# Alinhamento de sequências

Alinhamento múltiplo

Prof. Dr. Tetsu Sakamoto Instituto Metrópole 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

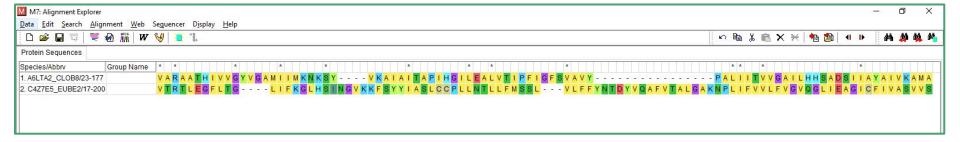
```
: : :
                        -M<mark>P</mark>REDRATW<mark>KSNYFLKIIQLLDDYP</mark>KCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--<mark>P</mark>ALE
                                                                                                                     76
               ------MPREDRATWKSNYFLKIIOLLDDYPKCFIVGADNVGSKOMOOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
                                                                                                                     76
                                                                                                                     76
                     ---MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
                  -----MPREDRATWKSNYFLKIIOLLDDYPKCFIVGADNVGSKOMOOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
                                                                                                                     76
                                                                                                                     76
                       -MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
                                                                                                                     76
                    ----MPREDRATWKSNYFLKIIOLLDDYPKCFIVGADNYGSKOMOOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
                   ----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
                                                                                                                     76
                                                                                                                     76
                    ----M<mark>P</mark>REDRATW<mark>K</mark>SNYFLKIIQLLND<mark>YP</mark>KCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--<mark>P</mark>ALE
                 -----mvrenkaaw<mark>k</mark>aqyfikvvelfdef<mark>pkcfivgadnvgskqmqnir</mark>tsl<mark>rg</mark>l-avvlmgkntmmrkairghlenn--<mark>p</mark>qle
                                                                                                                     76
                  -----MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
                                                                                                                     75
                75
                                                                                                                     76
                -----MAKLSKQQK<mark>K</mark>QMYIEKLSSLIQQYSKILIYHVDNY<mark>GS</mark>NQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV-
                 MIGLAVITTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-
                                                                                                                     79
            ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-
                                                                                                                     80
RLAO SULSO ----MKRLALALKORKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG--
                                                                                                                     80
RLAO AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE
                                                                                                                     86
RLAO PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG--
                                                                                                                     85
RLAO METAC -----MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG--
                                                                                                                     78
            -----MAEERHHTEHIPQWKKDEIENIKELIQSHKYFGMYRIEGILATKIOKIRRDLKDV-AVLKYSRNTLTERALNQLG----
                                                                                                                     78
RLAO ARCFU -----MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG--
                                                                                                                     75
RLAO METKA MAVKAK<mark>GOPPSG</mark>YEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPOLOEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE
                                                                                                                     88
                                                                                                                     74
            -----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD
               ----mitaesehkiapwk<mark>ieevnklkellkngqivalvdmmevparqlqeir</mark>dki<mark>r-gtmtlkmsrntlier</mark>ai<mark>kevaeetgnp</mark>efa
                                                                                                                     82
                   -MIDAKSEHKIA<mark>PWK</mark>IEEVNALKELLKSANVIALIDMMEVPAV<mark>QLQEIR</mark>DKIR-DQMTLKMSRNTLIKRAVEEVAEET<mark>G</mark>NPEFA
                                                                                                                     82
                 ----METKYKAHVA<mark>PWK</mark>IEEVKTLK<mark>G</mark>LIKSK<mark>P</mark>VVAIVDMMDV<mark>PAPQLQ</mark>EIRDKIR-DKVKLRMSRNTLIIRAL<mark>K</mark>EAAEELNN<mark>P</mark>KLA
                                                                                                                     81
                                                                                                                     77
                          -MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAY<mark>P</mark>LSQMRRLIRENGGLLRVSRNTLIELAIKKAAQEL<mark>G</mark>KPELE
                   -----MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKEL<mark>G</mark>KPELE
                                                                                                                     77
                                                                                                                     77
                          ·MAHVAEW<mark>K</mark>KKEVEELANLIKSYPVVALVDVSSMPAY<mark>P</mark>LSQMRRLIRENNGLLRVSRNTLIELAIKKVAQEL<mark>G</mark>KPELE
                          -MAHVAEWKKKEVEELANIIKSYPVIALVDVA<mark>G</mark>VPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE
                                                                                                                     76
                 -MSAESERKTETI<mark>P</mark>EW<mark>K</mark>QEEVDAIV<mark>EMIESYESVGVVNIAGIPSRQLQDMR</mark>RDLH<mark>G</mark>T-AELRVSRNTLLE<mark>R</mark>ALDDVD-----D<mark>GLE</mark>
                                                                                                                     79
                 - MSESEVRQTEVI<mark>P</mark>QWKREEVDELVDFIESYESVGVVGVAGIPSRQL<mark>Q</mark>SMRRELHGS-AAVRMSRNTLVNRALDEVN----
                                                                                                                     79
                 MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----
                                                                                                                     79
                                                                                                                     72
            -----MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS
            72
            -----MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV
                                                                                                                     72
     ruler 1...... 10....... 20....... 30....... 40....... 50........ 60....... 70...... 80....... 90
```

### Para quê...

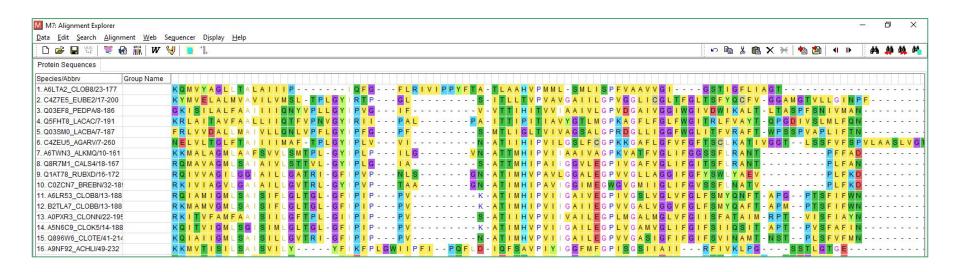
Alinhamento múltiplo de sequências

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

#### Para quê MSA



#### Para quê MSA



### Como fazer...

Alinhamento múltiplo de sequências

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

T C

Existem várias formas de alinhar múltiplas sequências:

AATGATC		Α	Т	G	Α
---------	--	---	---	---	---

$$A - G - A T C$$
  $A - G A T C$ 

Como escolher qual é o melhor?

No caso do alinhamento entre pares de sequências:

AATGATC

AAGG-TC

5 matches

1 mismatch

1 gap

AATGATC

AA-GGTC

5 matches

1 mismatch

1 gap

T C

Existem várias formas de alinhar múltiplas sequências:

AATGATC		Α	Т	G	Α
---------	--	---	---	---	---

$$A - G - A T C$$
  $A - G A T C$ 

Como escolher qual é o melhor?

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

#### Soma dos pares

Considera que o alinhamento múltiplo como  $inom{n}{2}$  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 A	A A A A I	A A I	A A I I	
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<sub>j</sub> = número de ocorrência do aminoácido j na coluna;

C = número de símbolos na coluna;

Α	Α	А	A	Α
Α	Α	A	Α	I
Α	Α	A	A	K
A	Α	Α	I	L
A	Α	I	I	S
Α	I	I	I	W
0	.44	.65	.69	1.79

Logarítmo neperiano

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

#### Programação dinâmica

$$S_{i,j} = max \left\{ egin{array}{l} S_{i-1,j-1} + s(x_i,y_i) \ S_{i-1,j} + g \ S_{i,j-1} + g \end{array} 
ight.$$

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

### Programação dinâmica com 3 sequências

$$S_{i,j} = max \left\{ egin{aligned} S_{i-1,j-1} + s(x_i,y_i) \ S_{i-1,j} + g \ S_{i,j-1} + g \end{aligned} 
ight.$$

$$S_{i-1,j,k} + \delta(x_i, -, -)$$

$$S_{i,j-1,k} + \delta(-, y_i, -)$$

$$S_{i,j,k-1} + \delta(-, -, z_k)$$

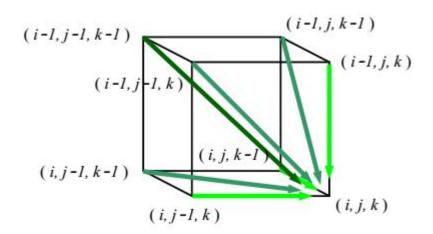
$$S_{i-1,j-1,k} + \delta(x_i, y_j, -)$$

$$S_{i-1,j,k-1} + \delta(x_i, -, z_k)$$

$$S_{i,j-1,k-1} + \delta(-, y_j, z_k)$$

$$S_{i-1,j-1,k-1} + \delta(x_i, y_j, z_k)$$





#### Programação dinâmica

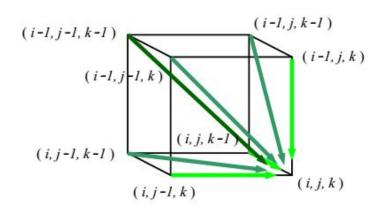
Custo para realizar a programação dinâmica:

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<sup>2</sup> operações.

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

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



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

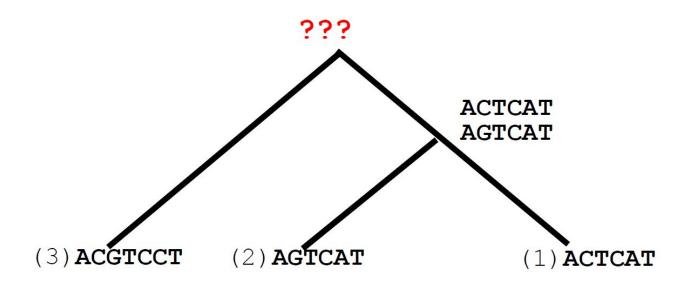
Alinhamento progressivo:

#### 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.

d(x,-) = 2

Construção de uma árvore guia (T).



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

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

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

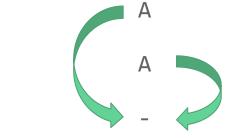
$$d(x,-) = 2$$

- (1) ACTCAT
- (2) AGTCAT
- (2) A-GTCAT
- (3) ACGTCCT

5

5

- (1) AC-TCAT
- (3) ACGTCCT



d(x,y)	=	3
d(x,-)	=	2

	-	Α	С	Т	С	Α	Т
	-	A	G	Т	С	A	Т
-	0						
A							
С							
G							
Т							
С							
С							
Т							



d(x,y)	= 3	
d(x,-)	= 2	

	-	A	С	Т	С	A	Т
	-	A	G	Т	С	A	Т
-	0	4					
A							
С							
G							
Т							
С							
С							
Т							

Alinhamento da sequência com o perfil.

d(x,-) = 2

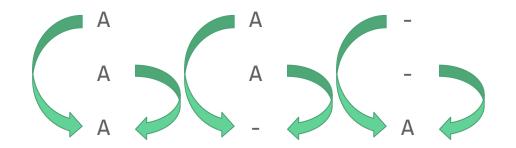


	-	A	С	Т	С	A	Т
	-	A	G	Т	С	A	Т
-	0	4	8				
A							
С							
G							
Т							
С							
С							
Т							

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

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

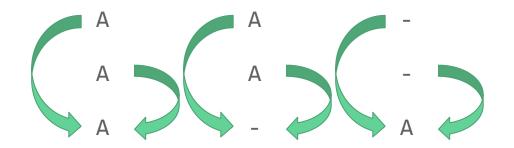
	-	A	С	Т	С	A	Т
	-	Α	G	Т	С	A	Т
-	0	4	8	12	16	20	24
A	4						
С	8						
G	12						
Т	16						
С	20						
С	24						
Т	28						



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

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

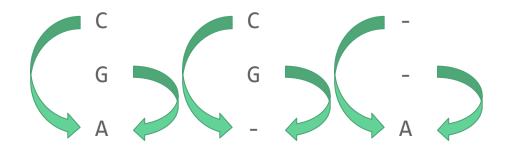
	-	Α	С	Т	С	Α	Т
	-	A	G	Т	С	A	Т
-	0	4	8	12	16	20	24
A	4						
С	8						
G	12						
Т	16						
С	20						
С	24						
Т	28						



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

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

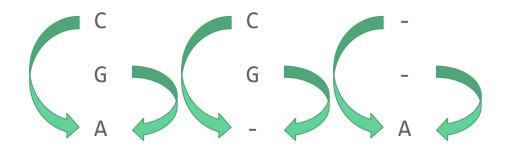
	-	A	С	Т	С	Α	Т
	-	A	G	Т	С	A	Т
-	0	4	8	12	16	20	24
A	4	0					
С	8						
G	12						
Т	16						
С	20						
С	24						
Т	28						



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

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

	-	A	С	Т	С	A	Т
	-	A	G	Т	С	A	Т
-	0	4	8	12	16	20	24
A	4	0					
С	8						
G	12						
Т	16						
С	20						
С	24						
Т	28						



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

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

	-	A	С	Т	С	A	Т
	-	A	G	Т	С	A	Т
-	0	4	8	12	16	20	24
Α	4	0	4				
С	8						
G	12						
Т	16						
С	20						
С	24						
Т	28						

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

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

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

- (1) AC-TCAT
- (2) AG-TCAT
- (3) ACGTCCT

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

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

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

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

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

	=	A	С	Т	С	A	Т
	-	A	G	Т	С	A	Т
-	0	4	8	12	16	20	24
A	4	0	4	8	12	16	20
С	8	4	3	7	8	12	16
G	12	8	7	9	12	14	18
Т	16	12	11	7	11	15	14
С	20	16	15	11	7	11	15
С	24	20	19	15	11	13	17
Т	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;