AAGB - TD1 - 21/10/18

1. Introduction

Pour chacun des problèmes de biologie computationnelle vus en cours, décrivez :

1. La problématique.
2. Les données en entrée.
3. D'où elles sont extraites (base de données ou expériences).
4. L'état de l'art et/ou les modèles.

2. Comparaison de séquences

1. Quel peut être l'intérêt de comparer des séquences ?
2. Qu'est ce que l'homologie ?
3. Comment comparer des séquences ?

2.1 Dot Plot (matrice de similarité)

On peut comparer facilement deux séquences en plaçant l'une en ordonnée et l'autre en abscisse, et en marquant les paires de lettres identiques. Pour mieux discriminer les régions conservées, on peut utiliser une fenêtre de taille n, afin de marquer uniquement les mots identiques de longueur $>$ n.

1. Etant données les deux séquences suivantes A = (GCACTAGACC) et B = (GCATCGAC), construire le dot plot correspondant, pour une fenêtre de longueur 3.
2. Quelle interprétation peut-on donner si l'on obtient une grande diagonale ? Une diagonale coupée en deux ?

2.2 Alignement de séquences : l'algorithme de Needleman & Wunsch

L'alignement de séquences est par définition leur mis en correspondance de manière à en faire ressortir les similitudes. On cherche à maximiser le nombre de coïncidences entre nucléotides/acides aminés (matches) et à minimiser les différences (mismatches et gaps).

Définition

Soient :

- $A = (a_1, a_2, ..a_n)$ et $B = (b_1, b_2, ..b_n)$,
- $S_{i,j}$ le score maximal correspondant à l'alignement entre $(a_1, a_2, ..a_i)$ et $(b_1, b_2, ..b_j)$,
- g le score attribué à un gap,
- $\sigma(a_i, b_j)$ le score de match ou de mismatch.

Algorithme

- Initialisation : $\begin{cases} S_{i,0} = i \times g \\ S_{0,j} = j \times g \end{cases}$
- Récurrence : $S_{i,j} = max \begin{cases} S_{i-1,j-1} + \sigma(a_i, b_j) \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{cases}$
- Traceback : à chaque étape, on garde en mémoire l'alignement qui donne le meilleur score. Cela donne le chemin qui donne le/les meilleurs alignements globaux.

On voudrait aligner les séquences A = (CATGAC) et B = (TCTGAAC), avec les scores suivants:

- $g = -1$ (gap)
- mismatch = -2
- match = 1

1. Remplir le tableau suivant l'algorithme de Needleman Wunsch.
2. Utiliser le traceback pour donner l'alignement correspondant.
3. Comment adapter cet algorithme afin d'obtenir un alignement local plutôt que global ?

		C	A	T	G	A	C
	0	-1	-2	-3	-4	-5	-6
T	-1						
C	-2						
T	-3						
G	-4						
A	-5						
A	-6						
C	-7						