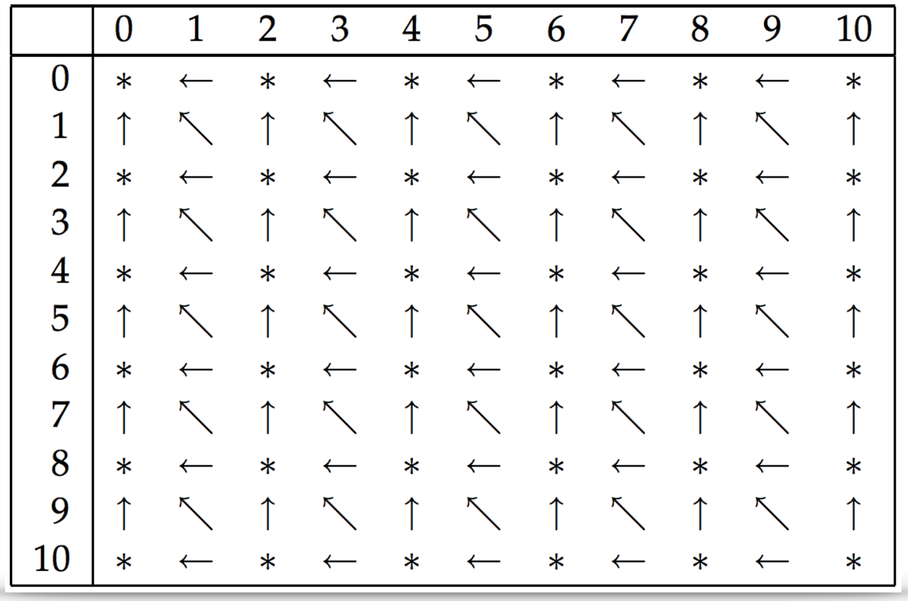
Profa. Dra. Raquel C. de Melo-Minardi
Departamento de Ciência da Computação
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais



	0	1	2	3	4	5	6	7	8	9	10
0	*	←	*	←	*	←	*	←	*	←	*
1	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
2	*	←	*	←	*	←	*	←	*	←	*
3	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
4	*	←	*	←	*	←	*	←	*	←	*
5	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
6	*	←	*	←	*	←	*	←	*	←	*
7	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
8	*	←	*	←	*	←	*	←	*	←	*
9	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
10	*	←	*	←	*	←	*	←	*	←	*

MÓDULO 4
ALGORITMOS PARA BIOINFORMÁTICA
Algoritmo de Smith-Waterman

ALGORITMO DE NEEDLEMAN-WUNSCH

- O algoritmo de Needleman-Wunsch que acabamos de apresentar é um algoritmo de alinhamento global par-a-par que busca similaridades entre duas sequências globalmente
  - Útil quando a similaridade entre sequências se estender por toda sua extensão
    - Exemplo: proteínas de uma mesma família que, normalmente, são conservadas, tem comprimentos próximos mesmo em organismos tão diversos quanto moscas e seres humanos

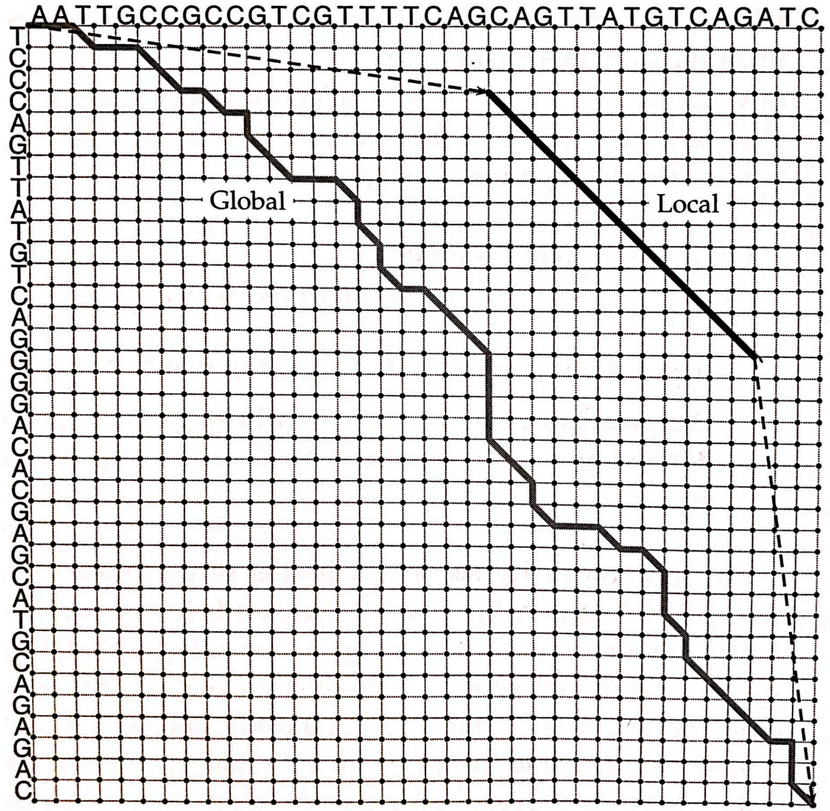
ALGORITMO DE SMITH-WATERMAN

- Entretanto, em diversas aplicações biológicas, isso não ocorre e 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
- Como podemos encontrar essas 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 e Waterman, 1981]

[Smith e Waterman, 1981] Smith, Temple F., and Michael S. Waterman. "Identification of common molecular subsequences." Journal of molecular biology147.1 (1981): 195-197.

ALGORITMO DE SMITH-WATERMAN

- A figura a seguir ilustra a **diferença conceitual** entre um **alinhamento global** e **local** e ainda nos dá uma primeira dica de como funciona o algoritmo de Smith-Waterman



Problema do Alinhamento Local de Sequências

Encontre o melhor alinhamento local entre duas sequências

Entradas: Duas sequências, v e w, e uma matriz de pontuação δ

Saída: Subsequências de v e w para os quais o **alinhamento global**, conforme a matriz de pontuação utilizada δ, é máximo entre todos os alinhamentos globais de subsequências de v e w

MATRIZES DE PONTUAÇÃO

- Matrizes para pontuar a similaridade entre sequências de DNA usualmente são definidas pelos parâmetros:
  - *Match* (M)
  - *Mismatch* (μ)
  - *Indel* (σ)
- No exemplo bastante simplificado que usamos na abordagem de alinhamento global o valor de M foi “+1”, μ e σ foram de “0”

MATRIZES DE PONTUAÇÃO

- Mutações aleatórias em sequências de nucleotídeos podem provocar mudanças na sequência de aminoácidos
  - Algumas dessas mutações podem não afetar a estrutura e a função de proteínas mas outras podem ser muito relevantes afetando a habilidade de sobrevivência do organismo
  - Há aminoácidos que são mais comumente mutados (ASN, ASP, GLU, SER) e outros raramente mutados (CYS e TRP)
- A probabilidade de se encontrar uma SER mutada por uma PHE é três vezes maior que de se encontrar um TRP mutado por uma PHE
- Esse tipo de conhecimento estatístico sobre as mutações que ocorrem nas proteínas dos seres vivos permitem elaborar matrizes de pontuação para alinhar adequadamente sequências de proteínas

MATRIZES DE PONTUAÇÃO

- Para alinhamento de sequências de **proteínas**, compostas por um alfabeto de 20 possíveis aminoácidos, usamos matrizes de pontuação δ
- Uma matriz de pontuação $\delta(i, j)$ traz a frequência na qual encontramos um aminoácido i substituído por um j
- As matrizes mais comumente utilizadas são
  - **PAM**: *Point Accepted Mutations*, desenvolvida por Margareth Dayhoff [Dayhoff et al., 1978]
  - **BLOSUM**: *Block Substitution*, desenvolvida por Steven e Joria Henikoff [Henikoff e Henikoff, 1992]

[Dayhoff et al., 1978] Dayhoff, M. O., R. M. Schwartz e B. C. Orcutt. "22 A Model of Evolutionary Change in Proteins." Atlas of protein sequence and structure. Vol. 5. National Biomedical Research Foundation Silver Spring, MD, 1978. 345-352.

[Henikoff e Henikoff, 1992] Henikoff, Steven, and Jorja G. Henikoff. "Amino acid substitution matrices from protein blocks." Proceedings of the National Academy of Sciences 89.22 (1992): 10915-10919.

ALGORITMO DE SMITH-WATERMAN

- Voltando ao Problema do Alinhamento Local de Sequências, Smith e Waterman, usando uma **matriz de substituição δ** para pontuar dissimilaridades entre aminoácidos substituídos em duas sequências de proteínas *v* e *w*, perceberam que, como uma **pequena e simples modificação** o algoritmo de Needleman-Wunsch poderia ser usado para buscar
  - **o melhor alinhamento global entre duas subsequências de *v* e *w***, ou em outras palavras,
  - **o máximo alinhamento local entre duas sequências**

ALGORITMO DE SMITH-WATERMAN

- Lembre que para resolver o LCS (alinhamento global), criamos uma matriz de programação dinâmica e a preenchemos seguindo o seguinte critério:
  - $s_{i,j} = max\ (s_{i-1,j},\ s_{i,j-1}\ e\ s_{i-1,j-1}+1)$ (desde que $v[i] = w[j]$)
- Há variações desse critério que pontuam **diferentemente** para *matches*, *mismatches* e *indels* como, por exemplo:
  - $s_{i,j} = max\ (s_{i-1,j}-\sigma,\ s_{i,j-1}-\sigma,\ s_{i-1,j-1}-\mu)$ (desde que $v[i]$ seja diferente $w[j]$)) e $s_{i-1,j-1}+M$ (desde que $v[i]$ igual a $w[j]$)):
  - onde
    - M: é a pontuação de um *match*
    - μ: é a penalidade de um *mismatch*
    - σ: é a penalidade de um *indel*

ALGORITMO DE SMITH-WATERMAN

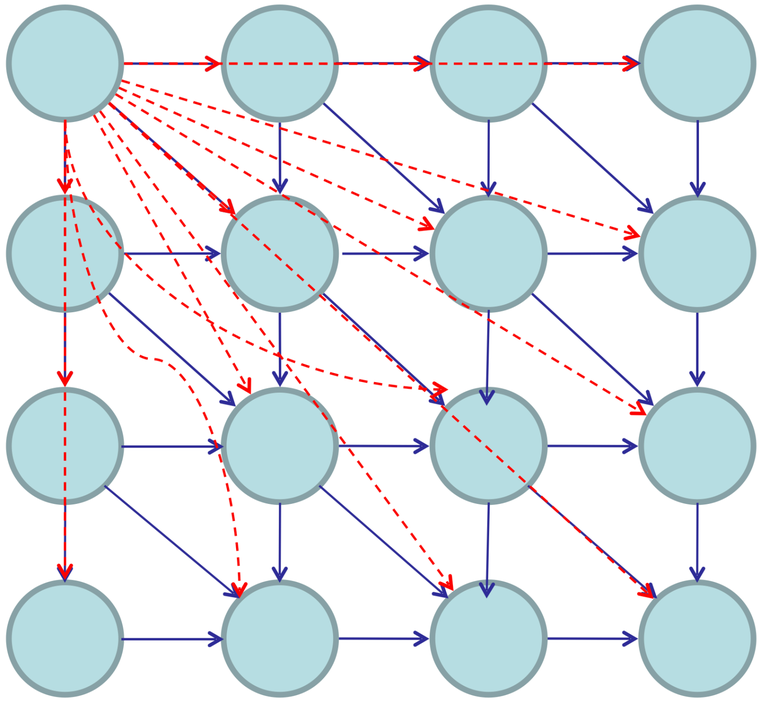
- Veja a seguir uma outra variação usando uma matriz de pontuação δ:
  - $s_{i,j} = max\ (\ s_{i-1,j}+\delta(v_i, -),\ s_{i,j-1}+\delta(-, w_j)$ e $s_{i-1,j-1}+\delta(v_i, w_j)\ )$
  - onde
    - $\delta(v_i, -)$ e $\delta(v_i, -)$: são as penalidade de *indels* (deleção e inserção, respectivamente).
    - $\delta(v_i, w_j)$: pode ser a pontuação ou a penalidade dependendo se é um *match* ou *mismatch*

ALGORITMO DE SMITH-WATERMAN

- O **Problema do Alinhamento Local** foi resolvido simplesmente a substituição do critério anterior
  - $s_{i,j} = max\ (\ s_{i-1,j}+\delta(v_i, -),\ s_{i,j-1}+\delta(-, w_j)$ e $s_{i-1,j-1}+\delta(v_i, w_j)\ )$
- pelo seguinte:
  - $s_{i,j} = \max\ (\mathbf{0},\ s_{i-1,j}+\delta(v_i,-),\ s_{i,j-1}+\delta(-, w_j)$ e $s_{i-1,j-1}+\delta(v_i, w_j)\ )$
  - Note que ele é idêntico ao critério anterior usado no alinhamento global exceto por adicionar “0" pontos no caso de a pontuação ter se tornado negativa

ALGORITMO DE SMITH-WATERMAN

- É como se adicionássemos **uma aresta de peso “0"** no *grid* ou, em outras palavras, conectássemos o nó fonte $s_{0,0}$ a todos os outros no $s_{i,j}$ do *grid*



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

ALGORITMO DE SMITH-WATERMAN

- No **alinhamento global**, a pontuação do melhor alinhamento sempre está no nó (ou célula) $s_{m,n}$
- No **alinhamento local**, buscamos a célula $s_{i,j}$ de **pontuação máxima** (que indica onde o alinhamento irá terminar nas sequências *v* e *w*)
- Dessa forma, caracteres após (ou à direita) de *i* e *j* não farão parte do alinhamento local máximo de *v* e *w*