

# Algorithmique avancée

## Recherche exhaustive

Denis TRYSTRAM  
TP ENSIMAG Alternants 2A

sept. 2022

## Comparaison de séquences biologiques

Beaucoup d'applications nécessitent la comparaison de séquences biologiques (ADN ou protéines).

- ADN, 4 "bases" : **A**dénine, **G**uaranine, **C**ytosine et **T**hymine, une séquence d'ADN s'écrit comme une suite de ces quatre caractères.
- But : Déterminer le degré de similarité entre deux séquences.

## La plus longue sous-séquence commune (PLSSC)

On considère deux séquences

$X = \langle x_1, x_2, \dots, x_n \rangle$  et  $Y = \langle y_1, y_2, \dots, y_m \rangle$

### Définition

On dit que  $Z$  est une sous-séquence commune de  $X$  et  $Y$  s'il existe une séquence –strictement croissante– d'indices  $i_1, i_2, \dots, i_k$  tels que  $\forall j \ 1 \leq j \leq k$ , on a  $z_j = x_{i_j} = y_{i_j}$ .

## Exemple

- $X = \langle \text{TCGCGAATGATC} \rangle$
- $Y = \langle \text{AGTGACGATCTGA} \rangle$
- $Z = \langle \text{GCATA} \rangle$  est une séquence commune

On cherche à déterminer la plus longue sous-séquence commune par une méthode exacte.

# Brute force

## Objectif

énumérer toutes les sous-séquences de  $X$   
vérifier pour chacune si elles sont des sous-séquences de  $Y$  en gardant la trace de la plus longue sous-séquence ainsi trouvée.

- Ecrire cet algorithme dans le langage de votre choix.
- Déterminer l'ordre de grandeur asymptotique de cette méthode.
- Faire des tests en fonction de  $n$  et  $m$ .

## Complexité

- Chaque sous séquence de  $X$  correspond à un sous-ensemble de l'ensemble  $\{1, 2, \dots, n\}$
- Donc,  $\mathcal{O}(2^n)$