# 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
```

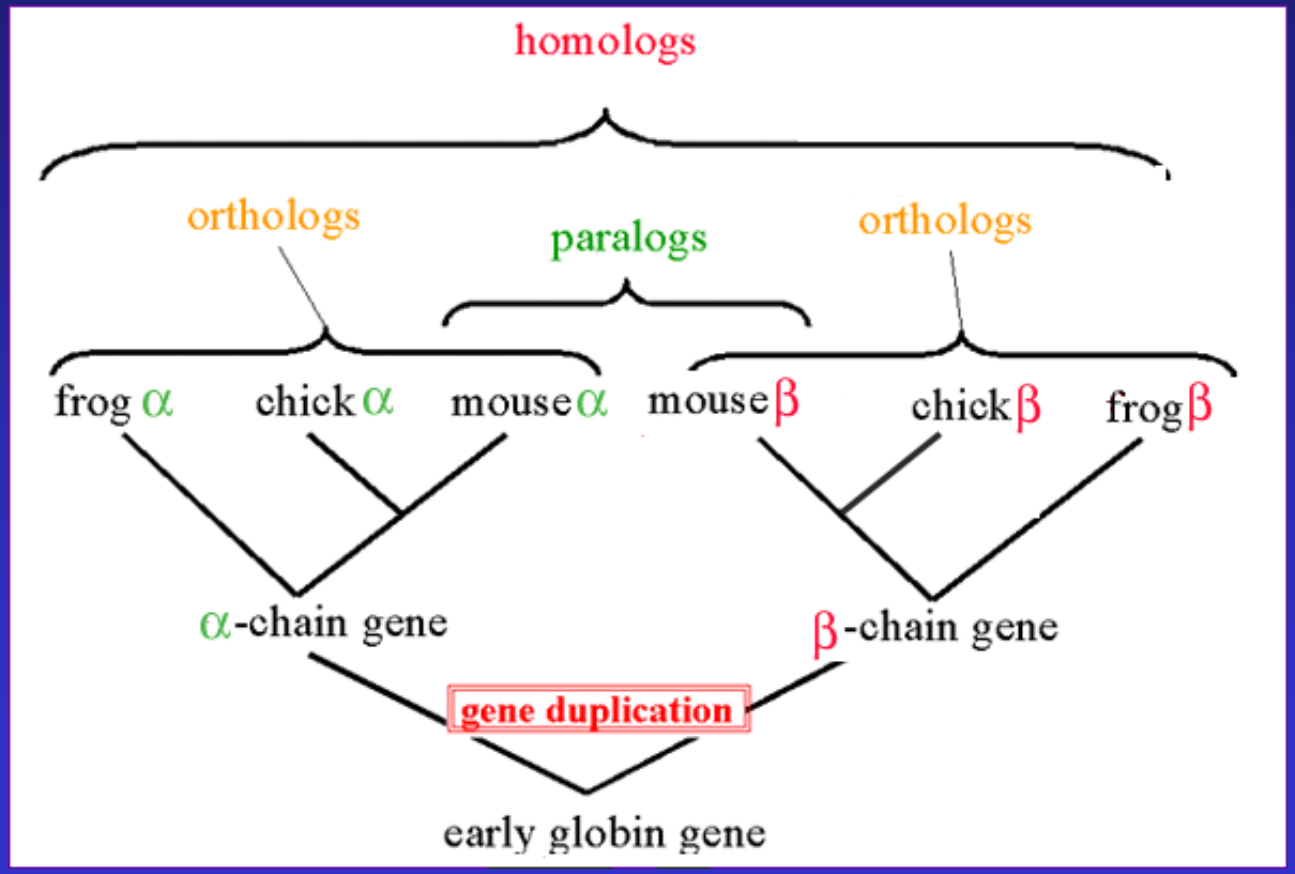
# Alinhamento global e local

- Global
  - o alinhamento se extende 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. ancentral.
- o alinhamento é muito útil na predição de função, estrutura e inferência filogenética.

# Relação entre as sequências

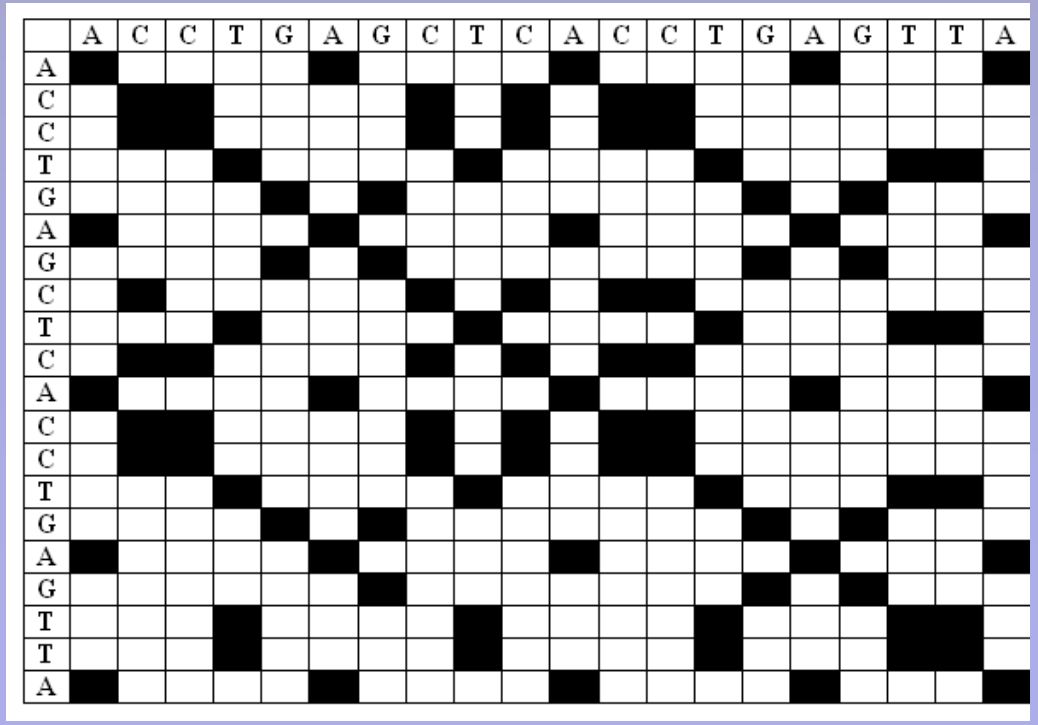


# Métodos de alinhamento de sequências

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

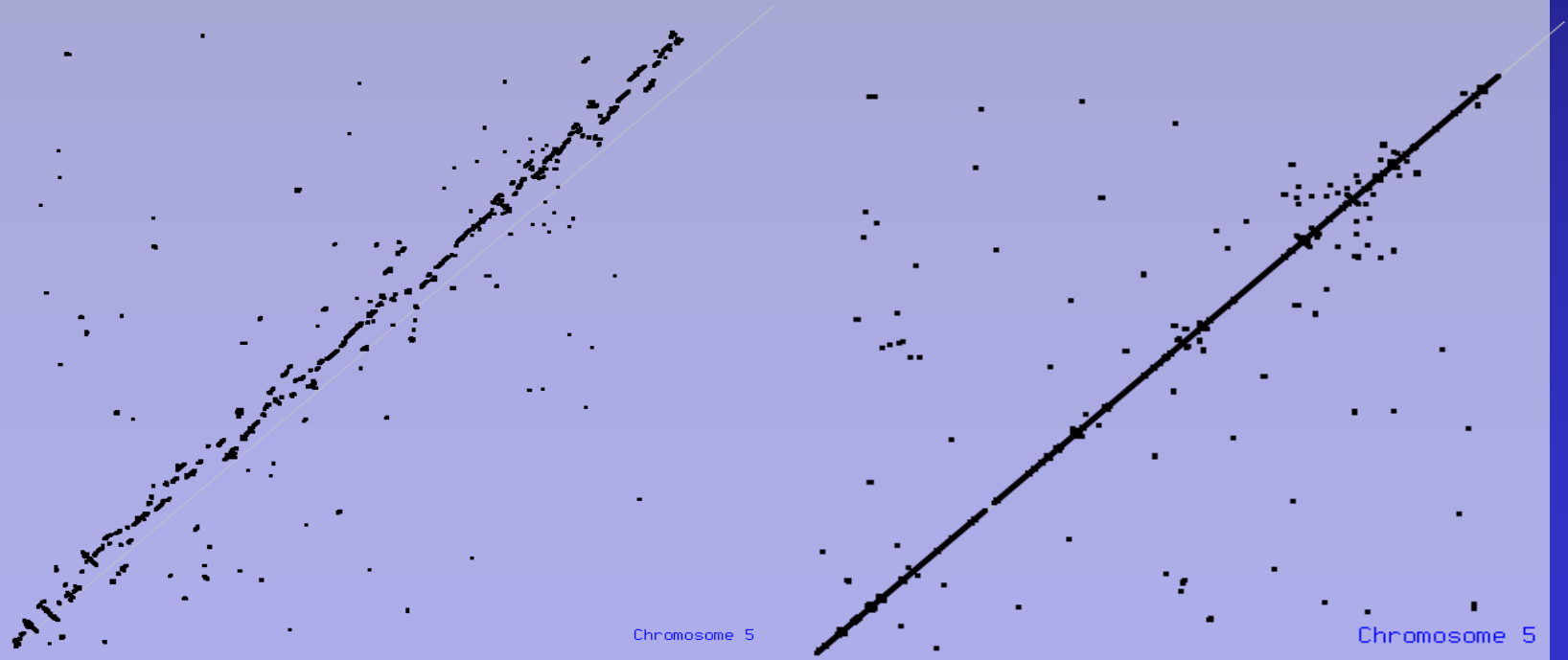
# Matriz de pontos (dot plot)

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



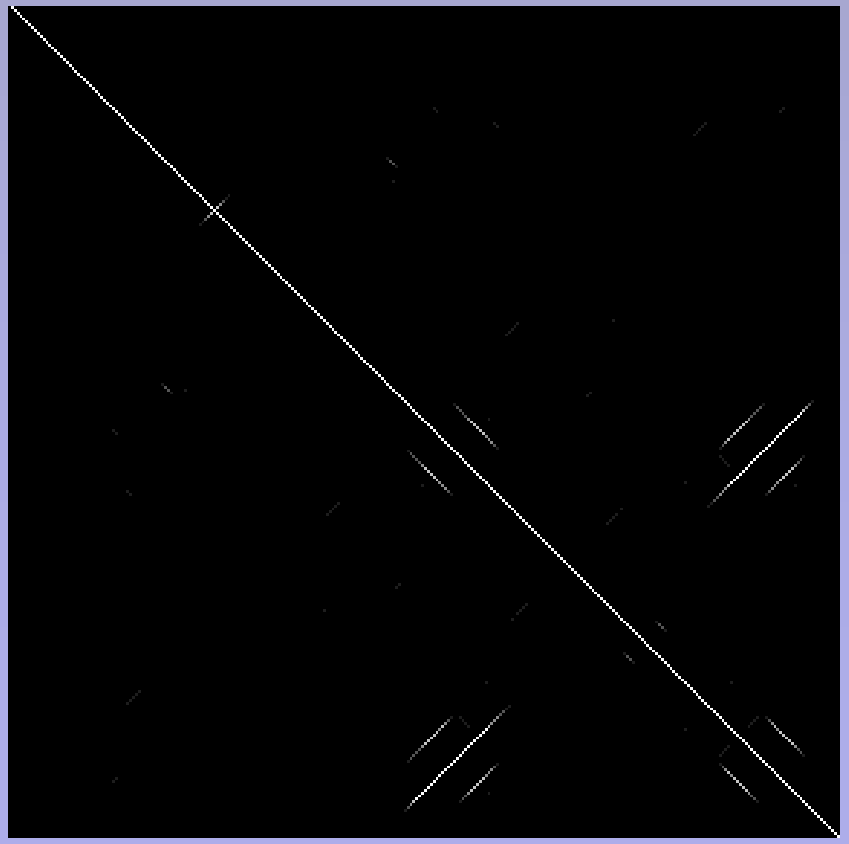
# Matriz de pontos (dot plot)

- 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

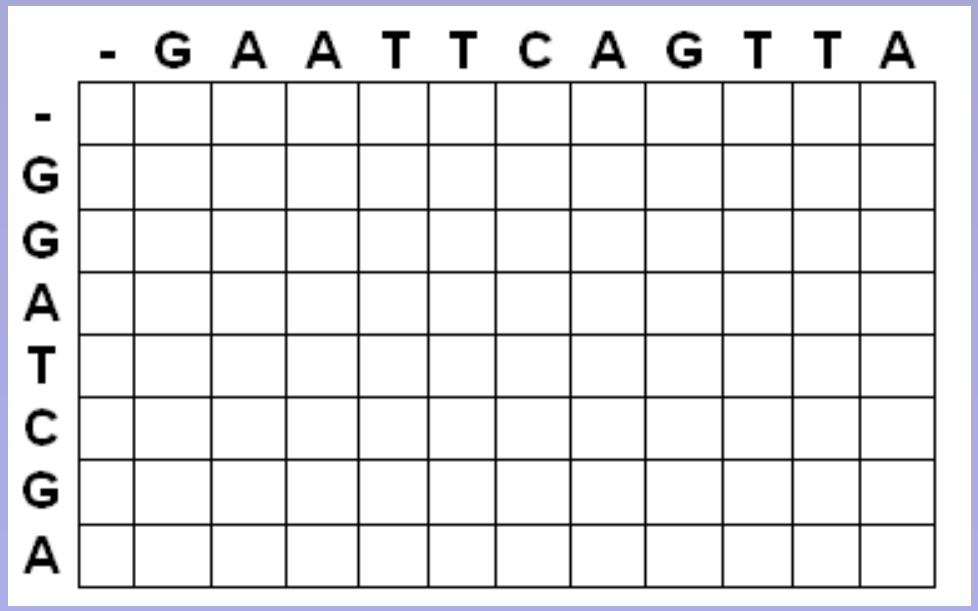
- Principais variáveis do programa:
  - match
  - mismatch
  - gap

# Example de uma Matriz PD

Sequence #1: GAATTCAGTTA; M = 11

Sequence #2: GGATCGA; N = 7

- Matriz PD:




M+1 linhas, N+1 colunas

# Descrição do algoritmo de PD

$S_{i,j}$ = **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)

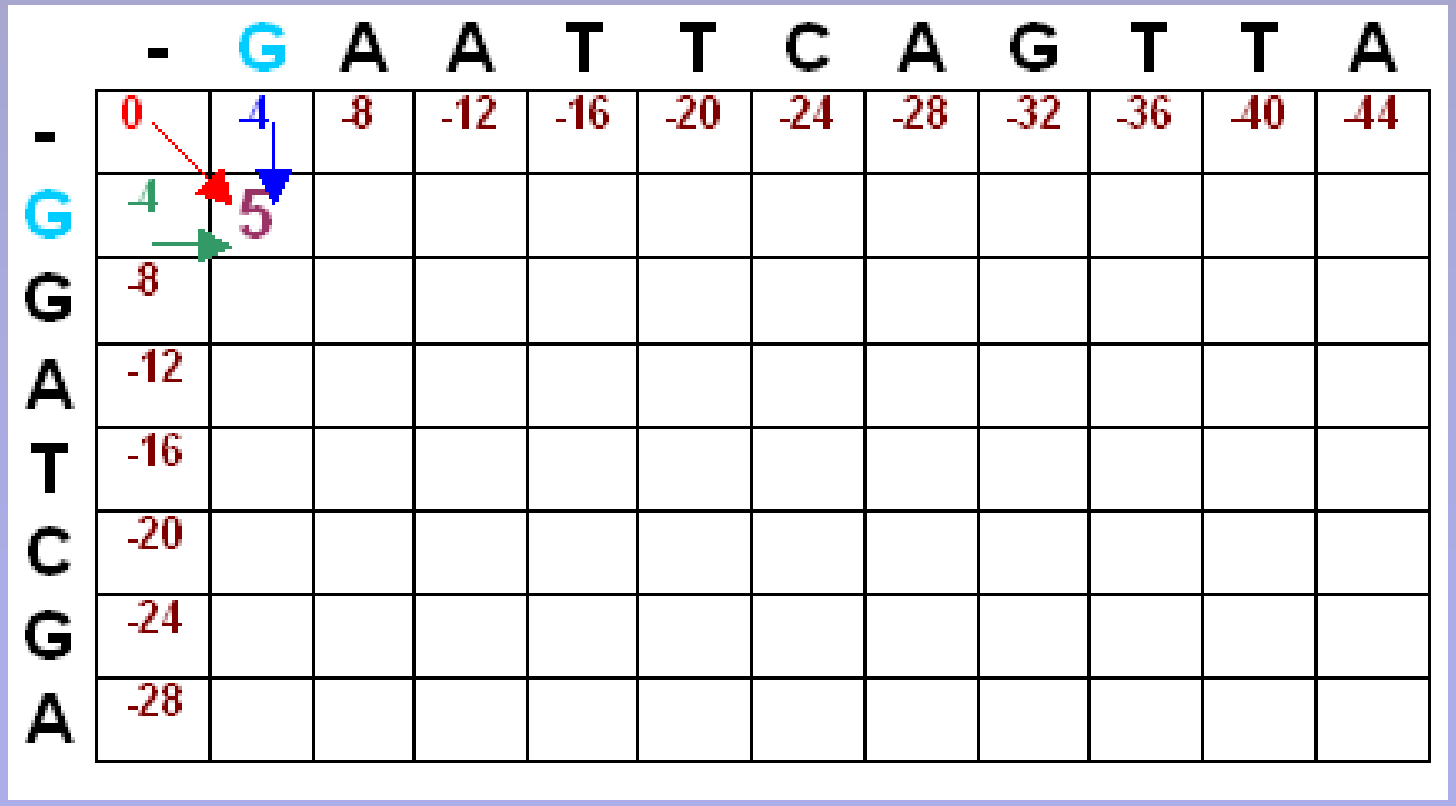
# Alinhamento global (Needleman-Wunsch)

- Inicialização da 1a. linha e 1a. coluna:
  - $S_{i,0} = w * i$
  - $S_{0,j} = w * j$

| | - | 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 | | | | | | | | | | | |
| G | -8 | | | | | | | | | | | |
| A | -12 | | | | | | | | | | | |
| T | -16 | | | | | | | | | | | |
| C | -20 | | | | | | | | | | | |
| G | -24 | | | | | | | | | | | |
| A | -28 | | | | | | | | | | | |

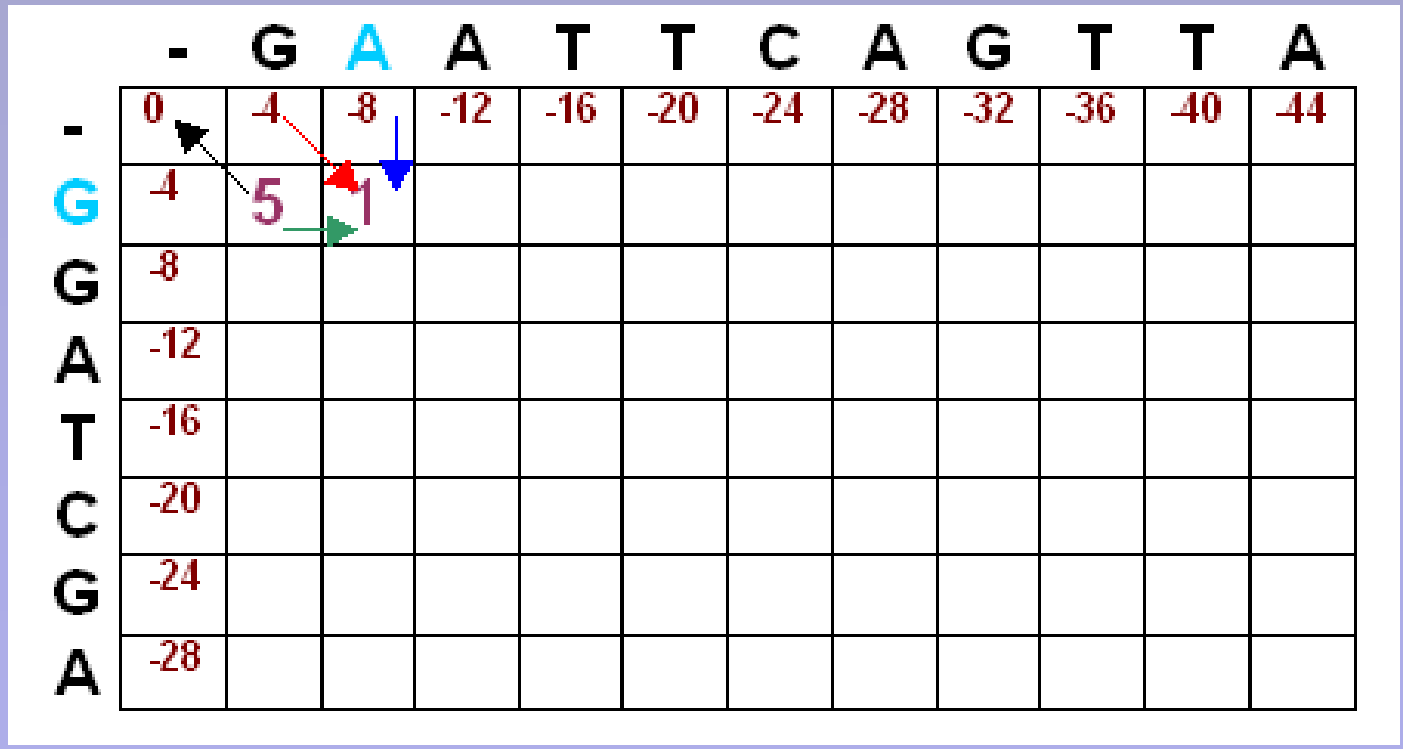
# Preenchendo a Matriz PD (alinhamento global)

- $S_{1,1} = MAX[S_{0,0} + 5, S_{1,0} - 4, S_{0,1} - 4] = MAX[5, -8, -8]$

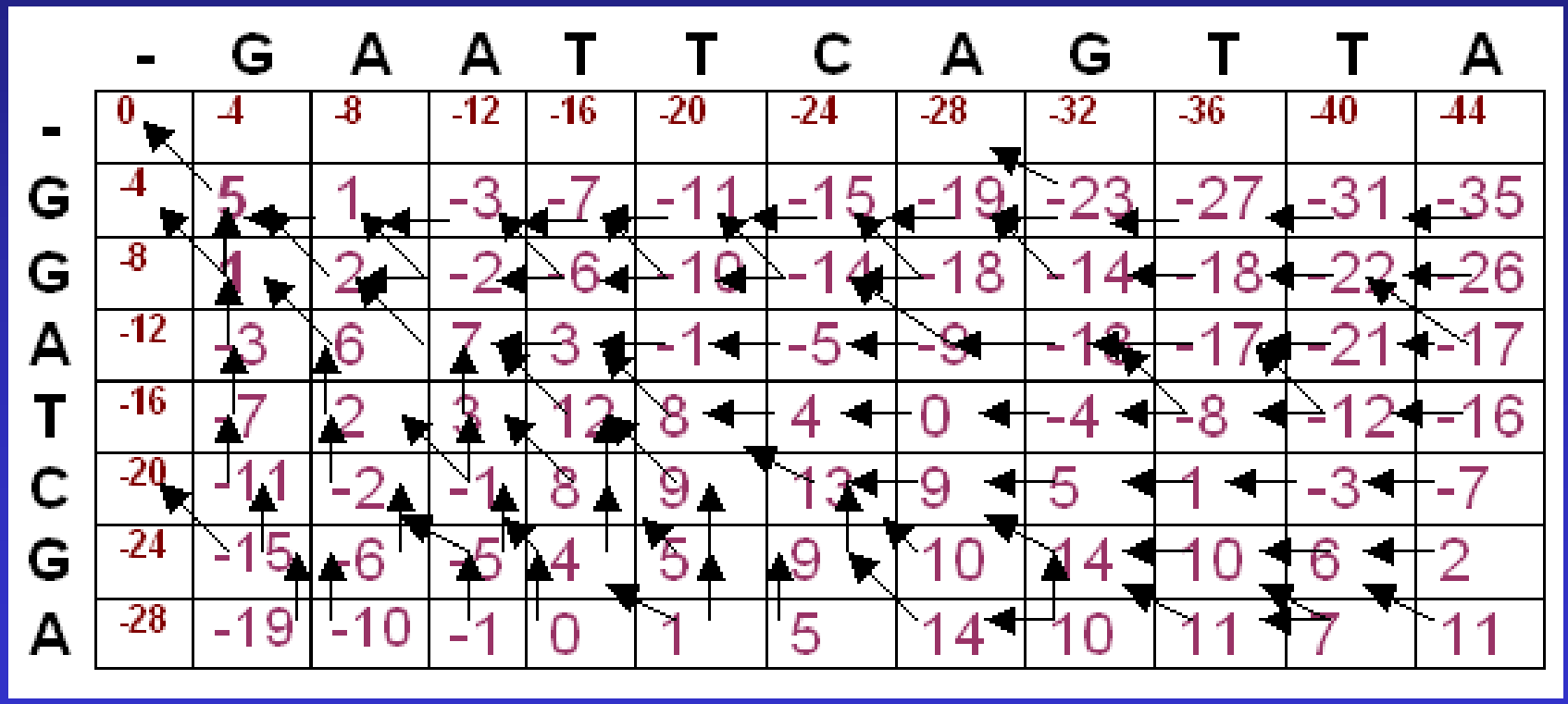


# Preenchendo a Matriz PD (alinhamento global)

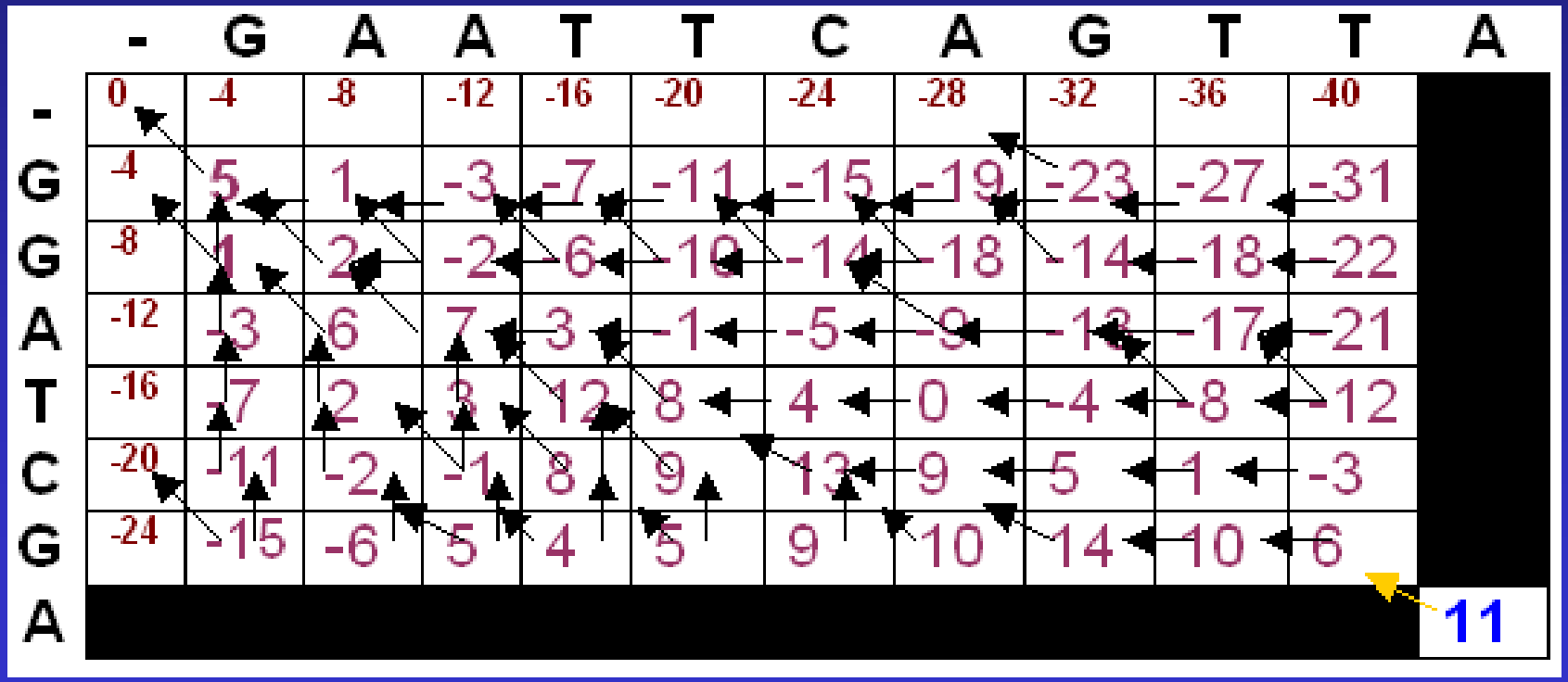
- $S_{1,2} = MAX[S_{0,1} - 3, S_{1,1} - 4, S_{0,2} - 4] = MAX[-4 - 3, 5 - 4, -8 - 4] = MAX[-7, 1, -12] = 1$



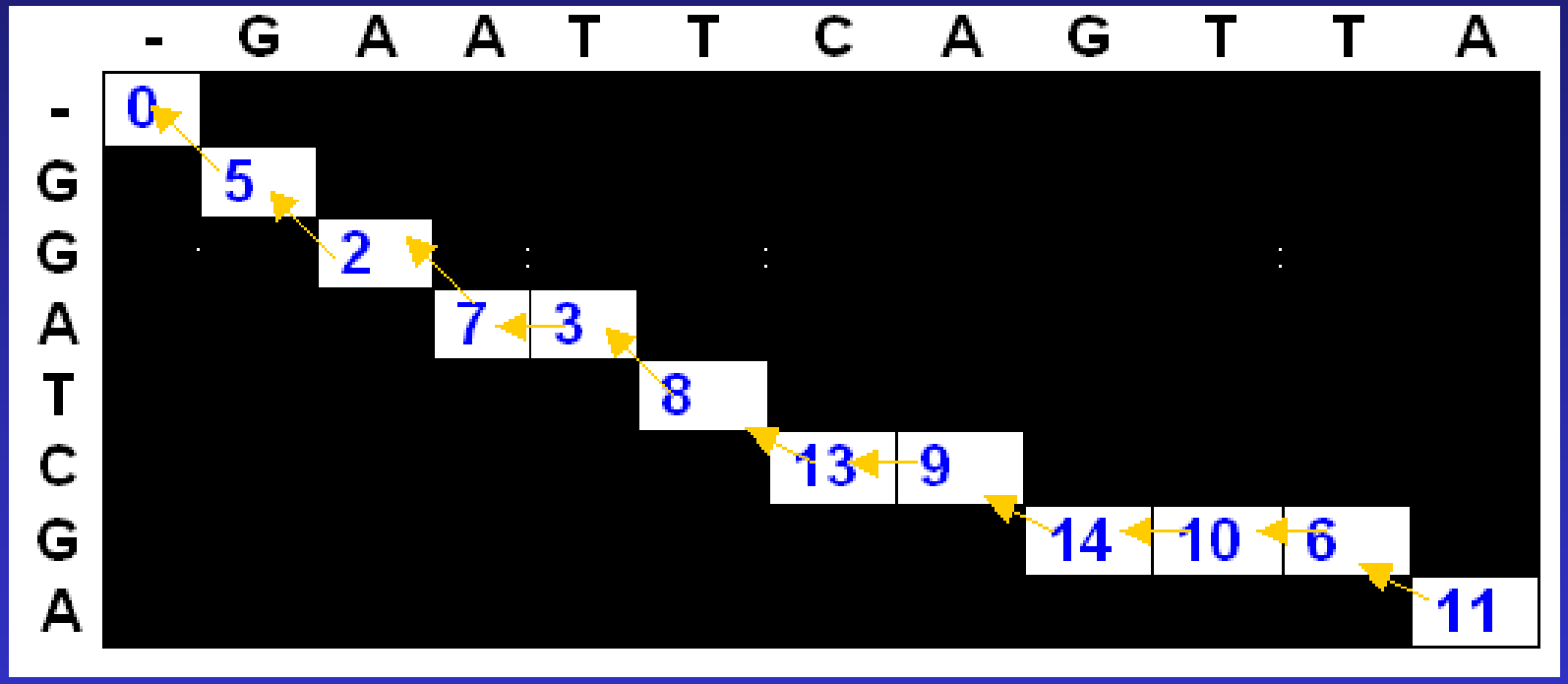
# Matriz PD preenchida (alinhamento global)



# 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
```

$5 - 3 + 5 - 4 + 5 + 5 - 4 + 5 - 4 - 4 + 5 = 11$ ✓

# Alinhamento local (Smith-Waterman)

- Variação do algoritmo de Needleman-Wunsch.

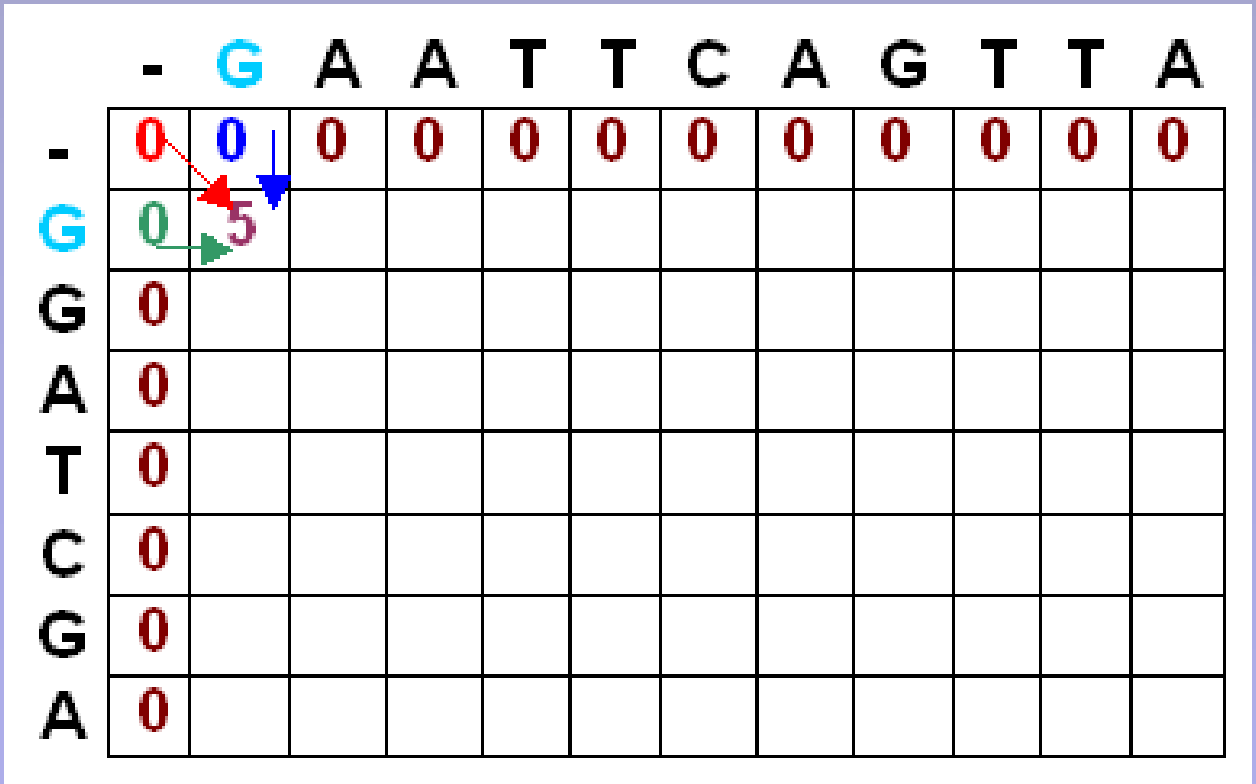
- Possui 2 modificações:
  - valor negativo para mismatch
  - valor da matriz de score negativo e trocado por zero (se inicia um novo alinhamento)

# Início do 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 | | | | | | | | | | | |
| G | 0 | | | | | | | | | | | |
| A | 0 | | | | | | | | | | | |
| T | 0 | | | | | | | | | | | |
| C | 0 | | | | | | | | | | | |
| G | 0 | | | | | | | | | | | |
| A | 0 | | | | | | | | | | | |

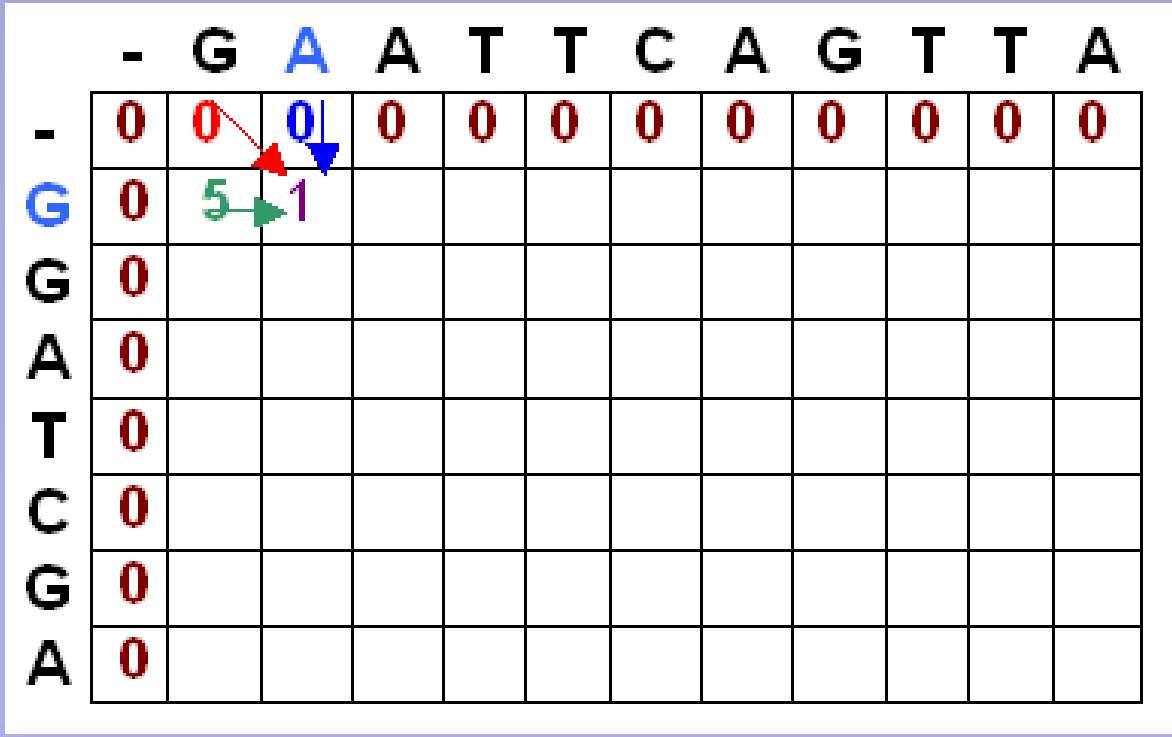
# Preenchendo a Matriz PD (alinhamento local)

- $S_{1,1} = MAX[S_{0,0} + 5, S_{1,0} - 4, S_{0,1} - 4,0] = MAX[5, -4, -4, 0] = 5$

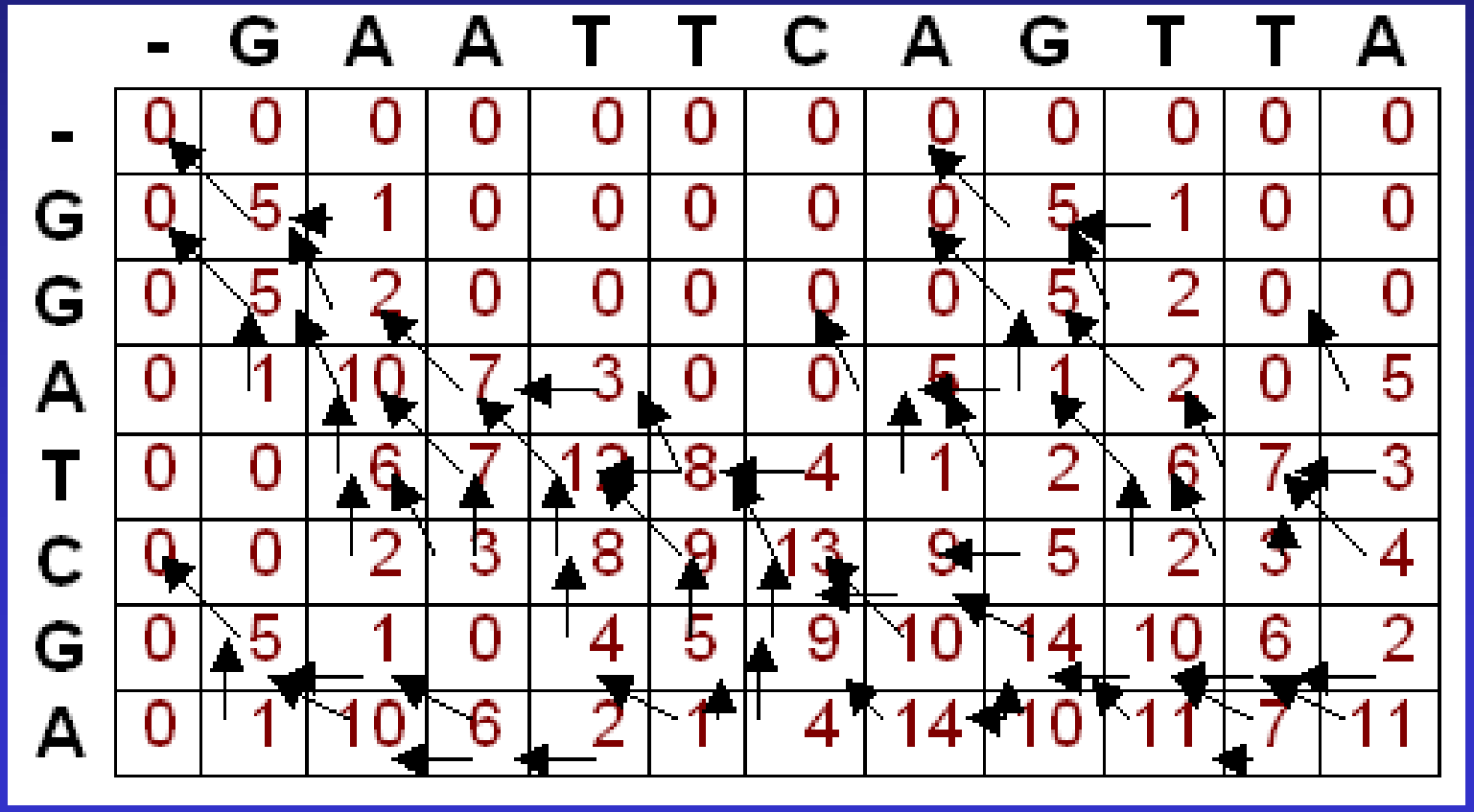


# Preenchendo a Matriz PD (alinhamento local)

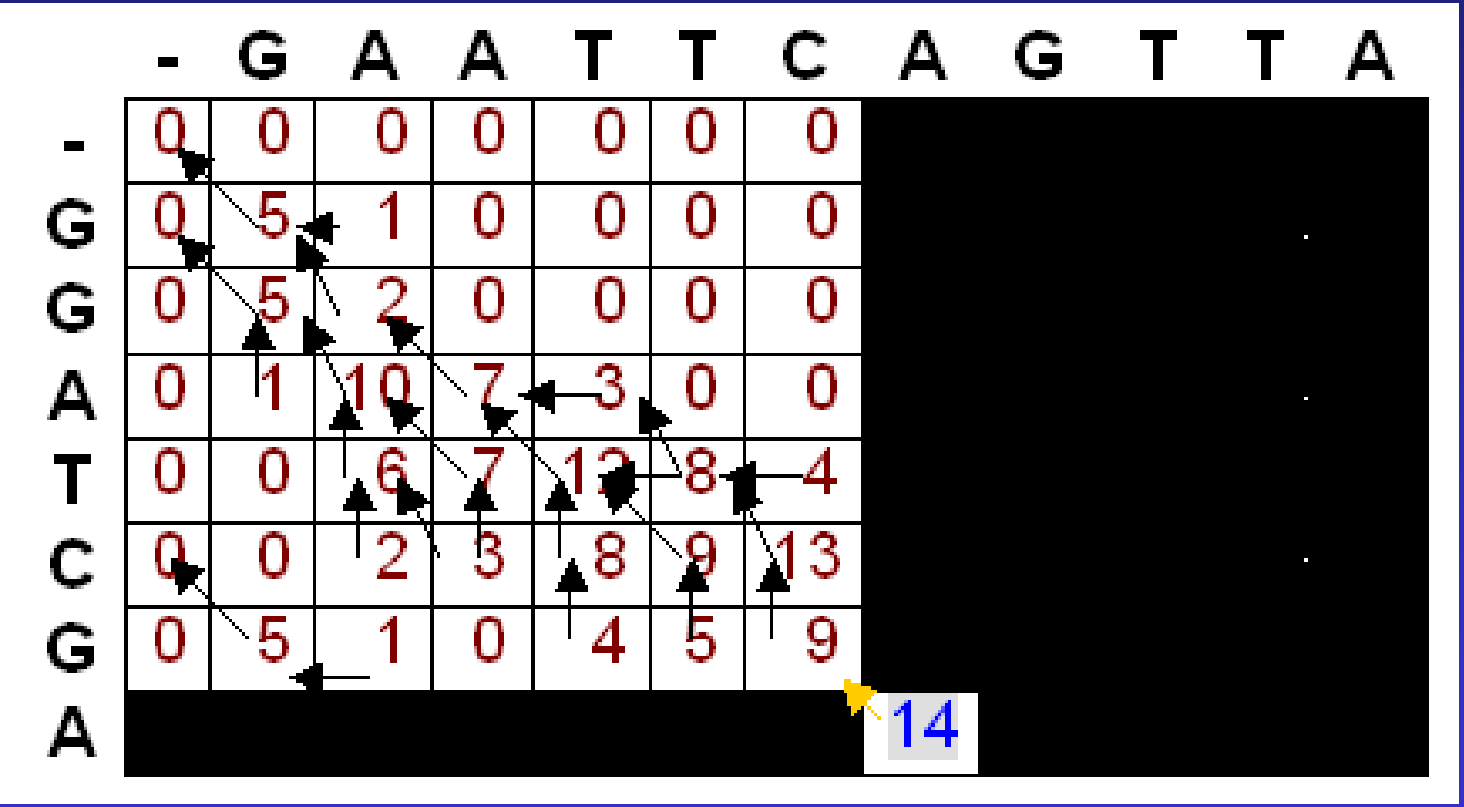
- $S_{1,2} = MAX[S_{0,1} - 3, S_{1,1} - 4, S_{0,2} - 4, 0] = MAX[0 - 3, 5 - 4, 0 - 4, 0] = MAX[-3, 1, -4, 0] = 1$



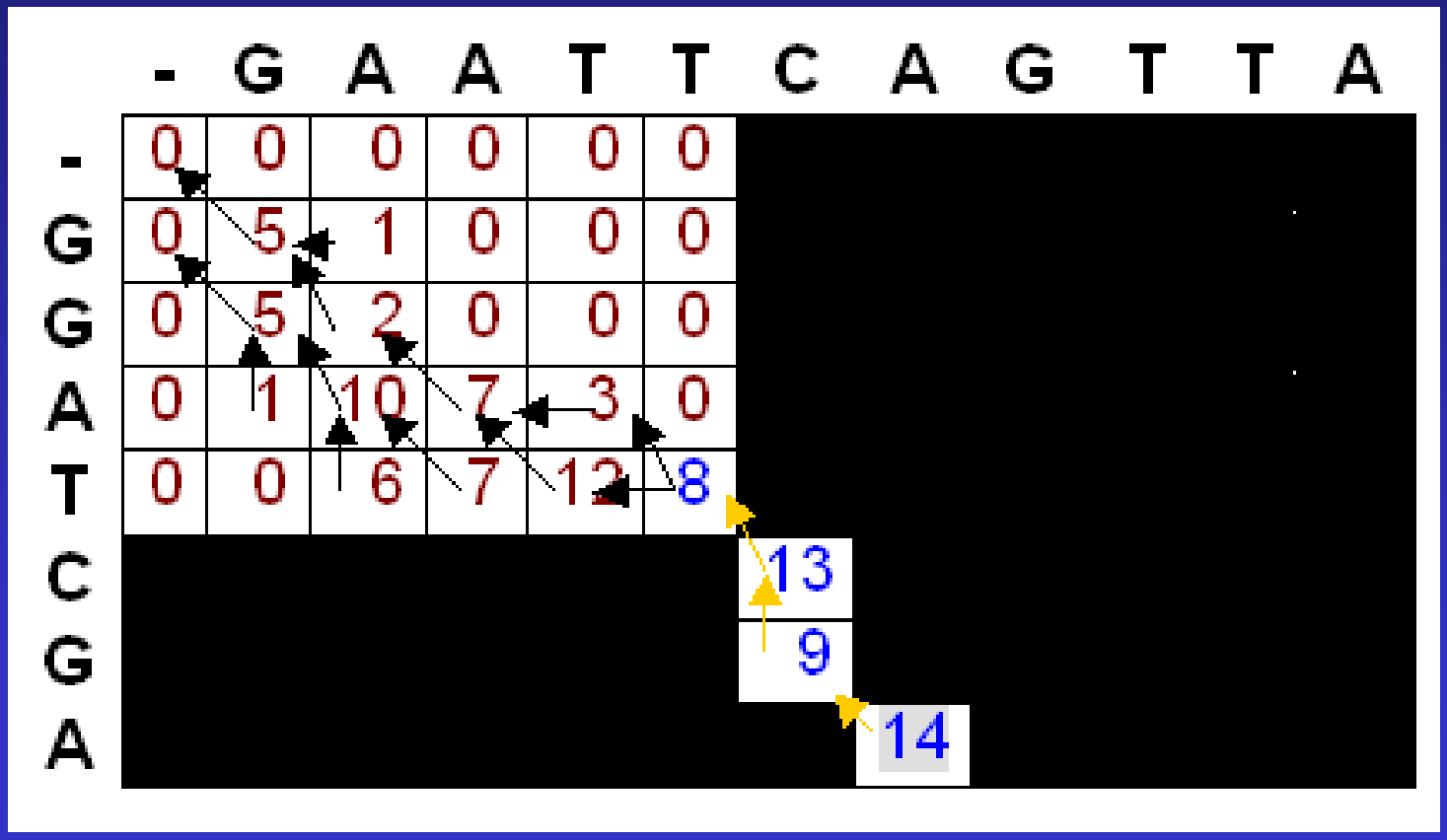
# Matriz PD preenchida (alinhamento local)



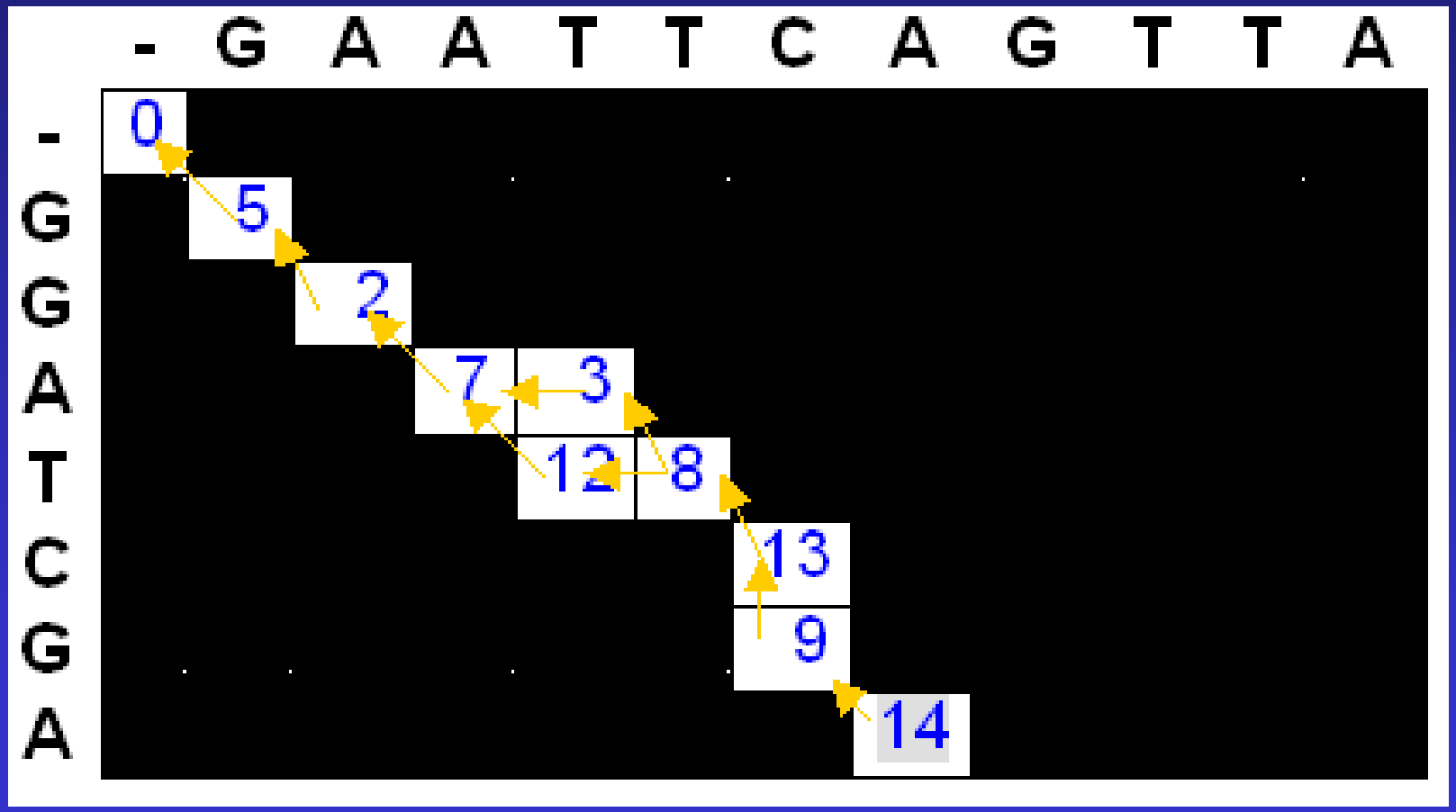
# Trace back (alinhamento global)



# Trace back (alinhamento global)



# Trace back (alinhamento global)



# Melhores alinhamentos locais

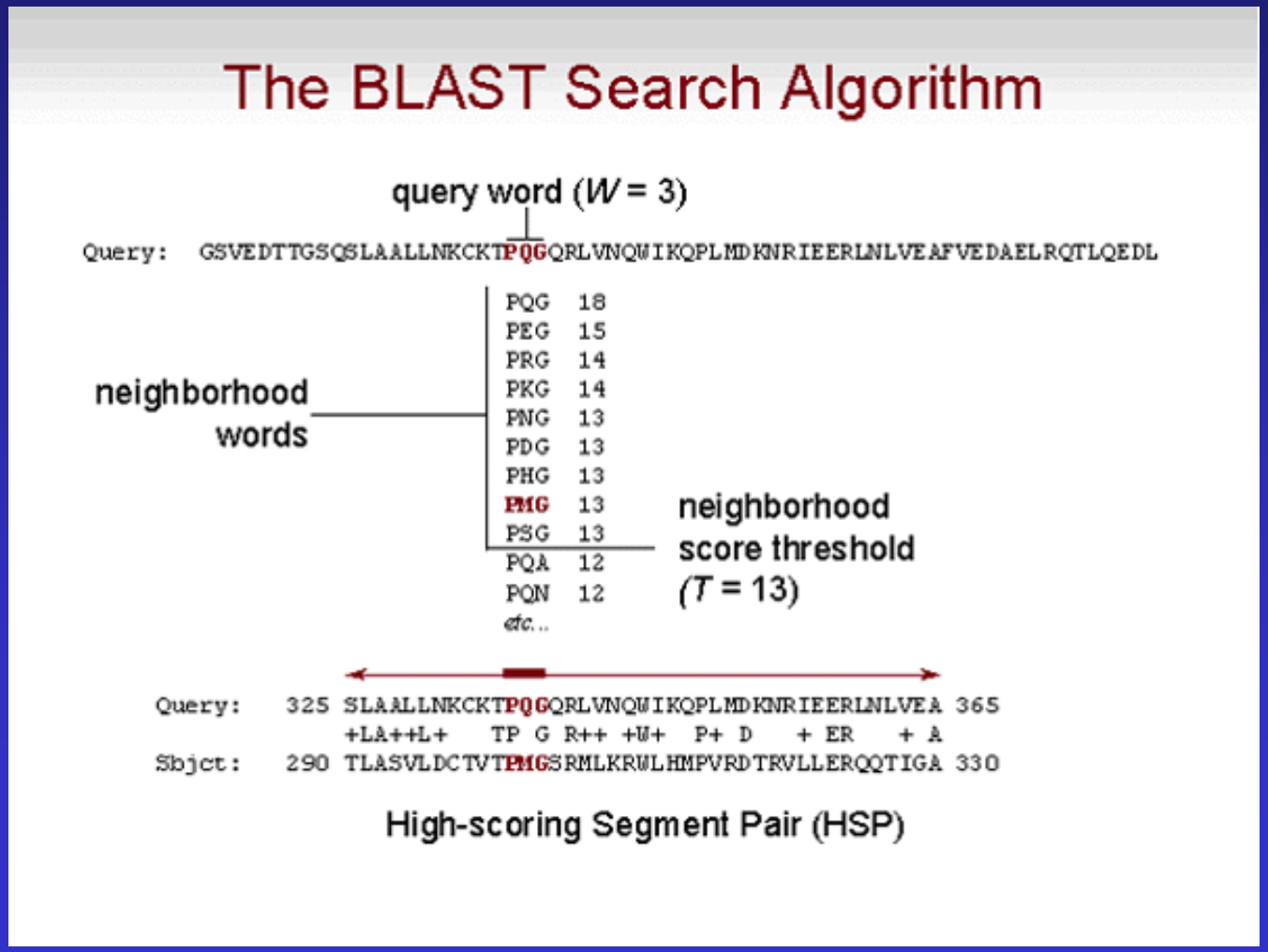
```
G A A T T C - A
|   | |   |   |
G G A T - C G A

+ - + + - + - +
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)



# 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