Projet de programmation PI4 2021/2022 : **Alignement de séquences**

Proposé par Alice Rogier

alice.rogier@inserm.fr

1 Contexte

L'ADN est une macromolécule constituée de quatre nucléotides (Adénine, Cytosine, Thymine et Guanine). Certains fragments d'ADN, appelés gènes, sont transcrits en ARN puis traduits en protéines, molécules qui assurent toutes les fonctions nécessaires au bon fonctionnement de l'organisme.
L'ordre des nucléotides d'un gène détermine entièrement la fonction de la protéine. Au cours de l'évolution, les gènes accumulent des mutations. Si deux séquences de gènes de différentes espèces sont similaires, elles dérivent certainement d'un ancêtre commun. On dit que ces séquences sont homologues.
L'alignement de séquences génomiques consiste à comparer deux séquences d'ADN pour évaluer un score optimal de similarité entre ces deux séquences. Cette technique est couramment utilisée par les bioinformaticiens pour identifier, comparer et/ou prédire la fonction de gènes. Elle est également utile en phylogénie pour comprendre leur évolution.

2 Objectif

Votre objectif est de développer une interface graphique permettant d'aligner deux courtes séquences d'ADN (pas plus de 20 nucléotides) selon l'algorithme d'alignement global de Needleman et Wunsch (https://fr.wikipedia.org/wiki/Algorithme_de_Needleman-Wunsch).
Vous pouvez vous appuyer sur l'application web "Global Alignment App" : https://bioboot.github.io/bimm143_W20/class-material/nw/

3 Principe de l'algorithme d'alignement

L'objectif de l'alignement de séquences est de rechercher le maximum d'appariements entre les nucléotides (=lettres) des deux séquences comparées. Pour ce faire, on va rechercher à maximiser un score d'alignement.
On s'appuiera sur l'alignement de la figure 1 pour expliquer le principe.
Pour aligner les séquences 1 ("GATTACA", longueur = 7) et 2 ("GTCGACGCA", longueur = 9) une matrice de taille 8*10 est remplie.
Vous remarquez que l'utilisateur définit des scores de "match", "miscmatch" et de "gap". Ces scores permettent de valoriser l'identité entre les éléments des deux séquences et de pénaliser la substitution et l'insertion de gap. L'identité signifie que les deux lettres comparées sont identiques (ex "G"" et "G""), la substitution signifie qu'elles sont distinctes (ex "T"" et "A""), le gap signifie

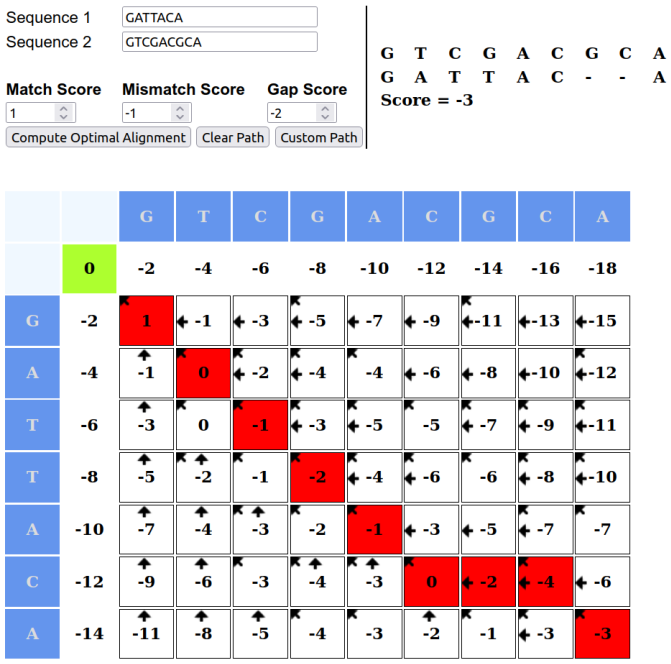

Figure 1: Capture d'écran de "Gobal Alignment App"

qu'un trou est inséré dans la séquence ("-").

3.1 Remplissage de la matrice : programmation dynamique

Soient $x[1..n]$ et $y[1..m]$ deux séquences à aligner (x horizontal, y vertical). On nomme S_{match} le score de match, $S_{mismatch}$ le score de mismatch et d la pénalité de gap. (dans l'exemple $S_{match} = 1$, $S_{mismatch} = -1$, et $d = -2$) . On nomme $s(a, b)$, la fonction de score telle que :

$$s(a,b) = \begin{cases} \mathrm{S}_{match} & \text{si a = b,} \\ \mathrm{S}_{mismatch} & \text{sinon} \end{cases} \tag{1}$$

Soit F la matrice d'alignement de $m+1$ lignes et $n+1$ colonnes. La première ligne et la première colonne de la matrice sont initialisées à l'aide de la pénalité de gap :
- $F[i][0] = d * i$
- $F[0][j] = d * j$

On a ensuite 3 façons d'étendre un alignement de x_i à y_i :
- x_i et y_i sont alignés : on a avancé les deux mots en alignant deux lettres
- x_i est aligné avec un gap : on a avancé dans x, mais pas dans y
- y_i est aligné avec un gap : on a avancé dans y, mais pas dans x

$$F[i][j] = MAX \begin{cases} F[i-1][j-1] + s(x_i, y_j), \\ F[i-1][j] + d, \\ F[i][j-1] + d \end{cases} \tag{2}$$

Dans l'exemple, on a :

$$F[1][1] = MAX \begin{cases} 0+1, \\ -2-1, \\ -2-1 \end{cases} \tag{3}$$

Donc $F[1][1] = 1$.

3.2 Backtracking

Une fois la matrice remplie, Le chemin pour obtenir le score final, c'est-à-dire le score correspondant à la case $F[m][n]$ est retrouvé. Dans l'exemple, ce score est de -3, et il correspond à l'alignement :

$$\begin{array}{c} G\ C\ G\ A\ C\ G\ C\ A \\ G\ T\ T\ A\ C\ -\ -\ A \end{array}$$

4 Fonctionnalités

Voici une liste de fonctionnalités à implémenter :

- (obligatoire) Possibilité d'aligner n'importe quelle paire de séquences de moins de 20 nucléotides
- (obligatoire) Possibilité de changer tous les scores
- (obligatoire) Boutons "Compute Optimal Alignment" et "Clear Path" permettant de mettre en évidence ou non le chemin optimal sur la matrice
- (obligatoire) Affichage de la matrice
- (obligatoire) Affichage de l'alignement
- (recommandé) Bouton "Custom Path"
- (recommandé) Fenêtre expliquant comment le score de la case a été trouvé
- (recommandé) Faire des icônes avec les nucléotides indexant la matrice que l'on peut sélectionner et interchanger

N'hésitez pas à faire preuve de créativité, à implémenter des choses n'apparaissant pas dans la liste.

5 Pour aller plus loin

Si votre groupe avance vite, je vous proposerai d'ajouter la possibilité d'aligner des séquences protéiques.