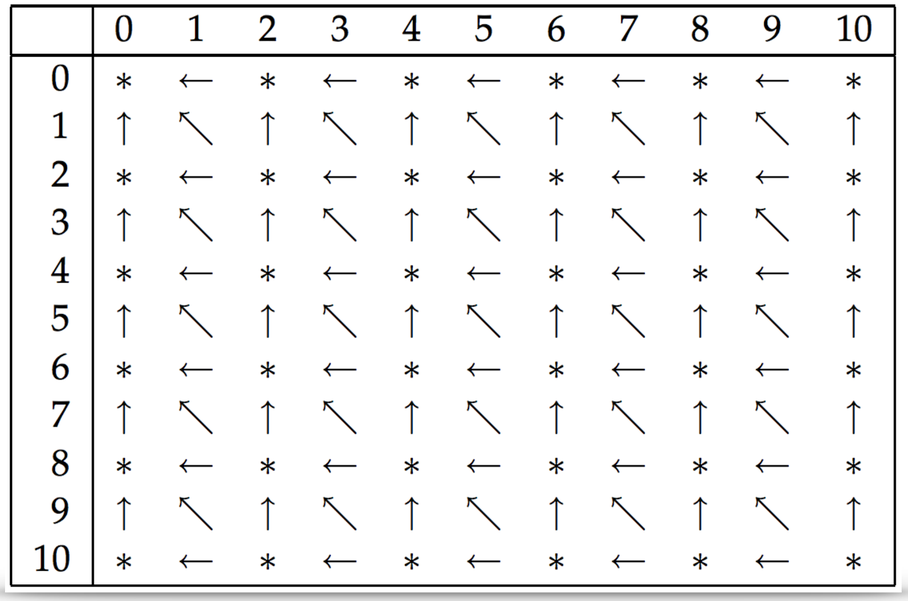

	0	1	2	3	4	5	6	7	8	9	10
0	*	←	*	←	*	←	*	←	*	←	*
1	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
2	*	←	*	←	*	←	*	←	*	←	*
3	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
4	*	←	*	←	*	←	*	←	*	←	*
5	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
6	*	←	*	←	*	←	*	←	*	←	*
7	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
8	*	←	*	←	*	←	*	←	*	←	*
9	↑	↖	↑	↖	↑	↖	↑	↖	↑	↖	↑
10	*	←	*	←	*	←	*	←	*	←	*

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

MÓDULO 4
ALGORITMOS PARA BIOINFORMÁTICA
Distâncias entre sequências

ALINHAMENTO DE SEQUÊNCIAS E DISTÂNCIA DE HAMMING

- É um problema genuinamente de bioinformática?
  - Não, já é um problema bem resolvido e mais antigo em ciência da computação
    - Uma das formas já utilizada para calcular a distância ou dissimilaridade de sequências era a chamada distância de ***Hamming***
    - Criada em 1950 para detecção e correção de erros em sequências

Na teoria da informação, a **distância de Hamming** entre duas cadeias de caracteres de mesmo comprimento é o número de posições nas quais elas diferem entre si. Vista de outra forma, ela corresponde ao menor número de substituições necessárias para transformar uma sequência na outra, ou o número de erros que transformaram uma na outra.

Exemplo:

As sequências

AC**T**G**A**CTG
TC**A**G**C**CTG

tem a distância de Hamming 3

Wikipedia

DISTÂNCIA DE EDIÇÃO OU DISTÂNCIA DE LEVENSHTEIN

- As sequências ATATATAT e TATATATA são muito diferentes em termos da distância de Hamming (8)
  - Poderiam ser consideradas extremamente similares se inseríssemos uma lacuna, arredando uma das sequências uma posição para a frente
  - É o que os biólogos chamam de ***gaps*** em um alinhamento de sequências
- Em 1966, Vladimir **Levenshtein** introduziu a noção da **distância de edição** entre duas sequências

A **distância de edição** ou **distância de Levenshtein** entre duas sequências é o número mínimo de edições dos seguintes tipos que são necessárias para transformar uma sequência na outra:

- **Inserção** de um caracter em uma sequência
- **Deleção** de um caracter em uma sequência
- **Substituição** de um caracter por outro em uma sequência

Exemplo:

As sequências

```
ATAGATAT-
-TACATATA
```

tem a distância de Levenshtein 3:

- A deleção de um “A” na 1a. posição
- A substituição de um “G" por um “C" na 4a. posição
- A inserção de um “A" na 9a. posição

Jones e Pevzner

- Ao contrário da distância de Hamming, a distância de **Levenshtein** permite **comparar sequências de tamanhos diferentes**

- Levenshtein propôs essa definição de distância sem nunca ter apresentado um algoritmo para calculá-la

Desafio

Implemente em Python duas funções que recebam como argumento duas sequências e retornem:

1. A distância de Hamming
2. A distância de Levenshtein