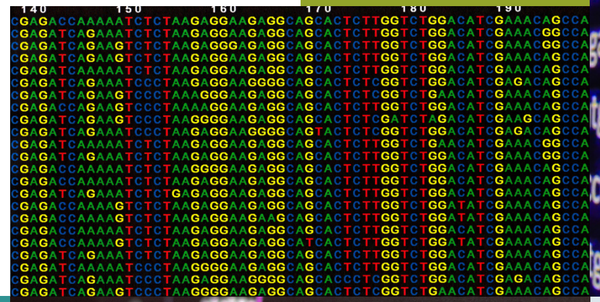
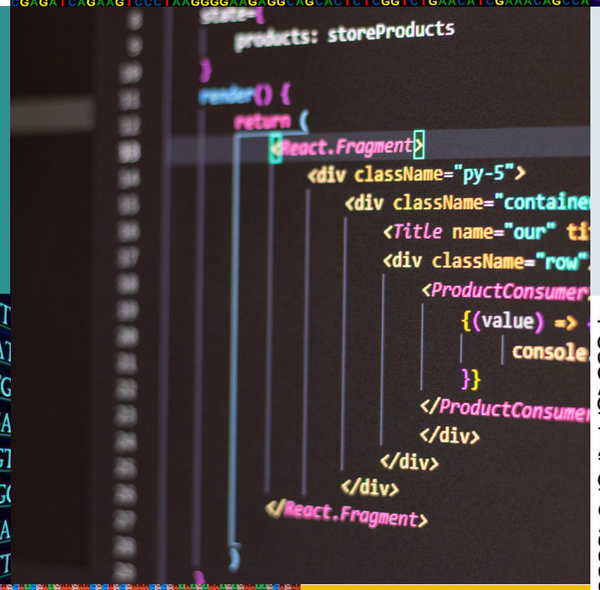
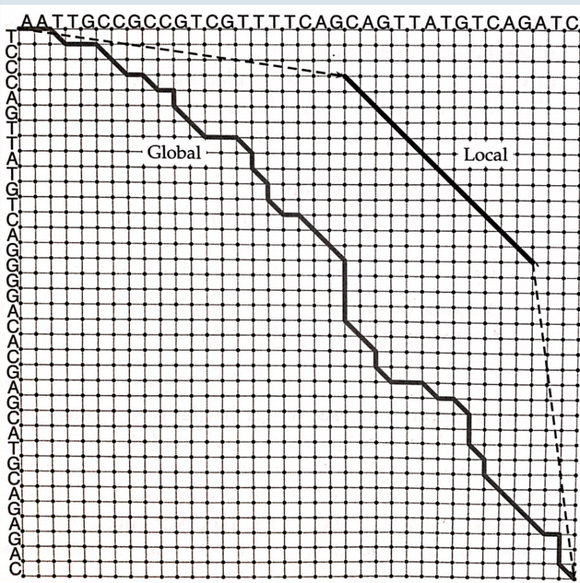
Algoritmo de Smith-Waterman







DESAFIO #1

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

```
>sequencia1
DRQTAKAAGTD

>sequencia2
ERQLAKAAGTD
```

9 pontos

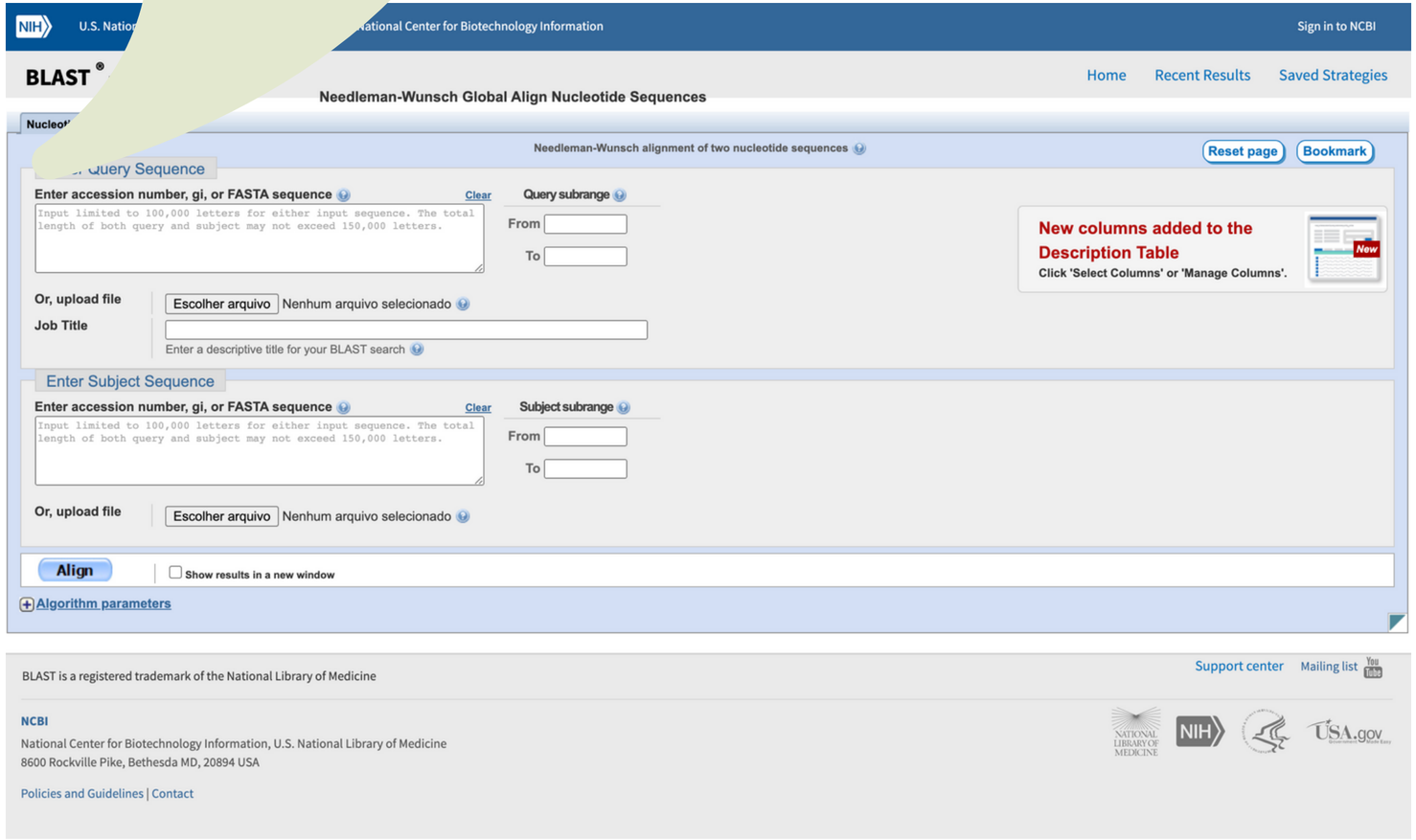
sequência 1 -DRQ-TAKAAGTD

sequência 2 E-RQL-AKAAGTD

		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_	0_
R	0|	0_	1\	1_	1_	1_	1_	1_	1_	1_	1_	1_
Q	0|	0_	1|	2\	2_	2_	2_	2_	2_	2_	2_	2_
L	0|	0_	1|	2|	2_	2_	2_	2_	2_	2_	2_	2_
A	0|	0_	1|	2|	2_	3\	3_	3\	3\	3_	3_	3_
K	0|	0_	1|	2|	2_	3|	4\	4_	4_	4_	4_	4_
A	0|	0_	1|	2|	2_	3\	4|	5\	5\	5_	5_	5_
A	0|	0_	1|	2|	2_	3\	4|	5\	6\	6_	6_	6_
G	0|	0_	1|	2|	2_	3|	4|	5|	6|	7\	7_	7_
T	0|	0_	1|	2|	3\	3_	4|	5|	6|	7|	8\	8_
D	0|	1\	1_	2|	3|	3_	4|	5|	6|	7|	8|	9\

DESAFIO #2

Tente realizar alinhamentos globais usando a ferramenta do NCBI*



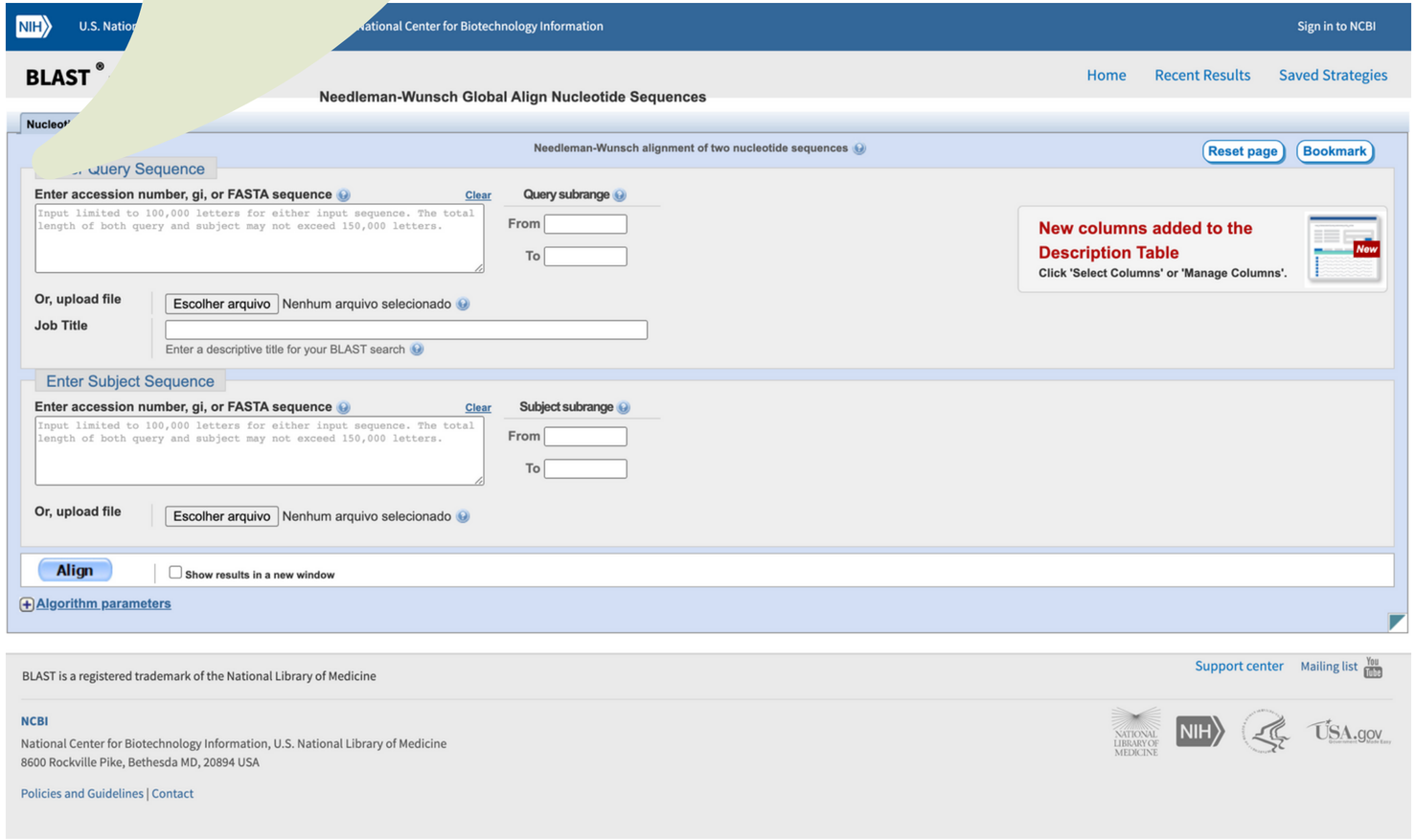

```
>2K7X_1|Chain A|SARS-CoV main
protease|SARS coronavirus
(255730)
DRQTAQAAGTDTTITLNVLAWLYAAVINGDRW
FLNRFTTTLNDFNLVAMKYNYEPLTQDHVDIL
GPLSAQTGIAVLDMCAALKELLQNGMNGRTIL
GSTILEDEFTPFDVVRQCSGVTFQ

>1RQ9_1|Chains
A,B|protease|Human
immunodeficiency virus 1 (11676)
PQITLWQRPIVTIKIGGQLKEALLNTGADDTV
LEEVNLPGRWKPKLIGGIGGFVKVRQYDQVPI
EICGHKVIGTVLVGPTPANVIGRNLMTQIGCT
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    DRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDH  60
             + T          +T+ +   L  A++N           T L + NL       ++P
Sbjct  1    PQITLWQR---PIVTIKIGGQLKEALLN-----TGADDTVLEEVNLPG---RWKP-----  44

Query  61   VDILGPLSAQTGIAVLDMCAALKELLQNGMNGRTILGSTILEDEFTPFDVVRQCSGVTFQ  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*




```
>2K7X_1|Chain A|SARS-CoV main
protease|SARS coronavirus (255730)
DRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNRFTT
TLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDM
CAALKELLQNGMNGRTILGSTILEDEFTPFDVVRQCSGV
TFQ

>6XCH_1|Chain A|3C-like
proteinase|Severe acute respiratory
syndrome coronavirus 2 (2697049)
SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCP
RHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVI
GHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLAC
YNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDC
VSFCYMHHMELPTGVHAGTDLEGNFYGPFVDRQTAQAAG
TDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVA
MKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQ
NGMNGRTILGSALLEDEFTPFDVVRQCSGVTFQ
```

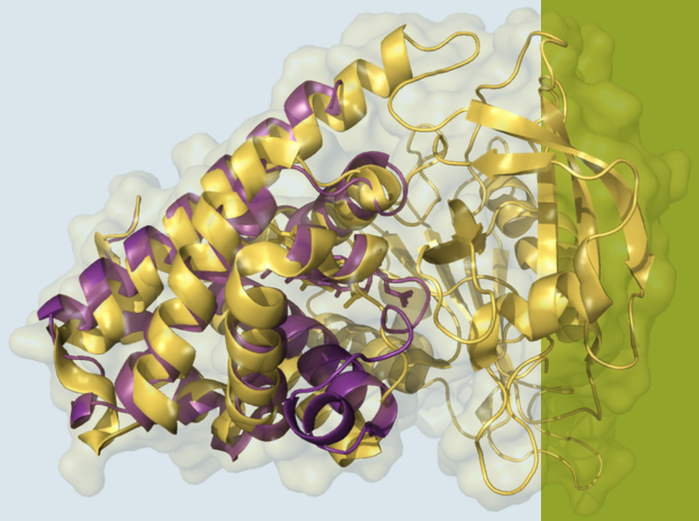
*National Center for Biotechnology Information

```
NW Score 413
Identities 116/306(38%)
Positives 119/306(38%)
Gaps 0/306(0%)

Sbjct  1    SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIR  60


Sbjct  61   KSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLACYNG  120


Sbjct  121  SPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGN  180


Query  1          DRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYE  54
                  DRQTAQAAGTDTTIT+NVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYE
Sbjct  181  FYGPFVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYE  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

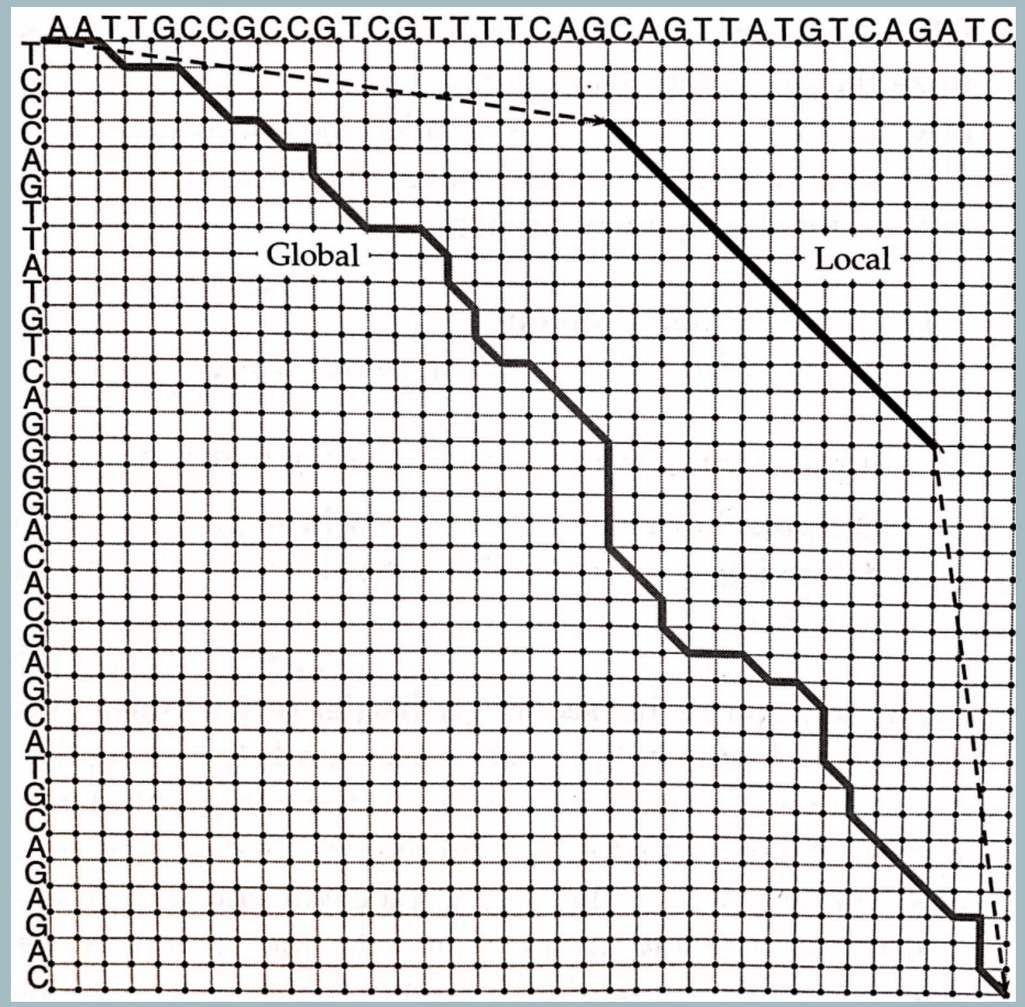


Algoritmo de Smith-Waterman

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 biology147.1 (1981): 195-197.

Esquema de pontuação

$$S_{i,j} = \max \begin{cases} S_{i-1,j} \\ S_{i,j-1} \\ S_{i-1,j-1} \textbf{ +1}, \text{ se } v[i]=w[j] \end{cases}$$

Novo esquema de pontuação

$$S_{i,j} = \max \begin{cases} S_{i-1,j} - \sigma & \text{penalidade de } \textit{indel} \\ S_{i,j-1} - \sigma & \\ S_{i-1,j-1} + M, \text{ se } v[i]=w[j] & \\ S_{i-1,j-1} - \mu, \text{ se } v[i]\neq w[j] & \text{penalidade do } \textit{mismacth} \end{cases}$$

Matrizes de substituição

$$S_{i,j} = \max \begin{cases} S_{i-1,j} + \delta(v_i, -) \\ S_{i,j-1} + \delta(-, w_j) \\ S_{i-1,j-1} + \delta(v_i, w_j) \end{cases}$$

onde δ é uma matriz de substituição como PAM ou BLOSUM

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

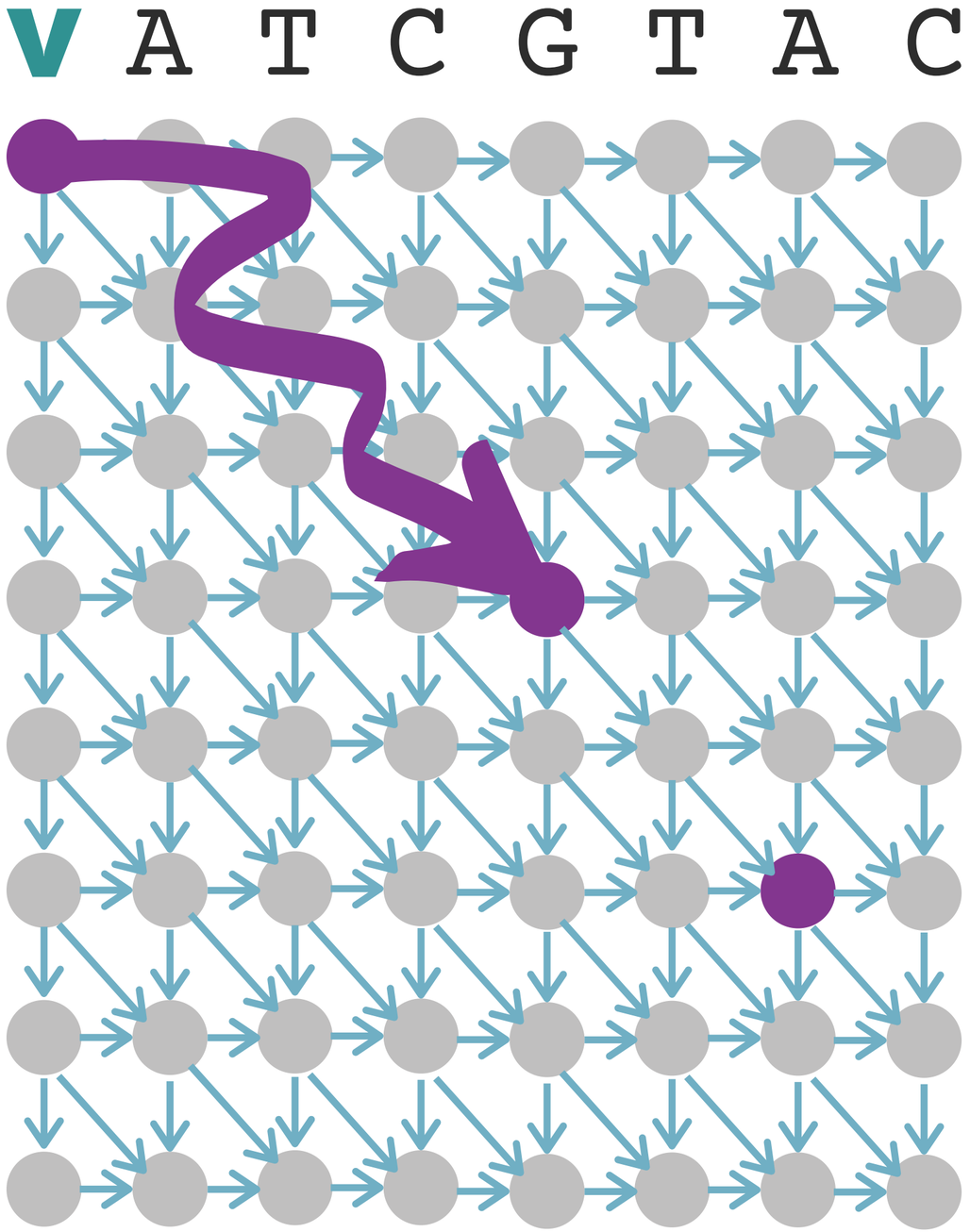
Fica negativa

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 *S0,0* 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 *Sn,m* mas no nó de maior pontuação

Algoritmo de Smith-Waterman

$$S_{i,j} = \max \begin{cases} \mathbf{0} \\ S_{i-1,j} + \delta(v_i, -) \\ S_{i,j-1} + \delta(-, w_j) \\ S_{i-1,j-1} + \delta(v_i, w_j) \end{cases}$$

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