IFT3295 - TP2

Par Zohreh Kheirinia et Tanguy Invernizzi

Repliement d'ARNs en tige-boucle

1. À quoi correspondent les cellules de M sur l'anti-diagonale (M[i, |S| – i]) ? Faut-il remplir toute la table ?

Les cellules de M sur l'antidiagonale correspondent à l'alignement d'un nucléotide avec lui-même. Il ne faut pas remplir toute la table, uniquement la moitié supérieure de l'antidiagonale.

2. Donnez les équations d'initialisation et de récurrence pour remplir la table M.

Les valeurs de la première ligne seront : $\forall i, j \qquad V_{(i,0)} = 0$

Les valeurs de la première colonne seront : $\forall i, j \qquad V_{(0,j)} = 0$

Quand à la récurrence, elle sera égale à :

$$\ddot{D}(i,j) = \max \begin{cases} D(i-1,j) \\ D(i,j-1) \\ \begin{cases} D(i-1,j-1) \text{ si mismatch} \\ D(i-1,j-1)+1 \text{ si match} \end{cases} \end{cases}$$

3. Décrivez comment on peut retrouver un repliement en tige-boucle maximisant les appariements de nucléotides.

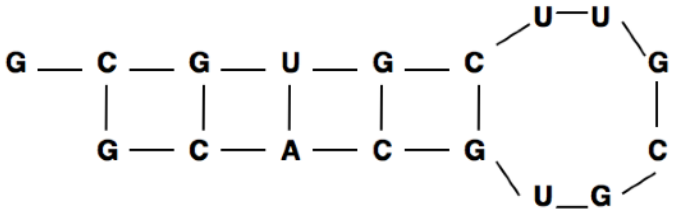
Si l'on veut maximiser les appariements de nucléotides, on peut simplement uniquement valoriser l'appariement. Cela veut dire que les structures qui ne comportent pas d'appariements, comme les bulge, ne seront pas pénalisante pour un alignement donné.

4. Appliquez votre algorithme à : *GCGUGCUUGCGUGCACG*. On vous demande la table de programmation dynamique, le score, ainsi que le repliement.

			S																
			G	C	G	U	G	C	U	U	G	C	G	U	G	C	A	C	G
Sr		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	X
	G	0	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	X	X
	U	0	0	0	2	3	3	3	3	3	3	3	3	3	3	3	X	X	X
	G	0	1	1	2	3	4	4	4	4	4	4	4	4	4	X	X	X	X
	C	0	0	2	2	3	4	5	5	5	5	5	5	5	X	X	X	X	X
	A	0	0	0	2	3	4	5	5	5	5	5	5	X	X	X	X	X	X
	C	0	0	1	2	3	4	5	5	5	5	6	X	X	X	X	X	X	X
	G	0	1	1	2	3	4	5	5	5	6	X	X	X	X	X	X	X	X
	C	0	0	2	2	3	4	5	5	5	X	X	X	X	X	X	X	X	X
	A	0	0	0	2	3	4	5	5	X	X	X	X	X	X	X	X	X	X
	A	0	0	0	2	3	4	5	X	X	X	X	X	X	X	X	X	X	X
	G	0	1	1	2	3	4	X	X	X	X	X	X	X	X	X	X	X	X
	C	0	0	2	2	3	X	X	X	X	X	X	X	X	X	X	X	X	X
	A	0	0	2	2	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	C	0	0	2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	G	0	1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	C	0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Figure 1 : Table de programmation dynamique pour GCGUGCUUGCGUGCACG

Pour retrouver un des repliements optimaux, on cherche le meilleur score d'alignement sur la grille et on remonte les flèches, comme dans un algorithme de distance d'édition. Un repliement optimal sera (chemin utilisé dessiné en rouge) :



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.

Pour assurer le fait qu'on ait au moins 3 nucléotides dans la boucle, on peut tout simplement retirer des antidiagonales. Ainsi, pour avoir au moins 3 nucléotides dans la boucle, on ignorerait deux antidiagonale après celle qui est déjà ignorée, car elle représente celle des appariements des nucléotides avec eux-mêmes. Cela empêchera l'algorithme d'apparier les 4 caractères centraux du repliement.

6. Décrivez une modification de votre l'algorithme qui permet les appariements "G-U"

Si un appariement est possible entre G et U, il suffit de modifier la récurrence pour la faire agir comme un match quand elle rencontre un appariement G-U (soit un couple C-U ou G-A quand l'une des deux séquences est complémentaire inversée)

$$\ddot{D}(i,j) = \max \begin{cases} D(i-1,j) \\ D(i,j-1) \\ \begin{cases} D(i,j-1) \text{ si mismatch} \\ D(i-1,j-1)+1 \text{ si match} \\ D(i-1,j-1)+1 \text{ si C-U ou G-A} \end{cases} \end{cases}$$

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

a) Décrivez un algorithme capable de retrouver le score maximal de repliement, ainsi que la structure secondaire correspondante.

Les valeurs de la première ligne seront : $\forall i,j \qquad V_{(i,0)} = 0$

Les valeurs de la première colonne seront : $\forall i,j \qquad V_{(0,j)} = 0$

La récurrence sera égale à :x

$$D(i,j) = \max \begin{cases} D(i-1,j) \\ D(i,j-1) \\ \begin{cases} D(i-1,j-1) \text{ si mismatch} \\ D(i-1,j-1)+1 \text{ si match} \\ D(i-1,j-1)+1+p(S[i],S_r[j]) \text{ si match avec S[i] = S[i-1] et Sr[j] = Sr[j-1]} \end{cases} \end{cases}$$

$$p(S[i],S[j]) =$$

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

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

La complexité de notre algorithme sera égale à $O(n^2)$ en espace et en temps, où n est la taille de la séquence étudiée.

Alignement multiple de séquences

Alignement avec arbre étoile

1. Grâce à l'algorithme d'alignement global, calculez la matrice des scores de similarité entre toutes les paires de séquences

	S1	S2	S3	S4	S5	Σ
S1	808	54	20	34	80	996
S2	54	983	94	143	77	1351
S3	20	94	737	86	16	953
S4	34	143	86	809	20	1092
S5	80	77	16	20	772	965

2. En déduire la séquence centrale S* de S

La séquence ayant le plus gros score de similarité avec les autres (c'est-à-dire la somme de tous les scores qu'elle a obtenus par rapport aux autres séquences et elles-mêmes) est la séquence centrale. Ici, c'est la séquence S2, avec un sigma de $\Sigma_{S2} = 1351$ qui est la séquence centrale.

3. Construire un alignement multiple de S en utilisant la méthode de la séquence centrale. Vous devez illustrer les différentes étapes pour la construction de votre alignement

Pour créer l'alignement multiple par la méthode de la séquence centrale, on aligne une à une les différentes séquences de l'arbre avec la séquence centrale, qu'on incorporer à chaque fois dans l'alignement à A. Si un nouvel indel a été rajouté dans le nouvel alignement, on rajoute un espace à chaque ligne à la colonne correspondante. Notre alignement sera donc :

```
MEKVPGEMEIERRERSEELSEAERKAVQATWARLYANCEDVGVAILVRFFVNFPSAKQYFSQFKHMEEPLEM
ERSPQLRKHACRVMG_ALNTVVENLHDPEK_V_SSVLSLVGKAHALKHKVEPVYFKILSGVILEVIAEEFANDF
PPETQRAWAKLRGLIYSHVTAAYKEVGWVQQVPNATTPPATLPSSGP_

M_____G_EIGF___TEKQ_EAL___VKESWEILKQDIPKYSLHFFSQILEIAPAAKGLFSFLRDSDE__VPHNNP
KLKAHAVKVFKMTCETAIQLREEGKVVVADTTLQYLGSIHLKSGVIDP_HFEVVKEALLRTLKEGLGEKYNEEV
EGAWS__Q__AYDHLALA______IK_________TEMKQEES_

M_________________VLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTY___FPHF__D_LSHGSA
QIKGHGKKVVAALIEAANHI_DD__IAGTLSKLSDLHAHKLRVDPVNFKLLGQCFLVVVAIHHPAALTPEVHASL
DKFLCAVGTVLTAKYR_____________________________

M_________________GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKA
SEDLKKHGATVLTALGGILKKKGHHE__AE_IKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQ
GAMNKALELFRKDMASNYKELGF_____________________QG____

M_____________ER___L_ESEL__IRQSWRAVSRSPLEHGTVLFSRLFALEPSLLPLFQYNGRQFSSPEDCLSS
PEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLATLGR____KHRAVGVRLSSFSTVGESLLYMLEKCLGPDFTPA
TRTAWS__Q_L________YGAV__VQAMSRGW________DGE
```

4. Quel est le score SP de votre alignement ?

Le score SP de l'alignement est égal somme des scores des alignements induits pour chaque paire de séquences dans A. Ici, il est égal à 3110.

5. Donnez la séquence consensus Z de l'alignement

La séquence consensus Z d'un alignement est le caractère apparaissant le plus de fois à chaque position de l'alignement. Dans notre cas, la séquence Z sera égale à :

M _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ K L S E A E K _ _ V K A S W A K L Y A D I P E Y G A E I L S R L F A I A P S A K Q Y F S Q F K H Q K S P _ E M H A S P Q L K K H A A K V V A A L I E A A V Q N G H D P _ I A E _ I S A L S Y L H A A H L K I A V V Y P K Y F K I C I G V V L A I I H E G A L G P D A Q E A Q R A A L S L Q G _ I Y A H K Y A A _

Outils bioinformatique

1. En utilisant BLASTP, identifier le nom du gène, l'espèce d'origine et la fonction de chacune des protéines de *sequences.fasta*.

>S1

Nom: hemoglobin 2

Gene : AHB2 GLB2, At3g10520, F13M14.20, F18K10.9

Organism: Arabidopsis thaliana (Mouse-ear cress)

Fonction :Peut ne pas agir comme une protéine de stockage ou de transport de l'oxygène, mais pourrait agir comme un capteur d'oxygène ou jouer un rôle dans le transfert d'électrons, éventuellement à une molécule d'oxygène liée. A une faible affinité pour O2.

Conserved domains on [gi|2581785|gb|AAB82770|]

>S2

Nom : Cytoglobin-like

Gene : CB1_000265021

Organism : Camelus ferus (Wild bactrian camel)

Fonction :est une globine respiratoire constituée, chez l'homme et la souris, par 190 acides aminés. Elle est produite par le gène (CYGB), localisé sur le chromosome 17q25. Contrairement à d'autres globines tissu-spécifiques (hémoglobine produite par les globules rouges, myoglobine dans les muscles, neuroglobine dans le système nerveux), elle semble ubiquitaire c'est-à-dire exprimée dans la totalité des types cellulaires de l'organisme2.

Conserved domains on [gi|296476010|tpg|DAA18125|]

>S3

Nom: Hemoglobin subunit alpha-A

Gene: HBAA

Organism: Gallus gallus (Chicken)

Fonction : Impliqué dans le transport de l'oxygène du poumon vers les divers tissus périphériques.

Conserved domains on [gi|63014|emb|CAA42606|]

>S4

Nom : myoglobin, partial

Gene : MB

Fonction : est une métalloprotéine contenant du fer présente dans les muscles des vertébrés, et particulièrement des mammifères. Elle est apparentée structurellement à l'hémoglobine, mais a pour fonction de stocker l'oxygène O2 plutôt que de le transporter. Comme l'hémoglobine, elle utilise l'hème comme groupe prosthétique, et est donc une hémoprotéine ; contrairement à l'hémoglobine, en revanche, la myoglobine est une protéine monomérique, c'est-à-dire qu'elle n'est formée que d'une seule sous-unité globine.

Conserved domains on [gi|127661|sp|P02144|]

>S5

Nom : neuroglobin

Gene : Ngb

Organism : Homo sapines

Fonction : La neuroglobine est une globine de stockage et de transport de l'oxygène dans le système nerveux. À ce titre, elle présente beaucoup d'analogies avec la myoglobine dont une structure monomérique et une forte affinité pour l'oxygène

2. Utilisez le programme Clustal 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.

Multiple sequence alignment

```
S3      ----------------MVLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYF
S2      MEKVPGEMEIERRERSEELSEAERKAVQATWARLYANCEDVGVAILVRFFVNFPSAKQYF
S4      ----------------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKF
S1      -------------MGEIGFTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKGLF
S5      ------------------MERLESELIRQSWRAVSRSPLEHGTVLFSRLFALEPSLLPLF
                         :   :   :   :  :       .   : :::   *     *

S3      PHFDLS-H------GSAQIKGHGKKVVAALIE------AANHIDDIAGTLSKLSDLHAHK
S2      SQFKHM-EEPLEMERSPQLRKHACRVMGALNTVV---ENLHDPEKVSSVLSLVGKAHALK
S4      DKFKHL-KSEDEMKASEDLKKHGATVLTALGGIL---KKKGHH---EAEIKPLAQSHATK
S1      SFLRDSDEVPHN---NPKLKAHAVKVFKMTCETAIQLREEGKVVVADTTLQYLGSIHLK-
S5      QYNGRQFSSPEDCLSSPEFLDHIRKVMLVID-AA--VTNVEDLSSLEEYLATLGRKHRA-
                      . .:  *   *.              .       :  :.  *

S3      LRVDPVNFKLLGQCFLVVVAIHHPAALTPEVHASLDKFLCAVGTVLTAKYR---------
S2      HKVEPVYFKILSGVILEVIAEEFANDFPPETQRAWAKLRGLIYSHVTAAYKEVGWVQQVP
S4      HKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG---
S1      SGVIDPHFEVVKEALLRTLKEGLGEKYNEEVEGAWSQAYDHLALAIKTE-----MKQEES
S5      VGVRLSSFSTVGESLLYMLEKCLGPDFTPATRTAWSQLYGAVVQAMSRG-----WDGE--
          :    :. :   ::  :           .. :  :    .   :

S3      --------------
S2      NATTPPATLPSSGP
S4      --------------
S1      --------------
S5      --------------
```

Matrice

	s1	s2	s3	s4	s5
1: S3	100.00	28.17	23.74	15.22	15.44
2: S2	28.17	100.00	27.92	17.53	20.81
3: S4	23.74	27.92	100.00	15.17	17.93
4: S1	15.22	17.53	15.17	100.00	23.65
5: S5	15.44	20.81	17.93	23.65	100.00

Phylogenetic Tree