

Alinhamento de sequências

Alinhamento múltiplo

Prof. Dr. Tetsu Sakamoto
Instituto Metr pole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

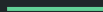
Objetivos

Alinhamento múltiplo de
sequências

O que é...

Para quê...

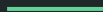
Como fazer...



O que é...

Alinhamento múltiplo de
sequências

“Alinhamento de três ou mais
sequências biológicas de forma a
maximizar a similaridade entre as
sequências.”



O que é MSA

[illegible]

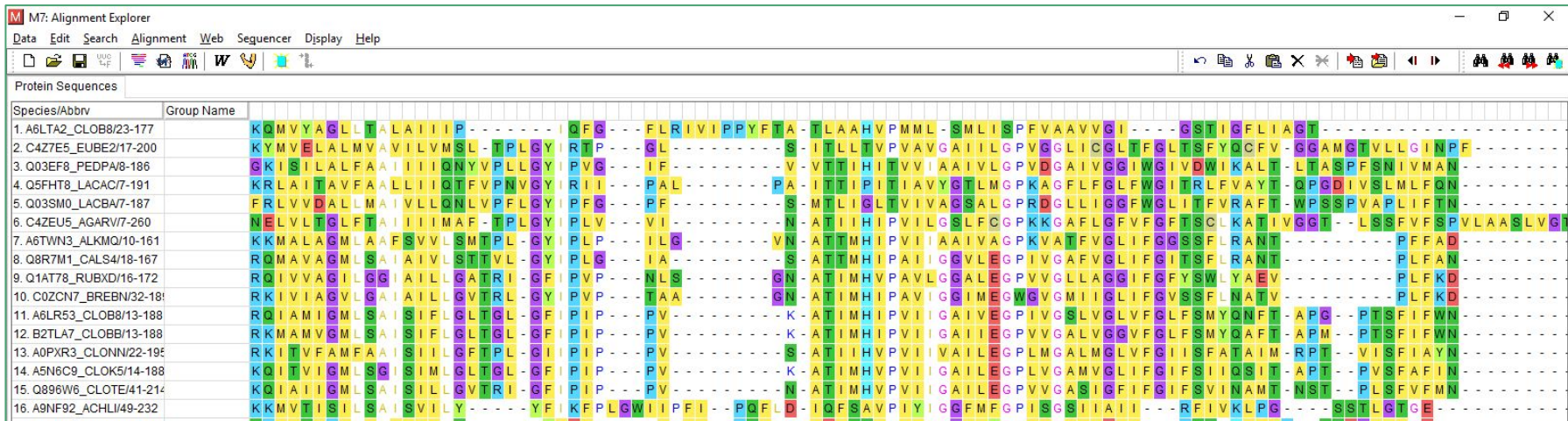
Para quê...

Alinhamento múltiplo de
sequências

“Vários padrões importantes
(motivos, domínios) não são
explícitos em um alinhamento
par-a-par.”



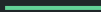
Para quê MSA



Como fazer...

Alinhamento múltiplo de
sequências

- Método de pontuação;
- Programação dinâmica;
- Métodos heurísticos;
 - Alinhamento progressivo;
 - Método iterativo.



Como avaliar um alinhamento múltiplo?

Existem várias formas de alinhar múltiplas sequências:

A A T G A T C

A A G G - T C

A - G - A T C

A A T G A T C

A A - G G T C

A - - G A T C

Como escolher qual é o melhor?

Como avaliar um alinhamento múltiplo?

No caso do alinhamento entre pares de sequências:

A A T G A T C

| | | | |

A A G G - T C

5 matches
1 mismatch
1 gap

A A T G A T C

| | | | |

A A - G G T C

5 matches
1 mismatch
1 gap

Como avaliar um alinhamento múltiplo?

Existem várias formas de alinhar múltiplas sequências:

A A T G A T C

A A G G - T C

A - G - A T C

A A T G A T C

A A - G G T C

A - - G A T C

Como escolher qual é o melhor?

Como avaliar um alinhamento múltiplo?

Não existe um critério matemático consenso que determine uma pontuação para o alinhamento.

Soma dos pares

Considera que o alinhamento múltiplo como $\binom{n}{2}$ pares de alinhamentos induzidos.

A ideia é utilizar o sistema de pontuação do alinhamento de um par de sequências.

Considere: match: 1; mismatch -1; gap: -1

A	A	T	G	A	T	C	}	3	}	1	=	5
A	A	G	G	-	T	C						
A	-	G	-	A	T	C						
3	-1	-1	-1	-1	3	3						

A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	I
A	A	I	I
A	I	I	I
<hr/>			
15	10	7	6

Soma dos pares

Exercício:

Qual a pontuação deste alinhamento?

Considere:

- Match: 1
- Mismatch: -1
- Gap: -2
- Gap-gap: 0

M	Q	P	I	L	L	L	V
M	L	R	-	L	L	-	-
M	K	-	I	L	L	L	-

Pontuação baseado em entropia

$$-\sum_j (c_j/C) \log (c_j/C)$$

c_j = número de ocorrência do aminoácido j na coluna;

C = número de símbolos na coluna;

A	A	A	A	A
A	A	A	A	I
A	A	A	A	K
A	A	A	I	L
A	A	I	I	S
A	I	I	I	W

0	.44	.65	.69	1.79
---	-----	-----	-----	------

Logarítmo
neperiano

Como achar um
alinhamento múltiplo
com pontuação
máxima (ou mínima)?

Programação dinâmica

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(x_i, y_i) \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{cases}$$

seqi T H I S - L I - N E -
 | | | | | | |
 seqj - - I S A L I G N E D

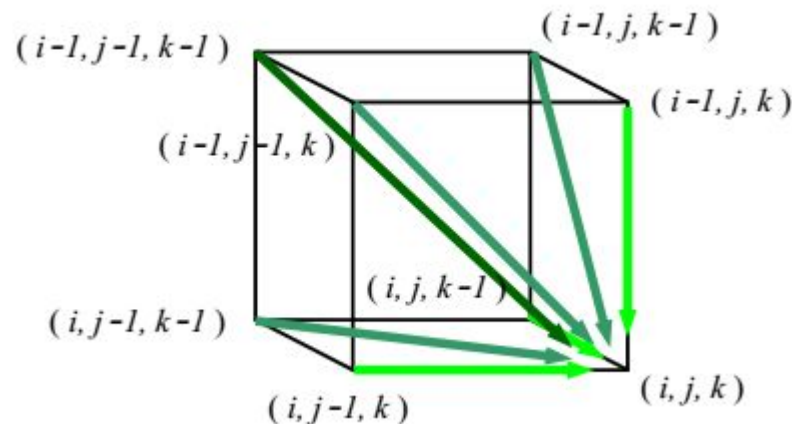
		I	S	A	L	I	G	N	E	D
T	0	→ -4	→ -8	→ -12	→ -16	→ -20	→ -24	→ -28	→ -32	→ -36
H	↓ -4	↓ -1	↓ -3	→ -7	→ -11	→ -15	→ -19	→ -23	→ -27	→ -31
I	↓ -8	↓ -5	↓ -2	→ -5	→ -9	→ -13	→ -17	→ -18	→ -22	→ -26
S	↓ -12	↓ -4	↓ -6	↓ -3	↓ -3	→ -5	→ -9	→ -13	→ -17	→ -21
L	↓ -16	↓ -8	↓ 0	→ -4	↓ -5	↓ -5	↓ -5	→ -8	→ -12	→ -16
I	↓ -20	↓ -12	↓ -4	↓ -1	↓ 0	↓ -3	↓ -7	↓ -8	↓ -11	↓ -15
N	↓ -24	↓ -16	↓ -8	↓ -5	↓ 1	↓ 4	↓ 0	↓ -4	↓ -8	↓ -12
E	↓ -28	↓ -20	↓ -12	↓ -9	↓ -3	↓ 0	↓ 4	↓ 6	↓ 2	↓ -2
D	↓ -32	↓ -24	↓ -16	↓ -13	↓ -7	↓ -4	↓ 0	↓ 4	↓ 11	↓ 7

Programação dinâmica com 3 sequências

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(x_i, y_i) \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{cases}$$



$$S_{i,j,k} = \max \begin{cases} S_{i-1,j,k} + \delta(x_i, -, -) \\ S_{i,j-1,k} + \delta(-, y_i, -) \\ S_{i,j,k-1} + \delta(-, -, z_k) \\ S_{i-1,j-1,k} + \delta(x_i, y_j, -) \\ S_{i-1,j,k-1} + \delta(x_i, -, z_k) \\ S_{i,j-1,k-1} + \delta(-, y_j, z_k) \\ S_{i-1,j-1,k-1} + \delta(x_i, y_j, z_k) \end{cases}$$



Programação dinâmica

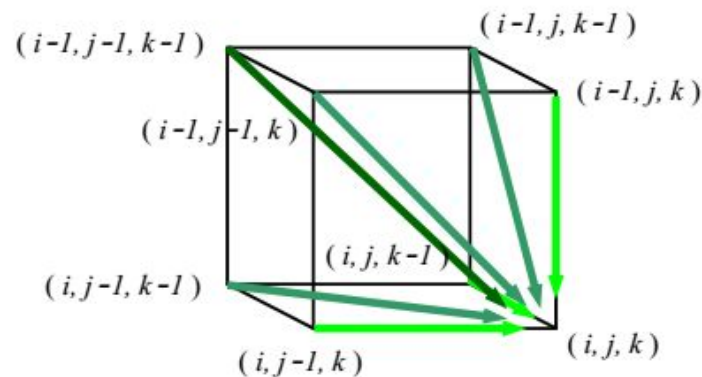
Custo para realizar a programação dinâmica:

Cada célula possui $2^k - 1$ ($O(2^k)$) células vizinhas
($k = \#$ de sequências);

Cálculo da soma dos pares de cada célula vizinha necessitaria de aproximadamente k^2 operações.

O tamanho da matriz é de n^k ($n = \#$ de resíduos);

Complexidade computacional = $O(n^k 2^k k^2)$;



Alinhamento progressivo

Complexidade do alinhamento por programação dinâmica cresce exponencialmente com o número de sequências e de resíduos → **heurística!**

Alinhamento progressivo:

Etapas:

1. Construção de uma matriz de distância D entre as sequências;
2. Construção de uma árvore guia T ;
3. Construção do alinhamento múltiplo através de alinhamento par-a-par ou alinhamento parcial (perfil) guiado por T ;
4. Melhorar o alinhamento.

Alinhamento progressivo

(1) ACTCAT
(2) AGTCAT
(3) ACGTCCT

$$d(x, y) = 3$$

$$d(x, -) = 2$$

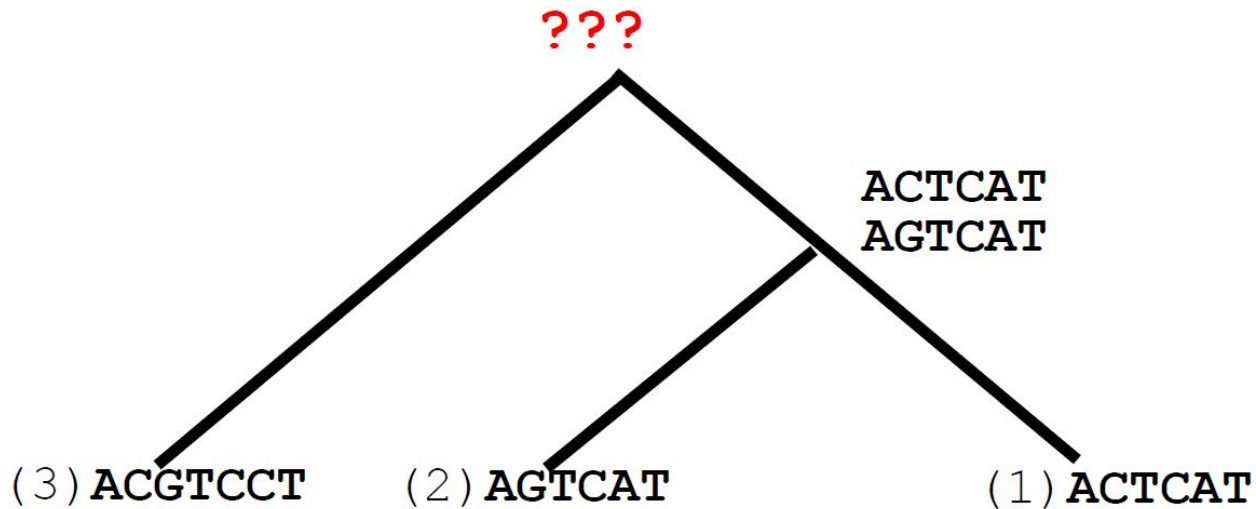
(1) A**C**TCAT
(2) A**G**TCAT **3**

(2) A-GTC**A**T
(3) A**C**GTC**C**T **5**

(1) AC-T**C**A**T**
(3) AC**G**TC**C**T **5**

Alinhamento progressivo

Construção de uma árvore guia (T).



Alinhamento progressivo

Alinhar (1) e (2) com (3).

(1) ACTCAT
(2) AGTCAT
(3) ACGTCCT

$$d(x, y) = 3$$

$$d(x, -) = 2$$

(1) ACTCAT	3
(2) AGTCAT	

(2) A-GTCAT
(3) ACGTCCT

5

(1) AC-TCAT
(3) ACGTCCT

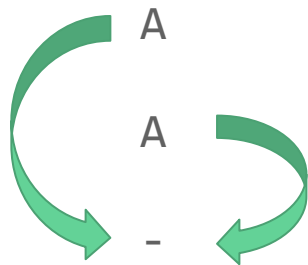
5

Alinhamento progressivo

Alinhamento da sequência com o perfil.

$$d(x, y) = 3$$

$$d(x, -) = 2$$



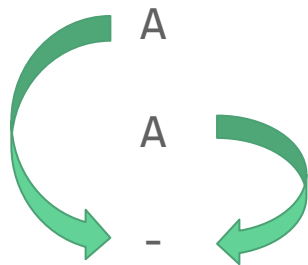
	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0						
A							
C							
G							
T							
C							
C							
T							

Alinhamento progressivo

Alinhamento da sequência com o perfil.

$$d(x, y) = 3$$

$$d(x, -) = 2$$



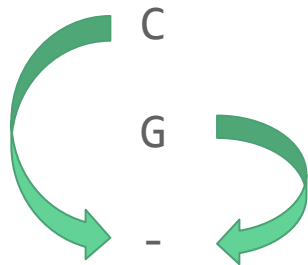
	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4					
A							
C							
G							
T							
C							
C							
T							

Alinhamento progressivo

Alinhamento da sequência com o perfil.

$$d(x, y) = 3$$

$$d(x, -) = 2$$



	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8				
A							
C							
G							
T							
C							
C							
T							

Alinhamento progressivo

Alinhamento da sequência com o perfil.

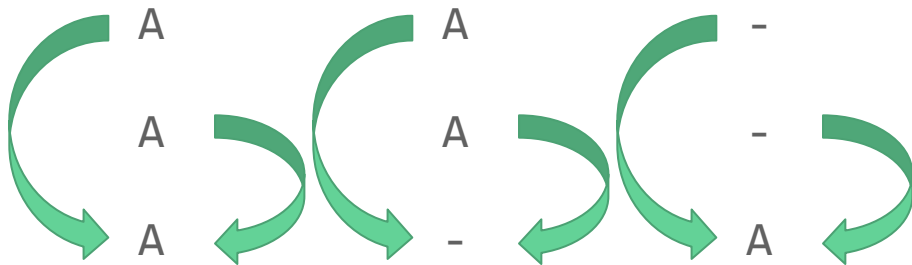
$$d(x, y) = 3$$

$$d(x, -) = 2$$

	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8	12	16	20	24
A	4						
C	8						
G	12						
T	16						
C	20						
C	24						
T	28						

Alinhamento progressivo

Alinhamento da sequência com o perfil.



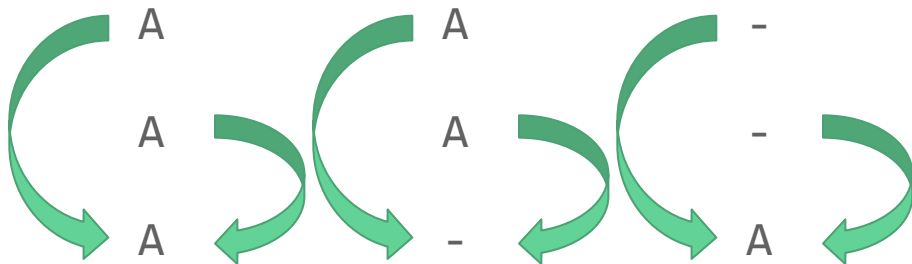
$$d(x, y) = 3$$

$$d(x, -) = 2$$

	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8	12	16	20	24
A	4						
C	8						
G	12						
T	16						
C	20						
C	24						
T	28						

Alinhamento progressivo

Alinhamento da sequência com o perfil.



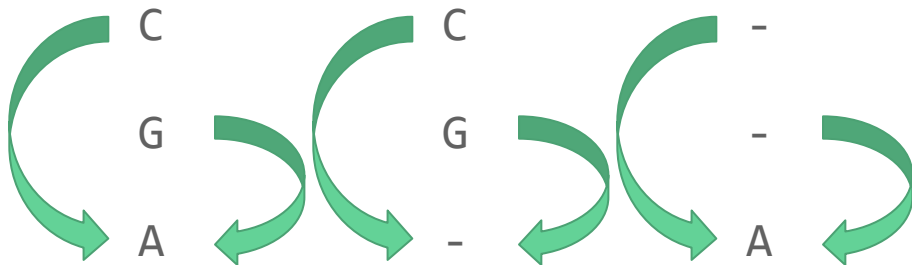
$$d(x, y) = 3$$

$$d(x, -) = 2$$

	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8	12	16	20	24
A	4	0					
C	8						
G	12						
T	16						
C	20						
C	24						
T	28						

Alinhamento progressivo

Alinhamento da sequência com o perfil.



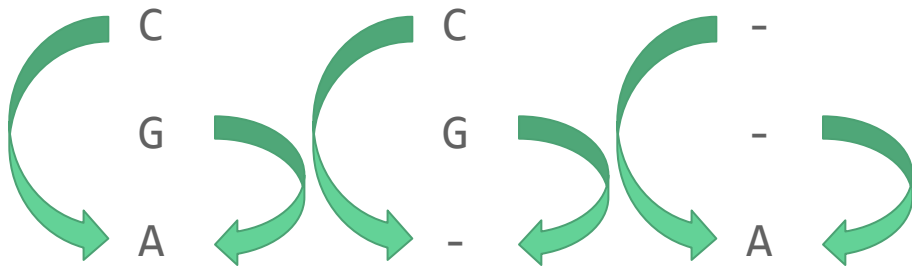
$$d(x, y) = 3$$

$$d(x, -) = 2$$

	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8	12	16	20	24
A	4	0					
C	8						
G	12						
T	16						
C	20						
C	24						
T	28						

Alinhamento progressivo

Alinhamento da sequência com o perfil.



$$d(x, y) = 3$$

$$d(x, -) = 2$$

	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8	12	16	20	24
A	4	0	4				
C	8						
G	12						
T	16						
C	20						
C	24						
T	28						

Alinhamento progressivo

Alinhamento da sequência com o perfil.

$$d(x, y) = 3$$

$$d(x, -) = 2$$

	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8	12	16	20	24
A	4	0	4	8	12	16	20
C	8	4	3	7	8	12	16
G	12	8	7	9	12	14	18
T	16	12	11	7	11	15	14
C	20	16	15	11	7	11	15
C	24	20	19	15	11	13	17
T	28	24	23	19	15	17	13

Alinhamento progressivo

Alinhamento da sequência com o perfil.

(1) AC-TCAT

(2) AG-TCAT

(3) ACGTCCT

$$d(x, y) = 3$$

$$d(x, -) = 2$$

	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8	12	16	20	24
A	4	0	4	8	12	16	20
C	8	4	3	7	8	12	16
G	12	8	7	9	12	14	18
T	16	12	11	7	11	15	14
C	20	16	15	11	7	11	15
C	24	20	19	15	11	13	17
T	28	24	23	19	15	17	13

Alinhamento progressivo

Alinhamento da sequência com o perfil.

(1) AC-TCAT

(1) AC-TCAT

(2) AG-TCAT

(2) A-GTCAT

(3) ACGTCCT

(3) ACGTCCT

$$d(x, y) = 3$$

$$d(x, -) = 2$$

	-	A	C	T	C	A	T
	-	A	G	T	C	A	T
-	0	4	8	12	16	20	24
A	4	0	4	8	12	16	20
C	8	4	3	7	8	12	16
G	12	8	7	9	12	14	18
T	16	12	11	7	11	15	14
C	20	16	15	11	7	11	15
C	24	20	19	15	11	13	17
T	28	24	23	19	15	17	13

Revisão

Alinhamento múltiplo

Programação dinâmica;

Métodos heurísticos:

- Alinhamento progressivo;
 - Matriz de distância;
 - Árvore guia;
 - Alinhamento entre sequências e perfis;
- Alinhamento iterativo;