IFT3295 - TP2

1er octobre 2017

Ce TP est à faire en équipe de deux. Vous devez le rendre au plus tard le lundi 16 Octobre avant la démo.

Repliement d'ARNs en tige-boucle (45pts)

Étant donné une séquence d'ARN S, on veut retrouver le repliement en tige-boucle (figure 1) de S qui maximise le nombre d'appariement de bases.

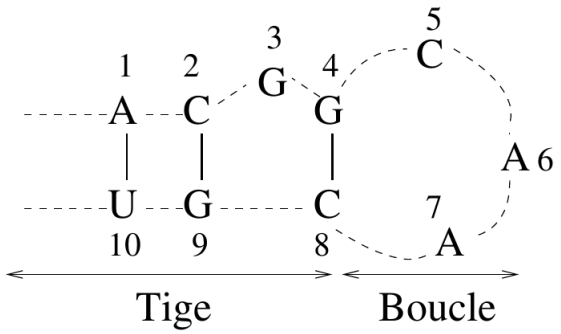

FIGURE 1 – Tige-boucle

Notez qu'un tige boucle peut contenir des nucléotides non-appariés et que seuls les appariements (A-U et G-C) sont considérés.

Ce problème peut être ramené à celui de l'identification du plus grand nombre de "match" entre une sous-séquence de S et une autre de la séquence inverse et complémentaire S_r de S. Il s'agit de la séquence obtenue en lisant S de droite à gauche et en remplaçant chaque nucléotide

par le nucléotide complémentaire (A par U ; C par G et inversement) Par exemple, si S est la séquence de la Figure 1 ($S = ACGGCAACGU$), alors $S_r = ACGUUGCCGU$.

Décrivez un algorithme de programmation dynamique qui remplit une table M (indicée à 0, où les lignes/colonnes ayant un indice à 0 représentent des chaînes vides) et permet de retrouver un repliement en tige-boucle maximisant le nombre d'appariements. Pour ce faire :

1. À quoi correspondent les cellules de M sur l'anti-diagonale ($M[i, |S| - i]$) ? Faut-il remplir toute la table ?
2. Donnez les équations d'initialisation et de récurrence pour remplir la table M.
3. Décrivez comment on peut retrouver un repliement en tige-boucle maximisant les appariements de nucléotides.
4. Appliquez votre algorithme à : $GCGUGCUUGCGUGCACG$. On vous demande la table de programmation dynamique, le score, ainsi que le repliement.
5. Pour être stable, la boucle de la tige-boucle doit contenir au moins 3 nucléotides. Décrivez une façon de modifier votre algorithme afin d'avoir des tige-boucles avec au moins 3 nucléotides dans la boucle.
6. Décrivez une modification de votre l'algorithme qui permet les appariements "G-U"
7. Dans cette section, on désire améliorer l'algorithme précédent en considérant des scores d'empilement d'appariements, ainsi que des pénalités pour les nucléotides non-appariées (Considérez à nouveau que seul les appariements G-C et A-U sont permis). On vous demande de garder des coûts unitaires pour les appariements de bases et considérer les coûts suivants pour l'empilement (N désigne tout nucléotide non-apparié) :

	G-C	A-U	N
G-C	2	1	0
A-U	1	0	0
N	0	0	0

Considérer également un coût de -1 par nucléotides non-appariés.

(a) Décrivez un algorithme capable de retrouver le score maximal de repliement, ainsi que la structure secondaire correspondante. Notez que vous pouvez proposer un algorithme totalement différent de celui des questions précédentes.

(b) Quelle est la complexité de votre algorithme en temps et espace ?

(c) Quel score obtenez vous pour la séquence GCGUGCUUGCGUGCACG ?

Alignement multiple de séquences (55pts)

Alignement avec arbre étoile (40pts)

Soit $\mathbb{S} = S_1, S_2, S_3, S_4, S_5$ un ensemble de séquences protéiques appartenant à la même famille de gène (voir fichier *sequences.fasta*).

On vous demande d'effectuer un alignement multiple de ces séquences par la méthode de l'étoile centrale. Pour ce faire, vous disposez du fichier *BLOSUM62.txt* qui contient le score de l'alignement entre chaque paire d'acide aminé. On vous demande d'utiliser la pénalité affine de gap : $-(a + kb)$ avec $a = 10$ et $b = 1$.

1. Grâce à l'algorithme d'alignement global, calculez la matrice des scores de similarité entre toutes les paires de séquences.
2. En déduire la séquence centrale S^* de $\mathbb{S}$
3. Construire un alignement multiple de $\mathbb{S}$ en utilisant la méthode de la séquence centrale. Vous devez illustrer les différentes étapes pour la construction de votre alignement.
4. Quelle est le score SP de votre alignement ?
5. Donnez la séquence consensus Z de l'alignement et son pourcentage d'identité avec chacune des séquences de $\mathbb{S}$. Notez que le pourcentage d'identité entre deux séquences **alignées** S_x et S_y désigne le pourcentage de positions comprenant des acides aminés identiques.

Outils bioinformatique (15pts)

On désire identifier la nature de chacune des séquences dans le fichier *sequences.fasta*.

1. En utilisant BLASTP, identifier le nom du gène, l'espèce d'origine et la fonction de chacune des protéines de *sequences.fasta*.
2. Utilisez le programme Clustal (vous pouvez aussi vous servir de ce site `http://www.ebi.ac.uk/Tools/msa/clustalo/`) avec les paramètres par défaut pour faire un alignement multiple des séquences, puis comparez le score SP entre cet alignement et celui que vous avez obtenu à la section précédente.