsch

- Passo 2: Traceback

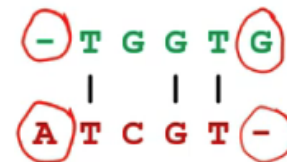
- Célula 1: não corresponde a nenhuma letra
 - Célula 2: 'A' na sequência S_2 e nada na sequência S_1
 - Célula 3: 'T' na sequência S_2 e 'T' na sequência S_1
 - Célula 4: 'C' na sequência S_2 e 'G' na sequência S_1
 - Célula 5: 'G' na sequência S_2 e 'G' na sequência S_1
 - Célula 6: 'T' na sequência S_2 e 'T' na sequência S_1
 - Célula 7: nada na sequência S_2 e 'G' na sequência S_1



Esquema de pontuação:

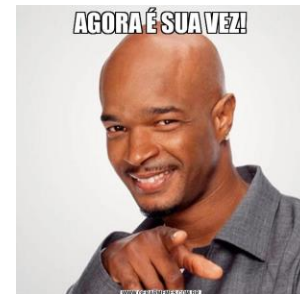
- +1 para todo match
- -1 para mismatches
- -2 para gaps

		sequence S_1					
		T	G	G	T	G	
sequence S_2	A	0	-2	-4	-6	-8	-10
	T	-2	-1	-3	-5	-7	-9
	C	-4	-1	-2	-4	-4	-6
	G	-6	-3	-2	-3	-5	-5
	G	-8	-5	-2	-1	-3	-4
	T	-10	-7	-4	-3	0	-2






Alinhamento de Sequências

Alinhe essas duas sequências usando o algoritmo Needleman-Wunsch



		A	T	G	C	T
A						
G						
C						
T						

Match : 1 
Mismatch : -1 
GAP : -2 

Sequences

Seq 1 = A T G C T

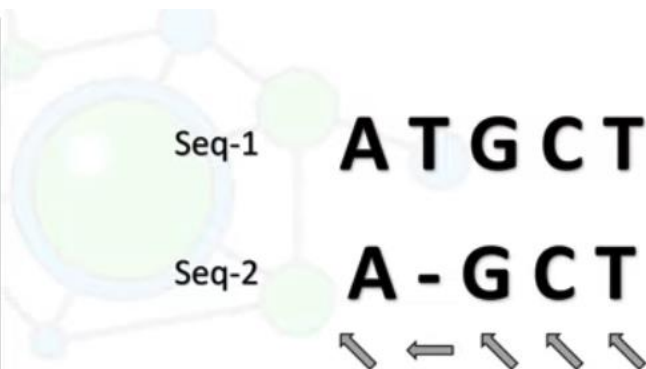
Seq 2 = A G C T

Alinhamento de Sequências

Alinhe essas duas sequências usando o algoritmo Needleman-Wunsch

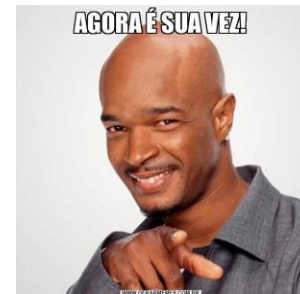
Resposta:

		A	T	G	C	T
	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
G	-4	-1	0	0	-2	-4
C	-6	-3	-2	-1	1	-1
T	-8	-5	-2	-3	-1	2



Alinhamento de Sequências

Pratique com outras sequências



Seq 1 = AATCG
Seq 2 = AACG

Seq 1 = CCGTCG
Seq 2 = CCGCG

Seq 1 = CGCTCGCT
Seq 2 = CACTCGT

Alinhamento de Sequências

- O algoritmo Smith-Waterman
 - Alinhamento entre partes de duas sequências
 - Ex: duas proteínas podem compartilhar um trecho de alta similaridade, mas ser muito diferentes fora desse região
 - Um alinhamento global dessas duas sequências teria:
 - Muitos matches na região de alta similaridade de sequência
 - Muitos mismatches e gaps (inserções/deleções) fora da região de similaridade
 - Nesse caso, faz mais sentido encontrar o melhor alinhamento local



Volume 147, Issue 1, 25 March 1981, Pages 195-197

Letter to the editor

Identification of common molecular subsequences

T.F. Smith, M.S. Waterman

[Show more](#) ▼

[+](#) Add to Mendeley [🔗](#) Share [🗣️](#) Cite

[https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5) ↗

[Get rights and content](#) ↗



Alinhamento de Sequências



- O algoritmo Smith-Waterman

- Veja por exemplo esse alinhamento global entre sequências de humano e *Drosophila*
- Espécies distantes
- Muitas regiões de alinhamento ruim, muitos gaps
- Poucas regiões de alta similaridade
- Mais sentido: alinhamento local

```
human/1-422 175 ..... DGCQQE...GGENTN 186
fly/1-898 400 SLSPNDIESLASIGHQRNCPVATEDIHLKKELGGHSDSETGSEGENSN 446

human/1-422 189 SISENGEDSDEAQMRLQLKRRKLQRNRTSFTQEDIEALEKEFERTHYPDVF 238
fly/1-898 450 GGASHIGNTEDDQARLLKRRKLQRNRTSFTNDQIDSLEKEFERTHYPDVF 499

human/1-422 239 ARERLAAKIDLPEARIQVWFSNRRAKWRREEKLNRQRQASNTPSHIPIS 288
fly/1-898 500 ARERLAGKIGLPEARIQVWFSNRRAKWRREEKLNRQRTPNSTGASATSS 549

human/1-422 289 SSFSTSVYQPIPQPTTPVSSFSSGMLORTDTALTNTYSALPPMPSFTMA 336
fly/1-898 550 TSATASLTDSRNSLSACSSLLSGSAGPSVSTINGLS-----PSTLST 594

human/1-422 339 N-NLP-----MQPPVPSQTSSYSCLMPTSPSVNGHSYD-----TYT 373
fly/1-898 595 NVNAPTLGAGIDSSSESTPIPHIRPSC---TSDNDNGRQSEDCRRVCSPC 641

human/1-422 374 PPHMQTHMNSQPMGTSSTSTGLISGVSVPVQVPGSEPDMSQYWPRQLQ- 422
fly/1-898 642 PLGVGGHQNTHHIQSNQHAQGHALVPAIS-----PRLNF 675

human/1-422 .....
fly/1-898 676 NSGSGFAMYSNMHTALSMDSYGAVTPIPSFNHSAVGPLAPPSPIPQQG 725

human/1-422 .....
fly/1-898 726 DLTSSLYPCHMTLRPPMAPAHHHIVPGDGGRPAGVGLSGSQSANLGAS 775

human/1-422 .....
fly/1-898 776 CSGSGYEVL SAYALPPPMASSSAADSSFAASSASANVTPHHTIAQESC 825

human/1-422 .....
fly/1-898 826 PSPCSSASHFQVAHSSGFSSDPI SPAVSSYAHMSYNYASSANTMT PSSAS 875

human/1-422 .....
fly/1-898 876 GTSAHVAPGKQQFFASCIFYSPWV 898
```

https://en.wikipedia.org/wiki/Drosophila_melanogaster

<https://www.britannica.com/science/human-body>

<https://pt.slideshare.net/slideshow/pairwise-sequence-alignment/16571651>

Alinhamento de Sequências



- O algoritmo Smith-Waterman

- Alinhamento local apresenta melhor resultado
- Há regiões bem conservadas entre as sequências
- Quais partes das sequências foram usadas no alinhamento local?



```
human/1-398 1 HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNQCVS 50
fly/1-573 1 HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNQCVS 50

human/1-398 51 KILGRYYETQSIRPRAIGGSKPRVATPEVVSKEADYKRECPISFAWEIRD 100
fly/1-573 51 KILGRYYETQSIRPRAIGGSKPRVATAEVVSKEADYKRECPISFAWEIRD 100

human/1-398 101 RLLSEGVCVTNDNIPSVSSINRVLRLNLASEKQDM..... 133
fly/1-573 101 RLLQENVCVTNDNIPSVSSINRVLRLNLAAKEQDQSTGSGSSSTSAGNSISA 150

human/1-398 134 ..... 135
fly/1-573 151 KVSVSIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGASNSSGEGSE 200

human/1-398 136 - DGMVDKLRMLNGDTG..... 150
fly/1-573 201 QEAIVKLRLLNTQHAAGPGPLEPARAAPLVGGSPNHLGTRSSHPQLVHG 250

human/1-398 151 ..... SWGTR... PGWYFGTSVPGQPTQ..... 170
fly/1-573 251 NHQALQHQHQQSWPPRHYSGSWYR... TSLSEIP... ISSAPNIASVTAYASGPS 299

human/1-398 171 ..... DGCCQDE... GGGEE 181
fly/1-573 300 LAHSLSPNDIESLASIGHQRNCPVATEDIHLKKELDGHQSDTGSDEGE 349

human/1-398 182 NTHSISNGEDSDEACMRLOLKRKLQRNRTSFTQEDIEALEKEFERTHYP 231
fly/1-573 350 NSNGGAENIGHTEDDARLILKRKLQRNRTSFTNDQIDSLKEKEFERTHYP 399

human/1-398 232 DVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRQASNTPSHI 281
fly/1-573 400 DVFARERLAGKIGLPEARIQVWFSNRRAKWRREEKLRNQRRTPNSTGASA 449

human/1-398 282 PISSSFSSTSVYQPIQPTTPVSSFTSGSMLRTDTALTNTYALPPMPSF 331
fly/1-573 450 TSSTSATASLTDSNLSACSSLLSGSAGOPSVSTINGLSE..... PST 494

human/1-398 332 TMAN... LPL..... MQPVPVPSQTSSYSCMLPTSPSVNGRSYD..... 366
fly/1-573 495 LSTNVMAPTLGAGIDSSSESTPIPHIRPC... TSDNDNGRQSEDCRRVC 541

human/1-398 367 TYTPPHMQTHMNSQPMGTSQTTSTGLISFGVS 398
fly/1-573 542 SPCPLGVGGHONTHIQSNHAQGHALVPAIS 573
```

https://en.wikipedia.org/wiki/Drosophila_melanogaster

<https://www.britannica.com/science/human-body>

<https://pt.slideshare.net/slideshow/pairwise-sequence-alignment/16571651>

Alinhamento de Sequências

- O algoritmo Smith-Waterman

- Encontra o melhor (maior pontuação) alinhamento local entre duas sequências
- Melhor alinhamento local: melhor alinhamento entre todas as possíveis subsequências (partes) das sequências S_1 e S_2
- Considere a tabela T , a coluna 0 e linha 0 são preenchidas com 0
- O restante da tabela usa a mesma regra de preenchimento do alinhamento global:

$$T(i, j) = \max \left\{ \begin{array}{l} T(i-1, j-1) + \sigma(S_1(i), S_2(j)) \\ T(i-1, j) + \text{gap penalty} \\ T(i, j-1) + \text{gap penalty} \\ 0 \end{array} \right.$$

A 4th possibility (unlike N-W)

- Diferença: O rastreamento **começa na célula com maior pontuação** na matriz T e vai para cima/esquerda **enquanto a pontuação ainda é positiva**

Alinhamento de Sequências

Esquema de pontuação:

- +2 para todo match
- -1 para mismatches
- -2 para gaps

● O algoritmo Smith-Waterman

- Ex: encontrar o melhor alinhamento local entre as sequências "ACCTAAGG" e "GGCTCAATCA"

		G	G	C	T	C	A	A	T	C	A
		0	0	0	0	0	0	0	0	0	0
A		0									
C		0									
C		0									
T		0									
A		0									
A		0									
G		0									
G		0									

$$T(i, j) = \max \begin{cases} T(i-1, j-1) + \sigma(S_1(i), S_2(j)) \\ T(i-1, j) + \text{gap penalty} \\ T(i, j-1) + \text{gap penalty} \\ 0 \end{cases}$$

A 4th possibility (unlike N-W)

Alinhamento de Sequências

Esquema de pontuação:

- +2 para todo match
- -1 para mismatches
- -2 para gaps

• O algoritmo Smith-Waterman

- Ex: encontrar o melhor alinhamento local entre as sequências "ACCTAAGG" e "GGCTCAATCA"
- $T(1,1)$:

		G	G	C	T	C	A	A	T	C	A
A	0	0	0	0	0	0	0	0	0	0	0
C	0	0	?								
C	0										
T	0										
A	0										
A	0										
G	0										
G	0										

$$T(i, j) = \max \begin{cases} T(i-1, j-1) + \sigma(S_1(i), S_2(j)) = 0 - 1 = -1 \\ T(i-1, j) + \text{gap penalty} = 0 - 2 = -2 \\ T(i, j-1) + \text{gap penalty} = 0 - 2 = -2 \end{cases}$$

0

Depois calculados $T(2,1)$
Assim por diante...

Alinhamento de Sequências

Esquema de pontuação:

- +2 para todo match
- -1 para mismatches
- -2 para gaps

- O algoritmo Smith-Waterman

- Preenchemos toda a matriz T, sempre guardando a célula anterior (se existir) usada para calcular o valor de cada célula $T(i,j)$:
- Não há valores negativos
- Começamos o traceback do maior valor
- Percorre enquanto a pontuação é positiva



	G	G	C	T	C	A	A	T	C	A	
	0	0	0	0	0	0	0	0	0	0	
A	0	0	0	0	0	2	2	0	0	2	
C	0	0	0	2	0	2	0	1	1	2	0
C	0	0	0	2	1	2	1	0	0	3	1
T	0	0	0	0	4	2	1	0	2	1	2
A	0	0	0	0	2	3	4	3	1	1	3
A	0	0	0	0	0	1	5	6	4	2	3
G	0	2	2	0	0	0	3	4	5	3	1
G	0	2	4	2	0	0	1	2	3	4	2

Alinhamento de Sequências

Esquema de pontuação:

- +2 para todo match
- -1 para mismatches
- -2 para gaps

● O algoritmo Smith-Waterman

- Preenchemos toda a matriz T, sempre guardando a célula anterior (se existir) usada para calcular o valor de cada célula $T(i,j)$:
- Não há valores negativos
- Começamos o traceback do maior valor
- Percorre enquanto a pontuação é positiva



	G	G	C	T	C	A	A	T	C	A	
	0	0	0	0	0	0	0	0	0	0	
A	0	0	0	0	0	0	2	2	0	0	2
C	0	0	0	2	0	2	0	1	1	2	0
C	0	0	0	2	1	2	1	0	0	3	1
T	0	0	0	0	4	2	1	0	2	1	2
A	0	0	0	0	2	3	4	3	1	1	3
A	0	0	0	0	0	1	5	6	4	2	3
G	0	2	2	0	0	0	3	4	5	3	1
G	0	2	4	2	0	0	1	2	3	4	2

Alinhamento de Sequências

Esquema de pontuação:

- +2 para todo match
- -1 para mismatches
- -2 para gaps

- O algoritmo Smith-Waterman

- Você encontra o melhor alinhamento a partir do traceback, assim como Needleman-Wunsch

	G	G	C	T	C	A	A	T	C	A
	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	2	2	0	2
C	0	0	0	2	0	2	0	1	1	2
C	0	0	0	2	1	2	1	0	0	3
T	0	0	0	0	4	2	1	0	2	1
A	0	0	0	0	2	3	4	3	1	1
A	0	0	0	0	0	1	5	6	4	2
G	0	2	2	0	0	0	3	4	5	3
G	0	2	4	2	0	0	1	2	3	4



C T C A A
| | | |
C T - A A

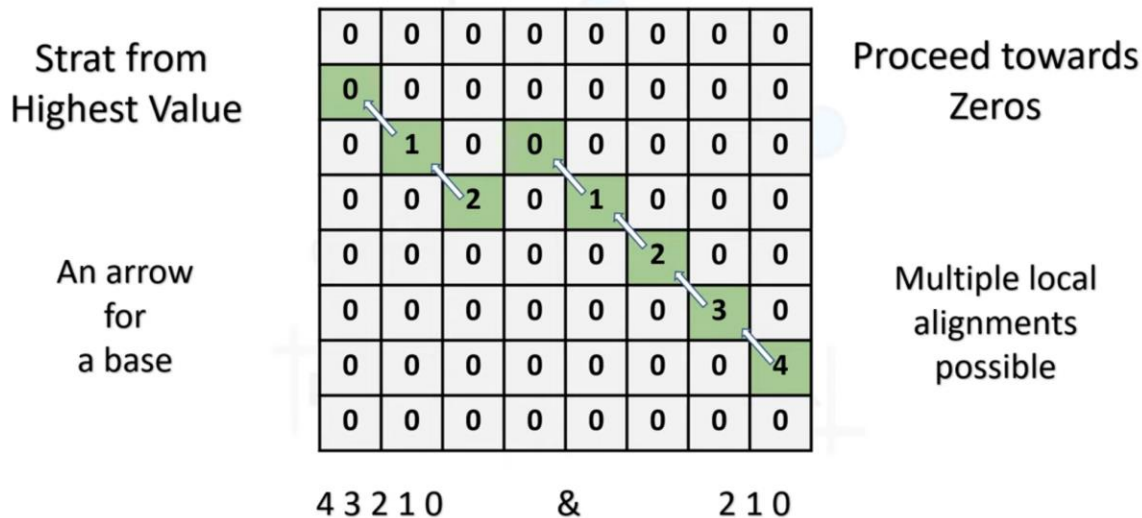


Alinhamento de Sequências

Esquema de pontuação:

- +2 para todo match
- -1 para mismatches
- -2 para gaps

- O algoritmo Smith-Waterman
 - Importante: podemos ter múltiplos alinhamentos locais



Alinhamento de Sequências

Encontre o melhor alinhamento local entre as sequências “AGCT” e “ATGCT”, com pontuações +1 (match), -1 (mismatch) e -2 (gap)

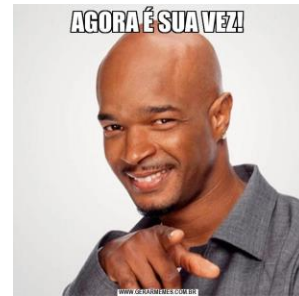
		A	T	G	C	T
A						
G						
C						
T						

Seq 1

Seq 1

ATGCT

AGCT



Alinhamento de Sequências

Encontre o melhor alinhamento local entre as sequências “AGCT” e “ATGCT”, com pontuações +1 (match), -1 (mismatch) e -2 (gap)

		A	T	G	C	T
	0	0	0	0	0	0
A	0	1	0	0	0	0
G	0	0	0	1	0	0
C	0	0	0	0	2	0
T	0	0	0	0	0	3

GCT

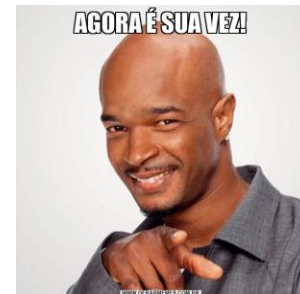


ATGCT
| | |
AGCT



Alinhamento de Sequências

Encontre o melhor alinhamento local entre as sequências “TCAGTTGCC” e “AGGTTG”, com pontuações +1 (match), -2 (mismatch) e -2 (gap)



Alinhamento de Sequências

Encontre o melhor alinhamento local entre as sequências “TCAGTTGCC” e “AGGTTG”, com pontuações +1 (match), -2 (mismatch) e -2 (gap)



	T	C	A	G	T	T	G	C	C
	0	0	0	0	0	0	0	0	0
A	0	0	0	1	0	0	0	0	0
G	0	0	0	0	2	0	0	1	0
G	0	0	0	0	1	0	0	1	0
T	0	1	0	0	0	2	1	0	0
T	0	1	0	0	0	1	3	1	0
G	0	0	0	0	1	0	1	4	2



G	T	T	G
G	T	T	G

É isso aí!

