

Principais algoritmos de alinhamento de sequências genéticas

Alexandre dos Santos Cristino

<http://www.ime.usp.br/~alexsc>

e-mail: alexsc@ime.usp.br

Definição de alinhamento de sequências

- Comparação de duas ou mais sequências por meio de buscas de uma série de caracteres ou padrões de caracteres que estão na mesma ordem.

A L I G N M E N T
| | | | | | | |
- L I G A M E N T

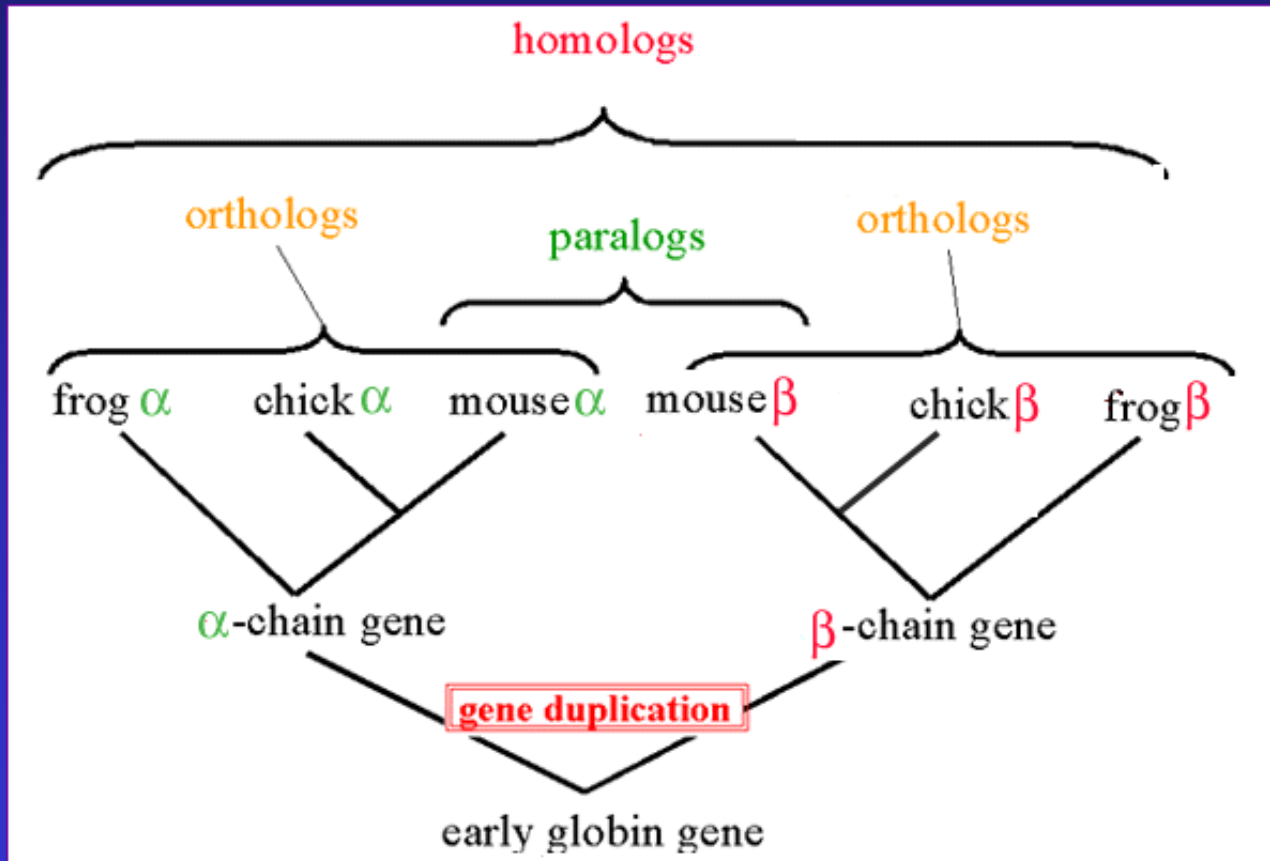
Alinhamento global e local

- Global
 - o alinhamento se estende por toda sequência
- Local
 - o alinhamento localiza fragmentos de sequências que são mais similares

Significado biológico do alinhamento de sequências

- Definindo 3 termos importantes:
 - **identidade** -> refere-se à presença do mesmo ac. nucléico (nt) ou aminoácido (aa) na mesma posição em 2 seqs. alinhadas.
 - **similaridade** -> porcentagem de nt idênticos ou de aa com propriedades químicas semelhantes.
 - **homologia** -> refere-se a relação evolutiva entre as seqs. Duas sequências homólogas derivam da mesma seq. ancestral.
- o alinhamento é muito útil na predição de função, estrutura e inferência filogenética.

Relação entre as sequências



Métodos de alinhamento de sequências

- Alinhamento de pares de seqs.
 - Matriz de pontos (dot matrix).
 - Programação dinâmica.
 - Dicionário de palavras ou k-tuplas (BLAST).
- Alinhamento de múltiplas seqs.

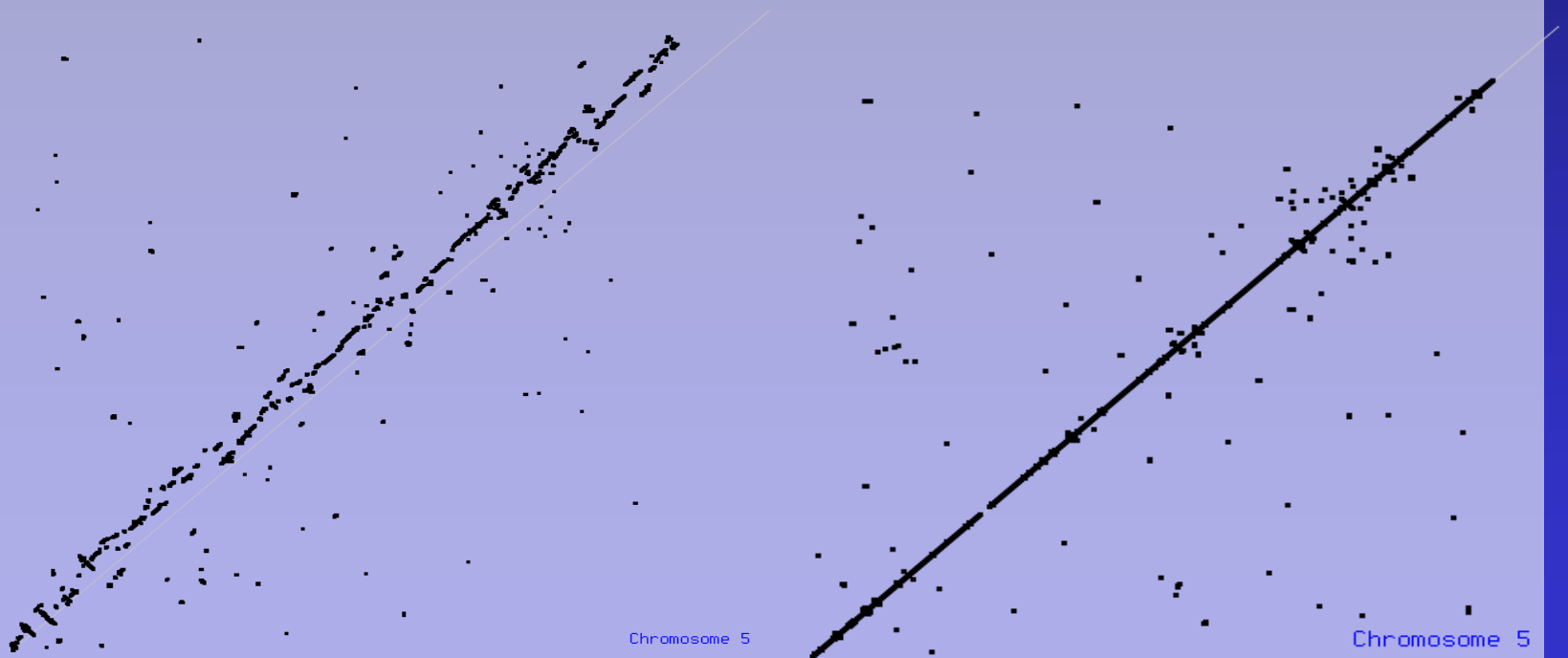
Matriz de pontos (dot plot)

- Comparar duas sequências buscando possíveis alinhamentos de caracteres entre as seqs.

	A	C	C	T	G	A	G	C	T	C	A	C	C	T	G	A	G	T	T	A
A	■					■					■					■				■
C		■	■					■		■		■	■							
C		■	■					■		■		■	■							
T				■					■					■				■	■	
G					■		■								■		■		■	
A	■					■					■					■				■
G					■		■								■		■		■	
C		■	■					■		■		■	■							
T				■					■					■				■	■	
C		■	■					■		■		■	■							
A	■					■					■					■				■
C		■	■					■		■		■	■							
C		■	■					■		■		■	■							
T				■					■					■				■	■	
G					■		■								■		■		■	
A	■					■					■					■				■
G					■		■								■		■		■	
T				■					■					■				■	■	
T				■					■					■				■	■	
A	■					■					■					■				■

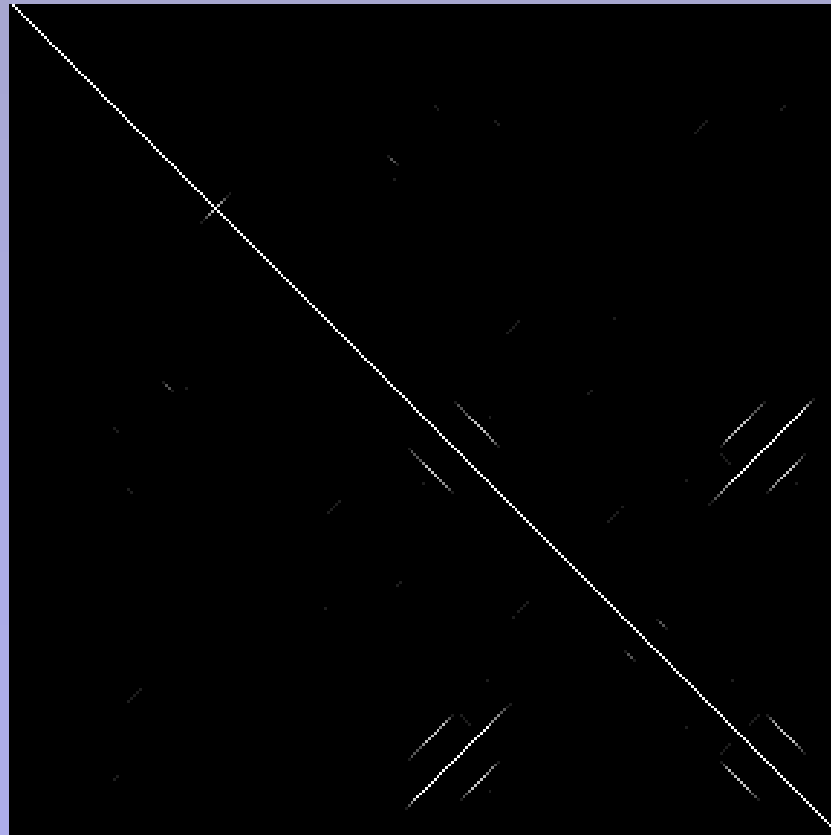
Matriz de pontos (dot plot)

- Comparação de sequências genômicas pareadas



Matriz de pontos (dot plot)

- Sequências repetitivas e inversões



Programação dinâmica (PD)

- Método computacional que calcula o melhor alinhamento possível entre sequências
- Principais variáveis do programa:
 - match
 - mismatch
 - gap

Example de uma Matriz PD

Sequence #1: GAATTCAGTTA; $M = 11$

Sequence #2: GGATCGA; $N = 7$

- Matriz PD:

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

$M+1$ linhas, $N+1$ colunas

Descrição do algoritmo de PD

$$S_{i,j} = \text{MAX}[$$

- $S_{i-1,j-1} + s(a_i, b_j)$ (match/mismatch),
- $S_{i,j-1} + w$ (gap seq #1),
- $S_{i-1,j} + w$ (gap seq #2)

$$]$$

Variáveis do programa:

- $s(a_i b_j) = +5$ if $a_i = b_j$ (match score)
- $s(a_i b_j) = -3$ if $a_i \neq b_j$ (mismatch score)
- $w = -4$ (gap penalty)

Preenchendo a Matriz PD (alinhamento global)

- $S_{1,2} = \text{MAX}[S_{0,1} - 3, S_{1,1} - 4, S_{0,2} - 4] = \text{MAX}[-4 - 3, 5 - 4, -8 - 4] = \text{MAX}[-7, 1, -12] = 1$

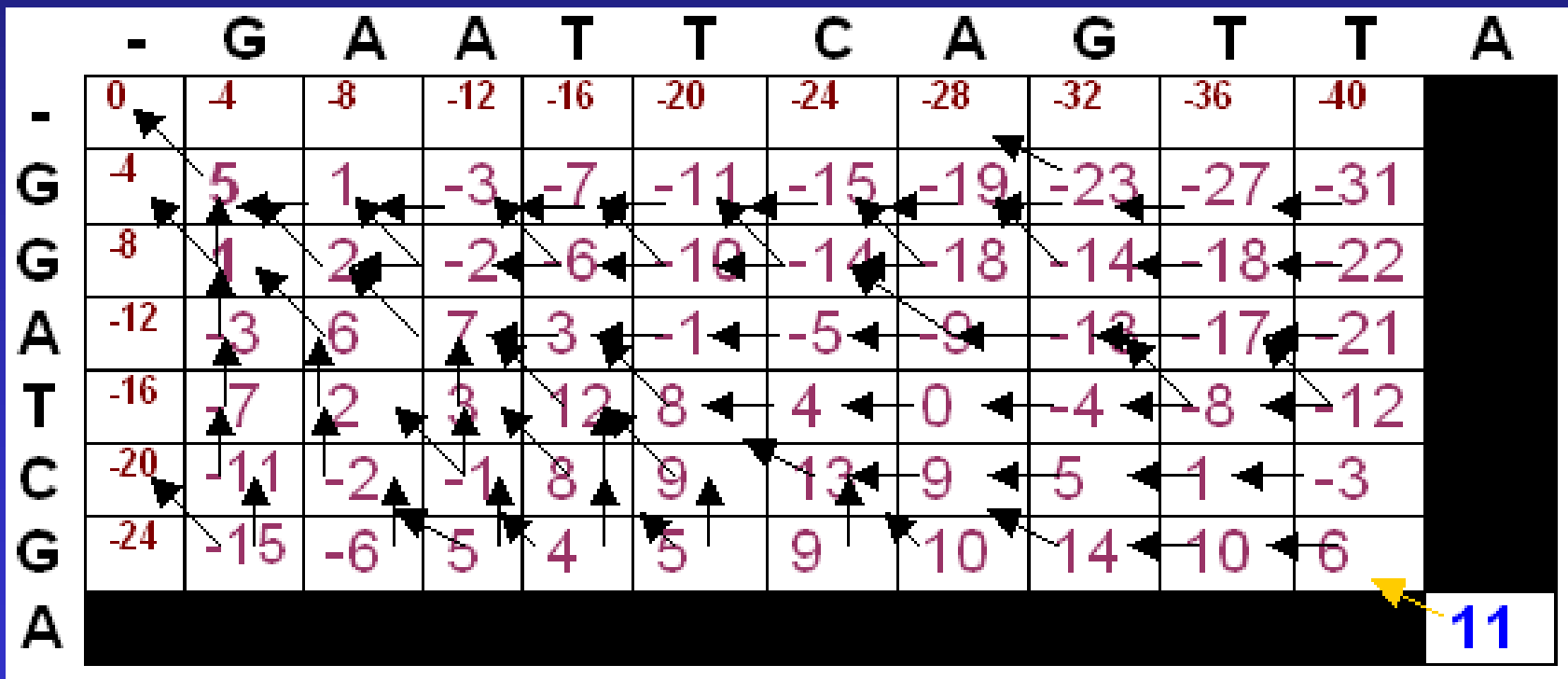
	-	G	A	A	T	T	C	A	G	T	T	A
-	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44
G	-4	5	1									
G	-8											
A	-12											
T	-16											
C	-20											
G	-24											
A	-28											

Diagram illustrating the Dynamic Programming matrix for global sequence alignment. The matrix shows scores for aligning the sequence "GATTCAGTTA" (rows) against the sequence "GATA" (columns). The top-left cell (0,0) is 0. The first row contains values: -4, -8, -12, -16, -20, -24, -28, -32, -36, -40, -44. The first column contains values: -4, -8, -12, -16, -20, -24, -28. The cell (1,1) contains 5, and the cell (1,2) contains 1. Colored arrows indicate transitions: a black arrow from (0,0) to (1,1), a red arrow from (0,1) to (1,2), a blue arrow from (1,1) to (1,2), and a green arrow from (1,1) to (1,2).

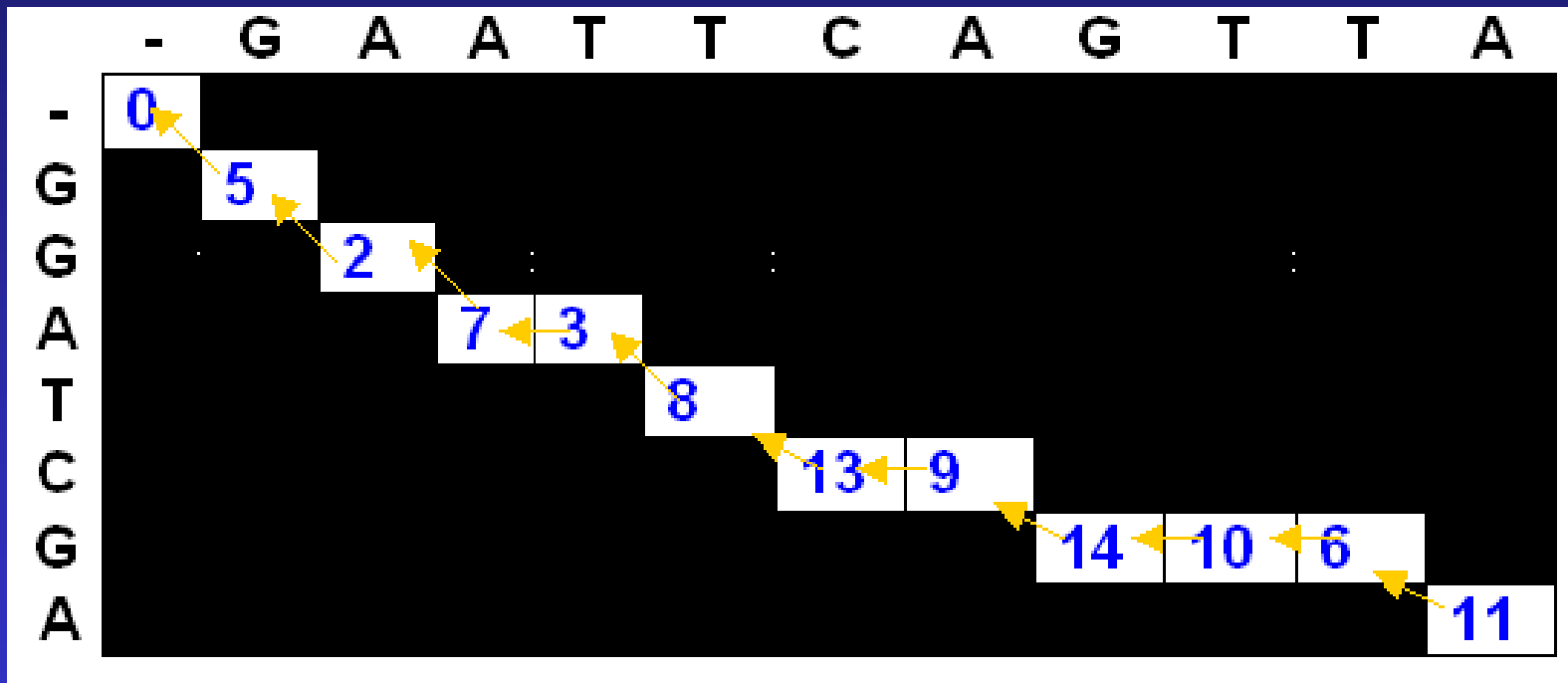
Matriz PD preenchida (alinhamento global)

	-	G	A	A	T	T	C	A	G	T	T	A
-	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44
G	-4	5	1	-3	-7	-11	-15	-19	-23	-27	-31	-35
G	-8	1	2	-2	-6	-10	-14	-18	-14	-18	-22	-26
A	-12	-3	6	7	3	-1	-5	-9	-13	-17	-21	-17
T	-16	-7	2	3	12	8	4	0	-4	-8	-12	-16
C	-20	-11	-2	-1	8	9	13	9	5	1	-3	-7
G	-24	-15	-6	-5	4	5	9	10	14	10	6	2
A	-28	-19	-10	-1	0	1	5	14	10	11	7	11

Trace back (alinhamento global)



Trace back (alinhamento global)



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

Verificando o score de alinhamento

G	A	A	T	T	C	A	G	T	T	A
G	G	A	-	T	C	-	G	-	-	A
+	-	+	-	+	+	-	+	-	-	+
5	3	5	4	5	5	4	5	4	4	5

$$5 - 3 + 5 - 4 + 5 + 5 - 4 + 5 - 4 - 4 + 5 = 11 \checkmark$$

Alinhamento local (Smith-Waterman)

- Variação do algoritmo de Needleman-Wunsch.
- Possui 2 modificações:
 - valor negativo para mismatch
 - valor da matriz de score negativo e trocado por zero (se inicia um novo alinhamento)

Preenchendo a Matriz PD (alinhamento local)

- $S_{1,2} = \text{MAX}[S_{0,1} - 3, S_{1,1} - 4, S_{0,2} - 4, 0] = \text{MAX}[0 - 3, 5 - 4, 0 - 4, 0] = \text{MAX}[-3, 1, -4, 0] = 1$

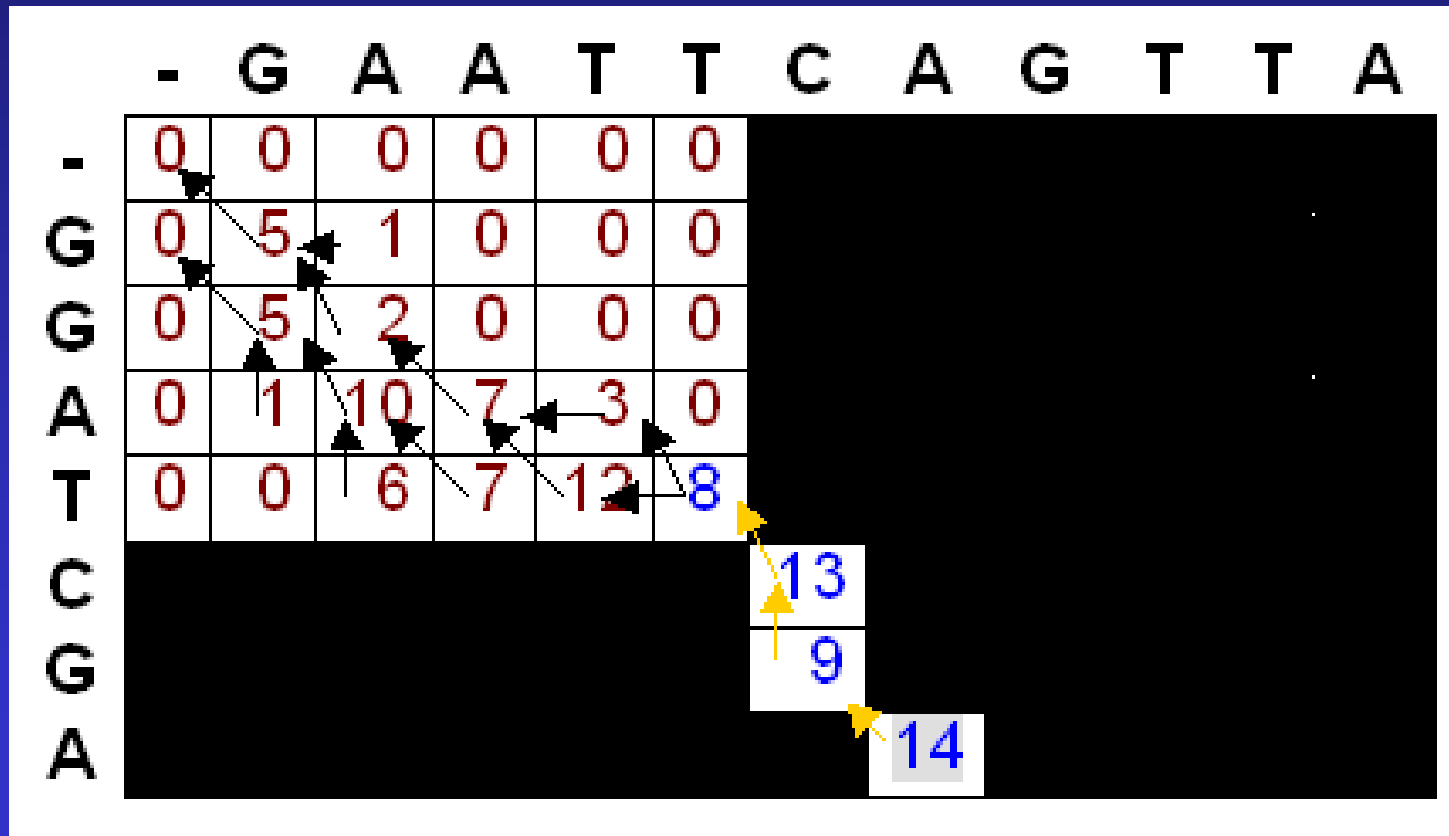
	-	G	A	A	T	T	C	A	G	T	T	A
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	1									
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

Diagram illustrating the dynamic programming matrix for local sequence alignment. The matrix shows scores for alignments between the sequences -GAGATTCAGTTA and -GAGATTCAGTTA. The cell (1,2) contains the score 5, and the cell (2,3) contains the score 1. Arrows indicate the backpointers: a red arrow from (1,2) to (1,1), a blue arrow from (1,2) to (2,3), a green arrow from (2,3) to (2,2), and a purple arrow from (2,3) to (1,3).

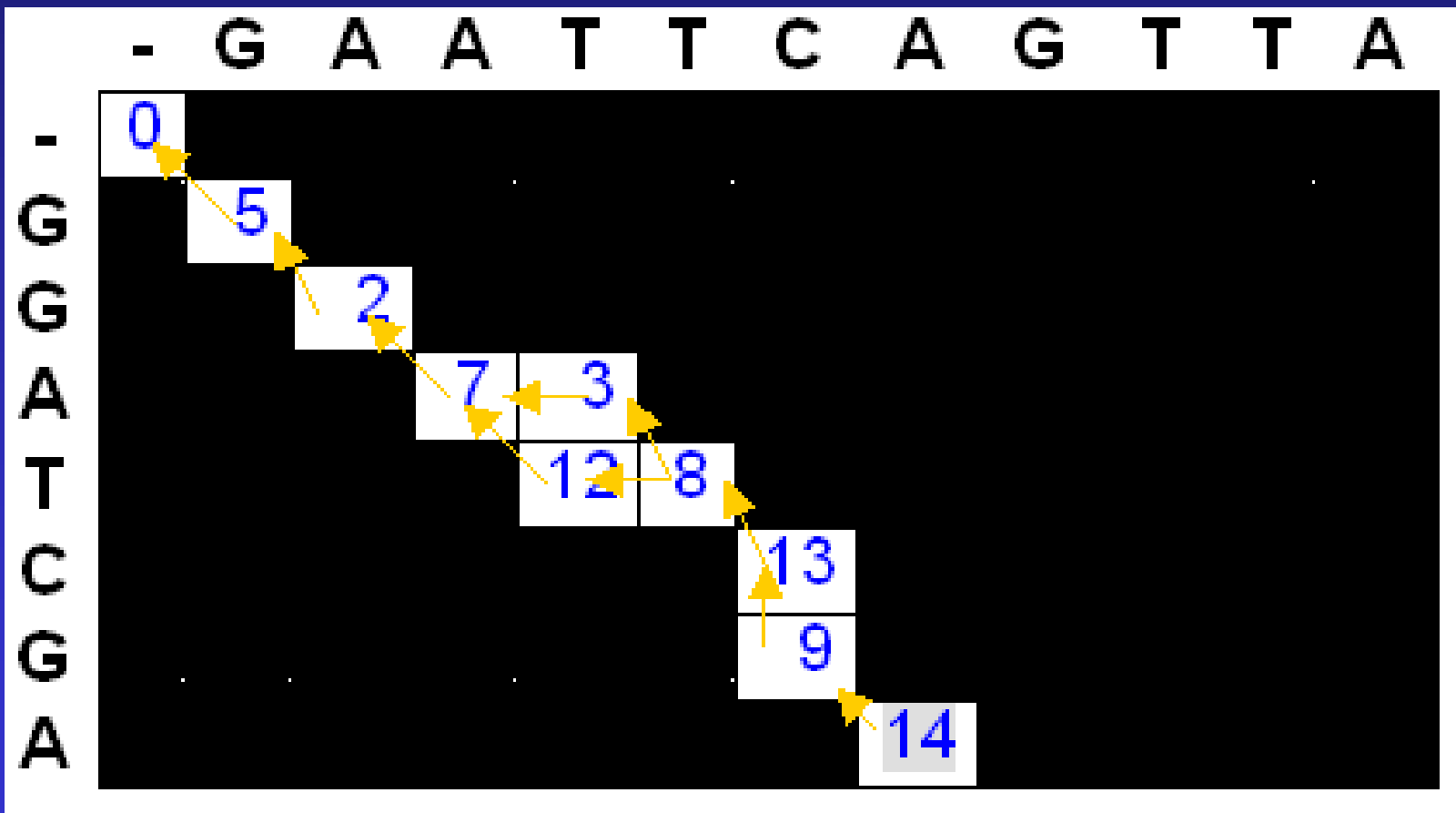
Matriz PD preenchida (alinhamento local)

	-	G	A	A	T	T	C	A	G	T	T	A
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	1	0	0	0	0	0	5	1	0	0
G	0	5	2	0	0	0	0	0	5	2	0	0
A	0	1	10	7	3	0	0	5	1	2	0	5
T	0	0	6	7	12	8	4	1	2	6	7	3
C	0	0	2	3	8	9	13	9	5	2	3	4
G	0	5	1	0	4	5	9	10	14	10	6	2
A	0	1	10	6	2	1	4	14	10	11	7	11

Trace back (alinhamento global)



Trace back (alinhamento global)



Melhores alinhamentos locais

G A A T T C - A

| | | | |

G G A T - C G A

+ - + + - + - +

5 3 5 5 4 5 4 5

G A A T T C - A

| | | | |

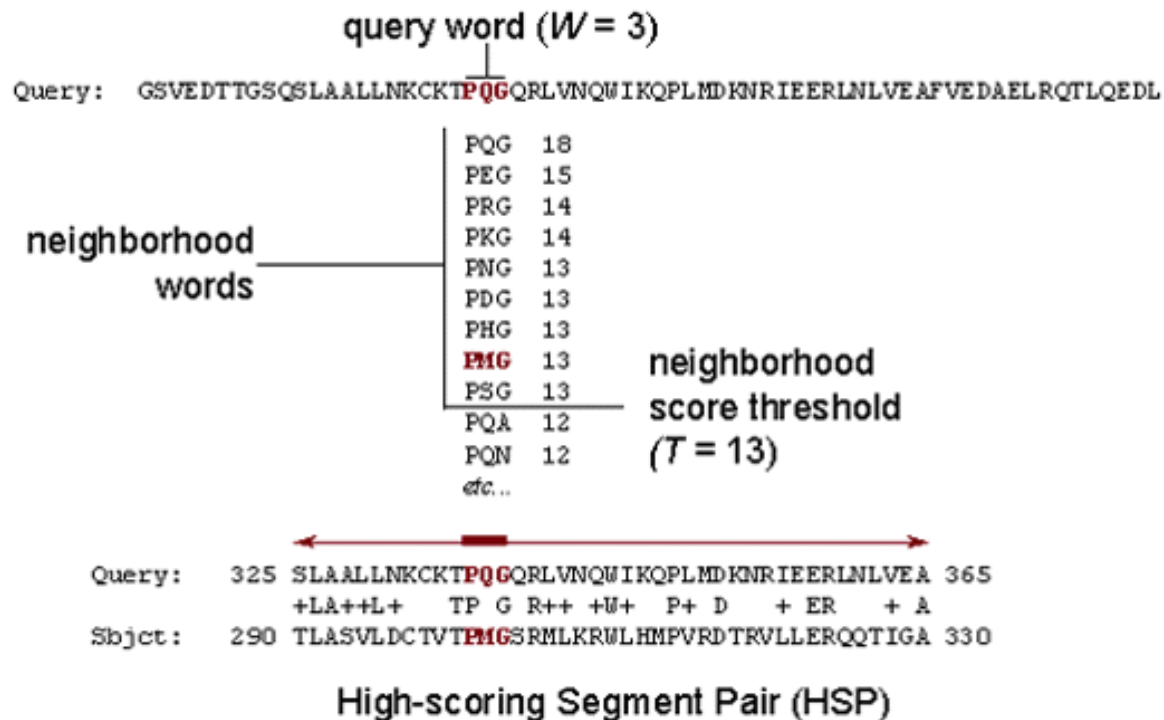
G G A - T C G A

+ - + - + + - +

5 3 5 4 5 5 4 5

K-tuplas (BLAST)

The BLAST Search Algorithm



Referências

- Gibbs, A. J. & McIntyre, G. A. (1970) The diagram method for comparing sequences. its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* **16**, 1-11.
- Mount, D. (?) Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Lab. Press.
- Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195-197.
- Needleman S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.
- Gibas, C. and Jambeck, P. (2001) Desenvolvendo bioinformática. O'reilly.

Sites

<http://www.ime.usp.br/~durham>

<http://kbrin.kwing.louisville.edu/~rouchka/CECS694/>

<http://www.lbm.fmvz.usp.br>