AAGB - TME1 - 23/11/18

1. Needleman & Wunsch

L'objectif de cet exercice est de recode l'algorithme de Needleman & Wunsch. Les différentes questions servent de guide par la mise en place de l'algorithme, mais ne doivent pas nécessairement être suivis à la lettre. Pour rappel, l'algorithme est le suivant :

- 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.

1. Une matrice de distance est une matrice D de taille n, où n est la taille de l'alphabet. Etant donnée deux lettres de l'alphabet A et B, $D(A, B)$ correspond à la distance entre les lettres A et B. Dans notre cas, l'alphabet peut être l'ensemble de nucléotides (A,C,G,T) ou l'ensemble des 20 acides aminés. La diagonale de la matrice D correspond à des matches, le reste à des mistmaches. Par exemple, pour une valeur de match de 1 et de mismatch de -1, D sera une matrice avec des 1 sur la diagonale et des -1 ailleurs. Ecrire une fonction renvoyant la matrice de distance, étant donnés un alphabet, un score de match et un score de mismatch.
2. Coder une fonction qui retourne la matrice de score (sans traceback) suivant l'algorithme de Needleman & Wunsch. Cette fonction peut par exemple prendre en entrée les deux séquences à aligner, la matrice de distance et une valeur correspondant au coût de gap. Quelle est la complexité de cet algorithme ?
3. Rajouter le traceback pour obtenir la fonction complète. La sortie de la fonction finale doit retourner l'alignement ainsi que le score associé. Etant données deux séquences, leur alignement est-il nécessairement unique ?
4. En général, on considère que l'introduction d'un gap est moins couteuse que le prolongement d'un gap. Modifier l'algorithme afin de pouvoir prendre en compte des valeurs différent pour l'ouverture et le prolongement de gaps.
5. On peut par exemple tester la fonction sur les séquences données en TD : A = (CATGAC) et B = (TCTGAAC).

2. Smith-Waterman

L'algorithme de Needleman & Wunsch est un alignement global. Il peut être plus utile de mettre en évidence des alignement locaux. Pour ce faire, on peut par exemple réutiliser l'algorithme précédent, mais en ajoutant les modifications suivantes :

- Récurrence : $S_{i,j} = max \begin{cases} 0 \\ S_{i-1,j-1} + \sigma(a_i, b_j) \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{cases}$

- Pour le traceback, au lieu de commencer en bas à droite du tableau, on part plutôt de la position qui donne le score le plus élevé.

1. Coder l'algorithme de Smith-Waterman. De la même façon, cet algorithme doit retourner l'alignement local entre les deux séquences données. On peut aussi retourner la localisation de cet alignement dans chaque séquence. Attention, il peut y avoir plusieurs maxima locaux : on peut au choix retourner un ou tous les meilleurs alignements.

2. Comparer le résultat des deux algorithmes sur les séquences données en TD.

3. Petit exemple

On aimerait voir ce que donne nos algorithmes sur des séquences protéiques réelles. On va prendre par exemple les protéines 2ABL et 1OPK, dont les séquences sont les suivantes:

- >2ABL:A|PDBID|CHAIN|SEQUENCE

MGPSENDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQ
TKNGQGWVPSNYITPVNSLEKHSWYHGPVSRN AAEYLLSSGINGSF
LVRESESSPGQRSISLRYEGRVYHYRINTASDGKLYVSSESRFNTLAELV
HHHSTVADGLITTLHYPAP

- >1OPK:A|PDBID|CHAIN|SEQUENCE

GAMDPSEALQRPVASDFEPQGLSEAARWNSKENLLAGPSENDPNLFV
ALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGW
VPSNYITPVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESE
SSPGQRSISLRYEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHHST
VADGLITTLHYPAPKRNKPTIYGVSPNYDKWEMERTDITMKHKLGGG
QYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNL
VQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVSAVVLLYMATQIS
SAMEYLEKKNFIHRNLAARNCLVGENHLVKVADFGLSRLMTGDTYTAH
AGAKFPIKWTAPESLAYNKFSIKS DVWAFGVLLWEIATYGMSPYPGIDL
SQVYELLEKDYRMERPEGCPEKVYELMRACWQWNPSDRPSFAEIHQAF
ETMFQES SISDEVEKELGKRGT

1. Utiliser la matrice BLOSUM62 comme matrice de distance, avec ouverture de gap de coût 11 et extension de coût 1, pour aligner ces deux séquences avec l'algorithme de Needleman Wunsch.

2. Comparer votre résultat avec l'algorithme Needleman Wunsch fourni par blast.

3. Comparer les structures de ces deux protéines sur le site Protein Data Bank. Que constatez vous ?

4. Sur quel algorithme se base BLAST général ?