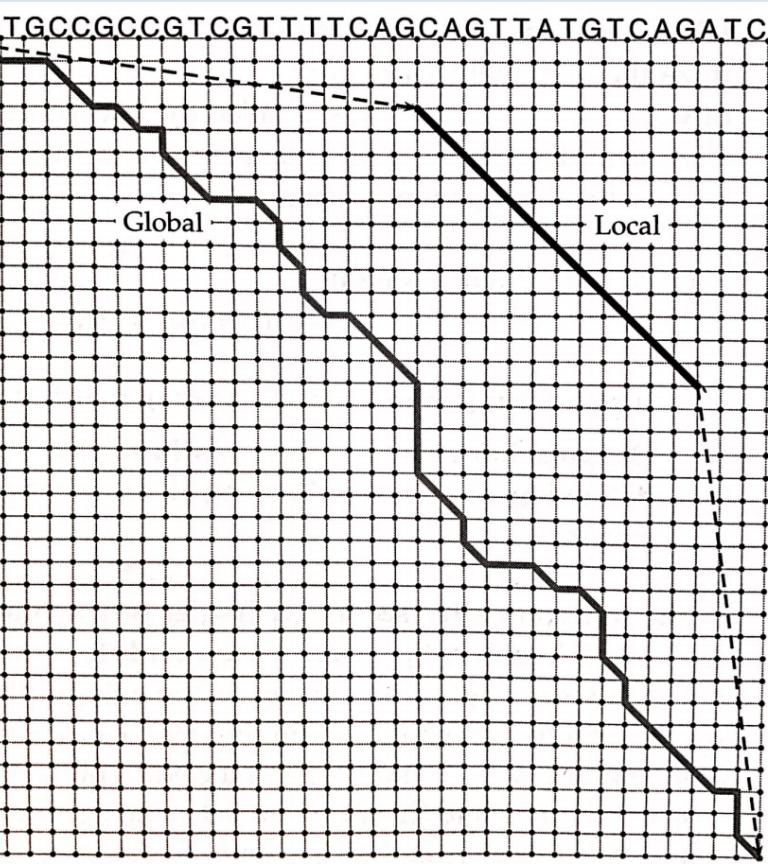


# Algoritmo de Smith-Waterman

CCCGCACCCT  
CTGAATGGCGAATG  
GAAGCGGTGCCGGA  
GAGGCCGATACTGT  
GCACGGTTACGATG  
TCCCCATTACGGTCA  
CCCGTTGT

```
products: storeProducts

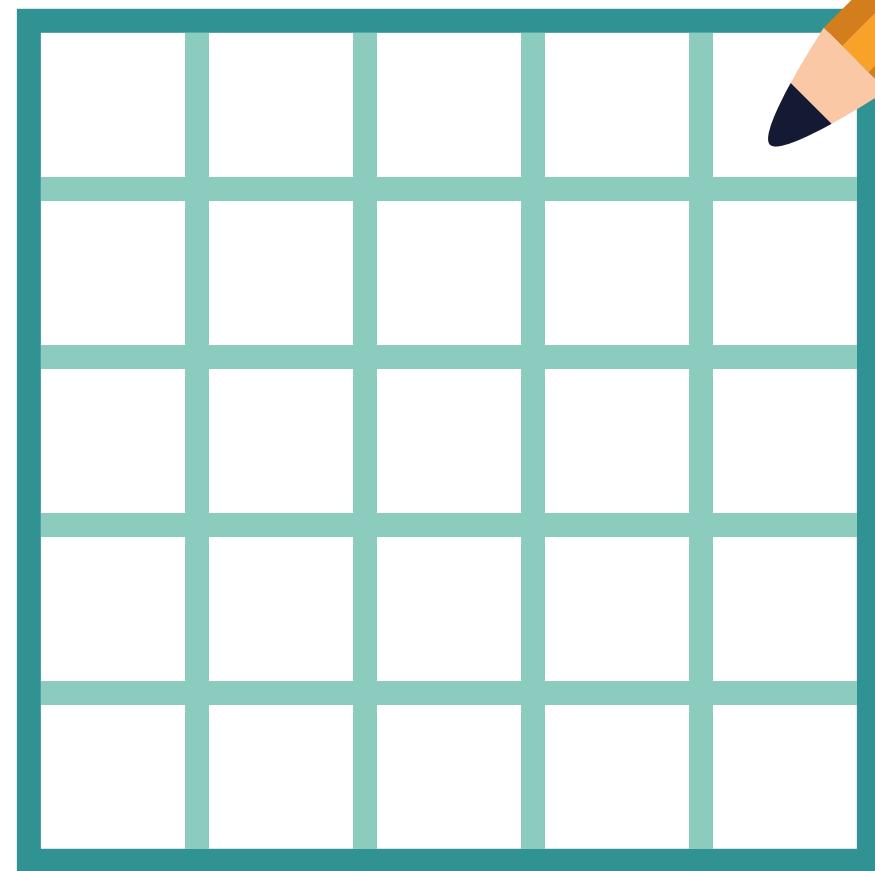
render() {
  return (
    <React.Fragment>
      <div className="py-5">
        <div className="container">
          <Title name="our" title="Our Products" />
          <div className="row">
            <ProductConsumer>
              {(value) => {
                console.log(value)
              }}
            </ProductConsumer>
          </div>
        </div>
      </React.Fragment>
```





## DESAFIO #1

Tente realizar alinhamentos globais usando o algoritmo de Needleman-Waterman manualmente



>sequencia1

DRQTAKAAGTD

>sequencia2

ERQLAKAAAGTD

9 pontos

sequência 1

sequência 2

- D **RQ** - T A K A A G T D

E - **RQL** - A K A A G T D

	D	R	Q	T	A	K	A	A	G	T	D
0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0
R	0	0	1\	1	1	1	1	1	1	1	1
Q	0	0	1	2\	2	2	2	2	2	2	2
L	0	0	1	2	2	2	2	2	2	2	2
A	0	0	1	2	3\	3	3\	3	3	3	3
K	0	0	1	2	2	3\	4\	4	4	4	4
A	0	0	1	2	2	3\	4	5\	5\	5	5
A	0	0	1	2	2	3\	4	5\	6\	6	6
G	0	0	1	2	2	3\	4	5	6	7\	7
T	0	0	1	2	3\	3	4	5	6	7	8\
D	0	1\	1	2	3	3	4	5	6	7	8\



## DESAFIO #2

Tente realizar alinhamentos globais usando a ferramenta do NCBI\*

National Center for Biotechnology Information

Sign in to NCBI

Home Recent Results Saved Strategies

Needleman-Wunsch Global Align Nucleotide Sequences

Reset page Bookmark

New columns added to the Description Table  
Click 'Select Columns' or 'Manage Columns'.

Query Sequence

Enter accession number, gi, or FASTA sequence  Clear  
Input limited to 100,000 letters for either input sequence. The total length of both query and subject may not exceed 150,000 letters.

Query subrange  From  To

Or, upload file  Nenhum arquivo selecionado   
Job Title   
Enter a descriptive title for your BLAST search

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence  Clear  
Input limited to 100,000 letters for either input sequence. The total length of both query and subject may not exceed 150,000 letters.

Subject subrange  From  To

Or, upload file  Nenhum arquivo selecionado

**Align**  Show results in a new window

**Algorithm parameters**

BLAST is a registered trademark of the National Library of Medicine

Support center Mailing list

NCBI National Center for Biotechnology Information, U.S. National Library of Medicine 8600 Rockville Pike, Bethesda MD, 20894 USA

Policies and Guidelines | Contact

NATIONAL LIBRARY OF MEDICINE USA.gov

>2K7X\_1 | Chain A | SARS-CoV main protease | SARS coronavirus (255730)  
DRQTAQAAGTDTTITLNVLAWL YAAVINGDRW FLNRFTTLNDFNLVAMKYNYEPLTQDHVDIL GPLSAQTGIAVLDMCAALKELLQNGMNGRTIL GSTILEDEFTPFDVVVRQCSGVTFQ

>1RQ9\_1 | Chains A, B | protease | Human immunodeficiency virus 1 (11676)  
PQITLWQRPIVTIKIGGQLKEALLNTGADDTV LEEVNLPGRWKPKLIGGIGGFVKVRQYDQVPI EICGHKVIGTVLVGPTPANVIGRNLMQTIGCT LNF

\*National Center for Biotechnology Information

### Alignment statistics for match

NW Score -38

Identities 17/120(14%)

Positives 38/120(31%)

Gaps 21/120(17%)

Query	1	DRQTAQAAGDTTITLNVLAWLYAAVINGDRWFLNRF <del>TTL</del> NDFNLVAMKYNYEPLTQDH	60
		+ T +T+ + L A++N T L + NL ++P	
Sbjct	1	PQITLWQR---PIVTIKIGGQLKEALLN-----TGADDTVLEEVNLPG---RWKP-----	44
Query	61	VDILGPLSAQTGIAVLDMCAALKELLQNGMNGRTILGSTILEDEF <del>TPFDVVRQCSGVTFQ</del>	120
		++G + + D E+ + + G ++G T +++ Q	
Sbjct	45	-KLIGGIGGFVKVRQYDQVPI--EICGHKVIGTVLVGPT--PANVIGRNLMTQIGCTLNF	99



## DESAFIO #3

Tente realizar alinhamentos globais usando a ferramenta do NCBI\*

NIH U.S. National Center for Biotechnology Information Sign in to NCBI

BLAST®

Nucleotide Query Sequence

Enter accession number, gi, or FASTA sequence  Clear  
Input limited to 100,000 letters for either input sequence. The total length of both query and subject may not exceed 150,000 letters.

Query subrange  From  To

New columns added to the Description Table  
Click 'Select Columns' or 'Manage Columns'.

Or, upload file  Nenhum arquivo selecionado  
Job Title   
Enter a descriptive title for your BLAST search

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence  Clear  
Input limited to 100,000 letters for either input sequence. The total length of both query and subject may not exceed 150,000 letters.

Subject subrange  From  To

Or, upload file  Nenhum arquivo selecionado

Show results in a new window

BLAST is a registered trademark of the National Library of Medicine

Support center Mailing list YouTube

NCBI National Center for Biotechnology Information, U.S. National Library of Medicine 8600 Rockville Pike, Bethesda MD, 20894 USA

Policies and Guidelines | Contact

NATIONAL LIBRARY OF MEDICINE NIH USA.gov

>2K7X\_1 | Chain A | SARS-CoV main protease | SARS coronavirus (255730)  
DRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNRFTT  
TLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLD  
CAALKELLQNGMNGRTILGSTILEDEFTPFDVVRQCSGV  
TFQ

>6XCH\_1 | Chain A | 3C-like proteinase | Severe acute respiratory syndrome coronavirus 2 (2697049)  
SGFRKMAFPSGKVEGCMVQVTCGTTLNGLWLDDVVYCP  
RHVICTSEDMLNPNEYEDLLIRKSNNHNFLVQAGNVQLRVI  
GHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLAC  
YNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDC  
VSFCYMHMELPTGVHAGTDLEGNFYGPFDVRQTAQAAG  
TDTTITVNVLAWLYAAVINGDRWFLNRFTTLNDFNLVA  
MKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLO  
NGMNGRTILGSALLEDEFTPFDVVRQCSGVTFQ

\*National Center for Biotechnology Information

NW Score 413

Identities 116/306(38%)  
Positives 119/306(38%)  
Gaps 0/306(0%)

Sbjct 1 SGFRKMAFPSGKVEGCMVQVTCGTTLNGLWLDDVVYCPRHVICTSEMLNPNYEDLLIR 60

Sbjct 61 KSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLACYNG 120

Sbjct 121 SPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHMELPTGVHAGTDLEGN 180

Query 1 DRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNRFTTLNDFNLVAMKYNYE 54  
DRQTAQAAGTDTTIT+NVLAWLYAAVINGDRWFLNRFTTLNDFNLVAMKYNYE

Sbjct 181 FYGPFVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTLNDFNLVAMKYNYE 240

Query 55 PLTQDHVDILGPLSAQTGIAVLDMCAALKELLQNGMNGRTILGSTILEDEFTPFDVVRQC 114  
PLTQDHVDILGPLSAQTGIAVLDMCA+LKELLQNGMNGRTILGS +LEDEFTPFDVVRQC

Sbjct 241 PLTQDHVDILGPLSAQTGIAVLDMCASLKELLQNGMNGRTILGSALLEDEFTPFDVVRQC 300

Query 115 SGVTFQ 120

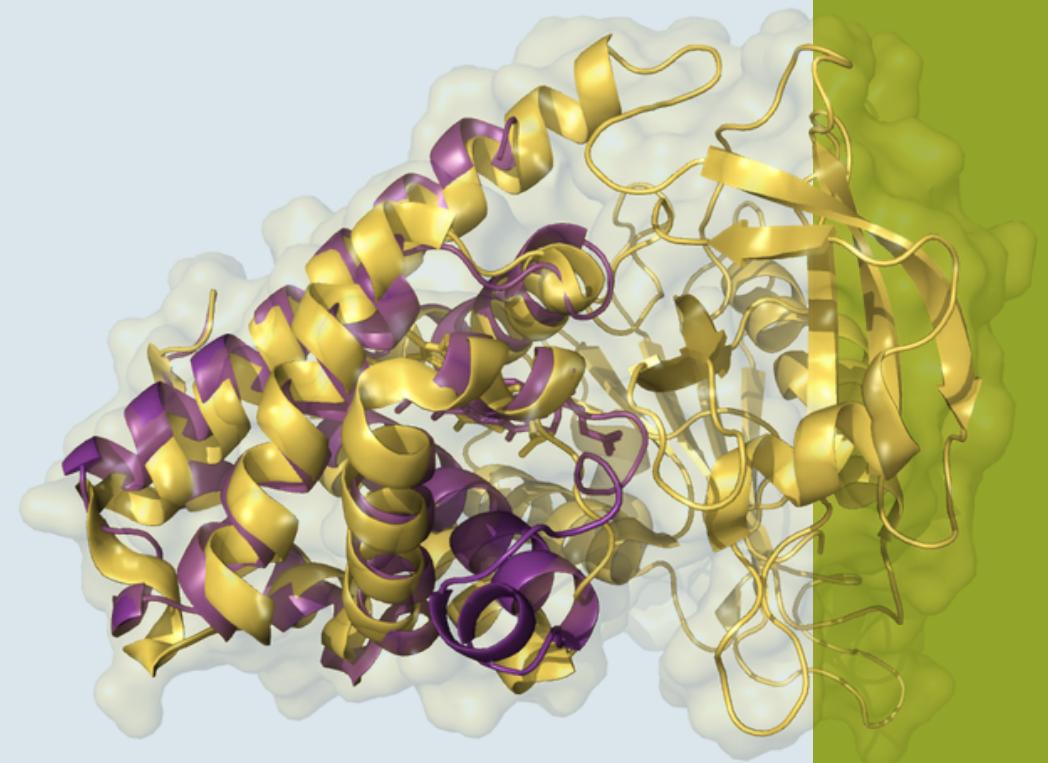
SGVTFQ

Sbjct 301 SGVTFQ 306

# Algoritmo de Needleman-Wunsch

Algoritmo calcula o alinhamento **global par-a-par** ótimo

- Útil quando a similaridade entre sequências se estender por toda sua extensão
  - **Proteínas de uma mesma** família que, normalmente, são conservadas, tem **comprimentos próximos** mesmo em organismos tão diversos quanto moscas e seres humanos
    - Exemplo: flavohemoglobina

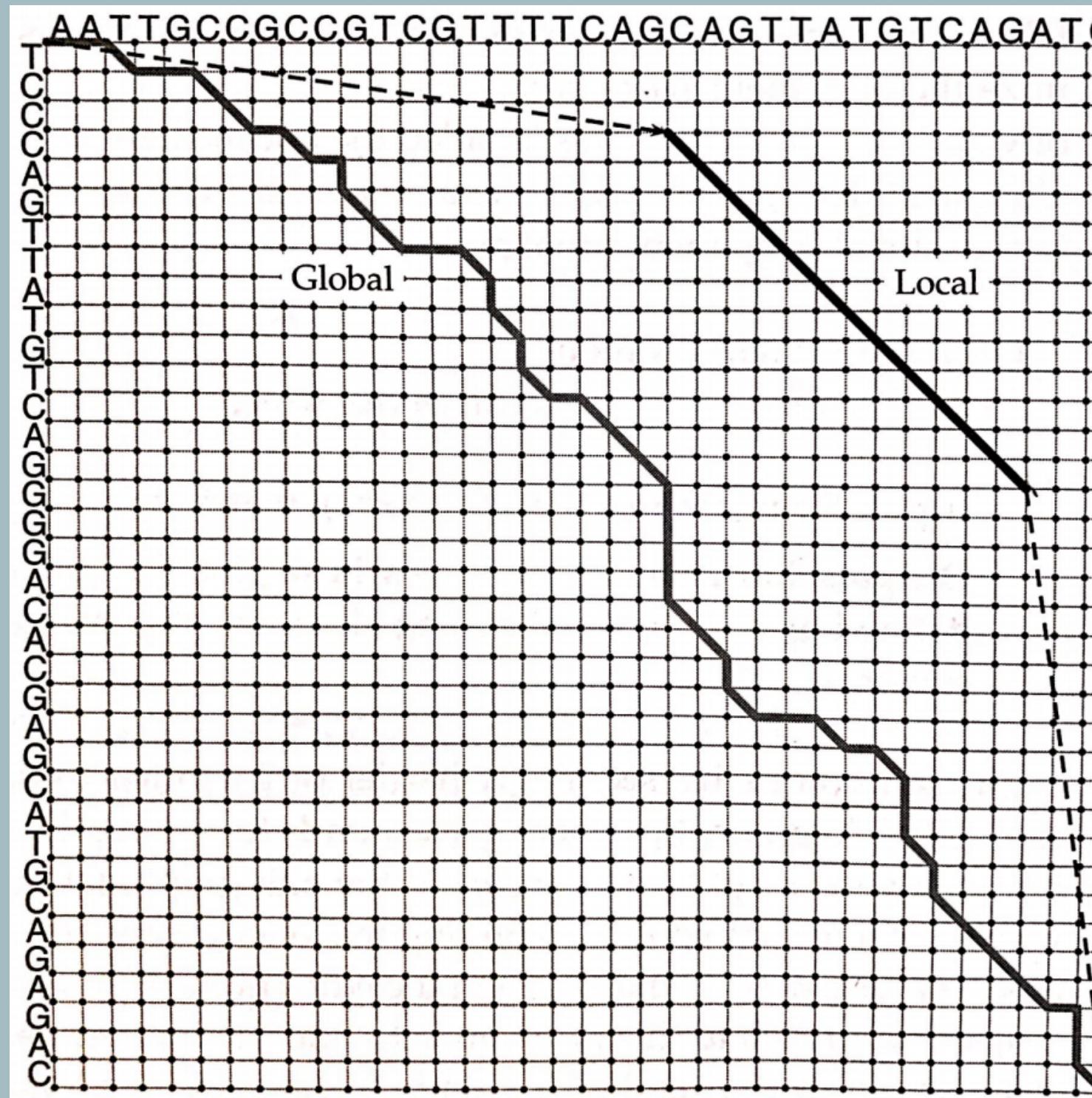


# Algoritmo de Smith-Waterman

Alinhamentos entre subsequências de v e w podem ter uma pontuação bem maior que a pontuação de v e w quando alinhadas globalmente

- Exemplo: proteínas que tem mais de um **domínio** altamente conservados mas a proteína por inteiro não é conservada
  - Hemoglobina e flavohemoglobina

# Como podemos encontrar regiões conservadas e ignorar as áreas de maior dissimilaridade?



Em 1981, Temple Smith e Michael Waterman propuseram uma elegante modificação no algoritmo de Needleman-Wunsch que resolve o alinhamento local e que ficou conhecido como o algoritmo de Smith-Waterman.



Smith, Temple F., and Michael S. Waterman. "Identification of common molecular subsequences." *Journal of molecular biology* 147.1 (1981): 195-197.

# Esquema de pontuação

$$S_{i,j} = \max \left\{ \begin{array}{l} \bullet S_{i-1,j} \\ \bullet S_{i,j-1} \\ \bullet S_{i-1,j-1} + 1, \text{ se } v[i] = w[j] \end{array} \right\}$$

# Novo esquema de pontuação

$$S_{i,j} = \max \left\{ \begin{array}{l} \bullet S_{i-1,j} - \sigma \text{ penalidade de } indel \\ \bullet S_{i,j-1} - \sigma \\ \bullet S_{i-1,j-1} + M, \text{ se } v[i] = w[j] \\ \bullet S_{i-1,j-1} - \mu, \text{ se } v[i] \neq w[j] \end{array} \right.$$

penalidade do *mismacth*

# Matrizes de substituição

$$S_{i,j} = \max \left\{ \begin{array}{l} \bullet S_{i-1,j} + \delta(v_i, -) \\ \bullet S_{i,j-1} + \delta(-, w_j) \\ \bullet S_{i-1,j-1} + \delta(v_i, w_j) \end{array} \right\}$$

onde  $\delta$  é uma matriz de substituição  
como PAM ou BLOSUM



# O que acontece com a pontuação quando alinhamos sequências muito diferentes?

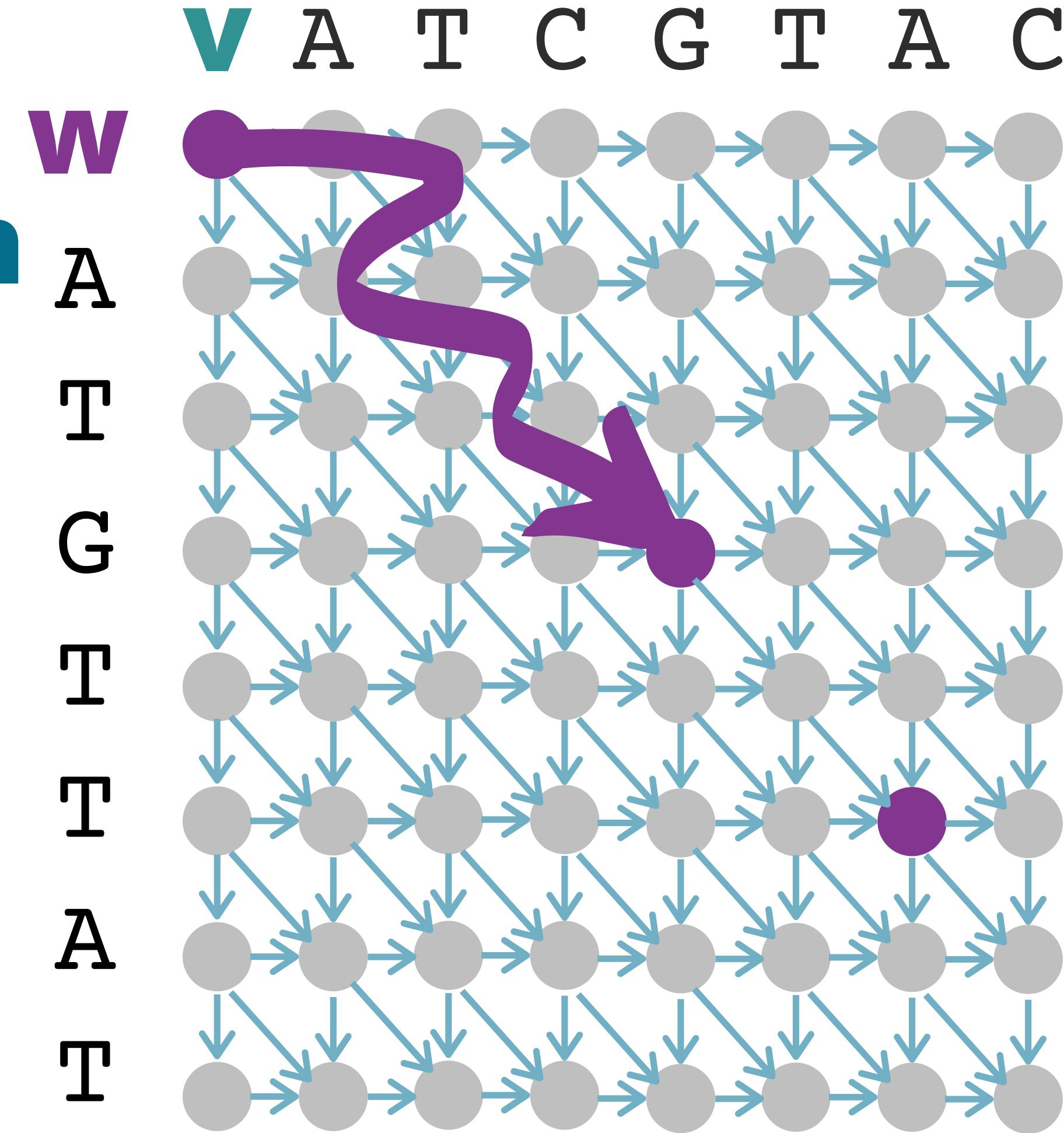
**Fica negativa**

Isso acontece também quando alinhamos sequências que tem um trecho em comum mas não similares como um todo

- Proteínas multidomínios por exemplo

Nesses casos, devemos preferir um algoritmo de **alinhamento local**

# Smith-Waterman



Adicionar uma nova seta entre o nó  $S_{0,0}$  e todos os outros de forma que, quando a pontuação ficar negativa, podemos eliminar o alinhamento até o momento

O melhor alinhamento não termina no nó  $S_{n,m}$  mas no nó de maior pontuação

# Algoritmo de Smith-Waterman

$$S_{i,j} = \max \left\{ \begin{array}{l} \bullet 0 \\ \bullet S_{i-1,j} + \delta(v_i, -) \\ \bullet S_{i,j-1} + \delta(-, w_j) \\ \bullet S_{i-1,j-1} + \delta(v_i, w_j) \end{array} \right.$$

Sempre que um alinhamento se torna muito ruim (pontuação negativa), pode-se **recomeçá-lo zerando a pontuação e ignorando a subsequência inicial**

# Diferenças do algoritmo de Smith-Waterman

