SORBONNE UNIVERSITÉ

Devoir Maison

Student name: *Durand Enzo 21107517*

Course: *Algorithmes sur les arbres et les graphes en bio-informatique (AAGB)*
Due date: *November 24th, 2021*

Exercice 1

1. Qu'est ce qu'une matrice de coût ? Quelles données pour la construire ? Comment la construire ? Donner un exemple.

- Une matrice de coût est une matrice qui contient les valeurs associées aux match/mismatch pour chaque lettre de l'alphabet étudié. Les indices de ligne et de colonne correspondent aux deux séquences étudiées.
- Pour construire une matrice de coût, on a donc besoin des différentes lettres contenues dans les séquences que l'on compare ainsi que des valeurs de match/mismatch associées aux lettres de l'alphabet.
- Pour exemple, on a la matrice blossum qui permet d'avoir les valeurs de match et mismatch pour un alphabet correspondant à des acides aminés:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

2. Comment comparer deux séquences ? Quelle est l'utilité de la matrice de coût pour cela ? Citer deux méthodes vues en TD pour comparer deux séquences.

- Pour comparer deux séquences, il faut trouver l'alignement global ou les alignements locaux de son alphabet. On les trouve en donnant des valeurs positives

à des lettres qui concordent et en donnant une pénalité aux lettres qui ne concordent pas. Lorsque l'on a pas de matrice de coût, on donne la même valeur positive et la même valeur de pénalité à toutes les lettres suivant les match/mismatch.

- La matrice de coût permet de donner une pénalité différente suivant la différence des lettres dans la séquence. En effet, il est parfois judicieux de punir plus fortement une différence d'acide aminé polaire/apolaire comparé à une différence d'acide aminé du même groupe.
- Les deux méthodes que nous avons étudié en TD sont le Dot Plot et l'algorithme de Needleman & Wunsch.

3. Définisser orthologie, paralogie, analogie et homologie. Pourquoi chercher des homologies entre séquences ?

- L'orthologie est le lien évolutif entre deux gènes présents chez deux espèces différentes. On dit de deux séquences qu'elles sont orthologues si elles descendent d'une séquence unique présente chez leur dernier ancêtre commun. Le gène impliqué résulte d'une spéciation.
- La paralogie est le lien évolutif entre deux gènes présents chez deux espèces différentes. On dit de deux séquences qu'elles sont paralogue si elles descendent d'une séquence unique présente chez leur dernier ancêtre commun mais contrairement à l'othologie, le gène impliqué résulte d'une duplication suivie d'une spéciation. Ce gène peut donc avoir de nouvelles fonctions.
- L'analogie est une similitude entre deux expressions du phénotype qui ont les mêmes fonctions biologiques chez deux espèces différentes. Ces deux expressions phénotypiques ne proviennent pas d'une descendance avec un ancêtre commun. Le gène est donc apparu une fois sur chacun des ancêtres communs des deux espèces.
- L'homologie est un lien évolutif entre expressions phénotypiques de deux espèces différentes, provenant d'une descendance avec un ancêtre commun.
- Il faut chercher les homologies pour pouvoir construire des arbres phylogénétiques afin de mieux comprendre l'évolution de certaines espèces. On peut faire cela avec des algorithmes comme UPGMA ou NJ.

Exercice 2

1. Définissez l'horlogie moléculaire et expliquer pourquoi on emploie ce terme pour l'algorithme UPGMA.

- L'horloge moléculaire est une hypothèse en biologie selon laquelle les mutations du génome se feraient à vitesse constante. Grâce à cette hypothèse, on peut donc mettre en corrélation le taux de mutation et les divergences génétiques de différentes espèces afin de trouver une chronologie d'apparition des espèces.

- On emploie ce terme pour UPGMA car cet algorithme construit un arbre phylogénétique à partir d'une matrice de distance. Il donne des tailles de branches équivalentes aux différentes espèces proportionnellement à leurs écarts de nucléotides. Il construit donc cet arbre en se basant sur l'hypothèse de l'horloge moléculaire.

2. Quelle est la partie de l'algorithme Neighbour-Joining qui permet une amélioration par rapport à UPGMA ? Pourquoi ?

- Contrairement à UPGMA, l'algorithme NJ permet aux branches de l'arbre phylogénétique de faire des tailles différentes au lieu d'une simple moyenne des distances.
- En réalité, les espèces ne sont pas soumises aux mêmes pressions de séléction, la théorie de l'horloge moléculaire n'est donc pas très précise. Le taux de mutation génétique varie aussi en fonction des différentes parties d'un génome. NJ est donc un algorithme plus précis et plus vraissemblable que UPGMA.

3. Dérouler l'algorithme Neighbour-Joining pour la matrice de distance suivante :

	A	B	C	D
A	0	5	4	7
B	5	0	5	6
C	4	5	0	3
D	7	6	3	0

- Etape 1 :
  - On calcule d'abord les valeurs $u_i = \frac{\sum_{j=0} d_{ij}}{n-2}$:

$$u_a = \frac{5+4+7}{2} = 8$$
$$u_b = \frac{5+5+6}{2} = 8$$
$$u_c = \frac{4+5+3}{2} = 6$$
$$u_d = \frac{7+6+3}{2} = 8$$

  - On calcule ensuite les valeurs Q : $q_{i,j} = d_{ij} - u_i - u_j$:

$$q_{a,b} = 5 - 8 - 8 = -11$$
$$q_{a,c} = 4 - 8 - 6 = -10$$
$$q_{a,d} = 7 - 8 - 8 = -9$$
$$q_{b,c} = 5 - 8 - 6 = -9$$
$$q_{b,d} = 5 - 8 - 8 = -10$$
$$q_{c,d} = 3 - 8 - 6 = -11$$

- On choisit la valeur Q minimale : Ici la valeur Q pour a,b.
- On recalcule les distances au nouveau noeud AB $u_{ij,k} = \frac{d_{ij}}{2}$:

$$d_{ab,c} = \frac{4+5-5}{2} = 2$$

$$d_{ab,d} = \frac{7+6-5}{2} = 4$$

- On calcule les distances des branches :

$$d_{a,ab} = \frac{5}{2} + \frac{(8+8)}{2} = 2.5$$

$$d_{b,ab} = \frac{5}{2} + \frac{(8+8)}{2} = 2.5$$

- Cela nous donne le nouveau tableau :

	AB	C	D
AB	0	2	4
C	2	0	3
D	4	3	0

- Étape 2 :
  - On calcule d'abord les valeurs U :

$$u_{ab} = \frac{2+4}{1} = 6$$

$$u_c = \frac{2+3}{1} = 5$$

$$u_d = \frac{4+3}{1} = 7$$

  - On calcule ensuite les valeurs Q :

$$q_{ab,c} = 2-6-5 = -9$$

$$q_{ab,d} = 4-6-7 = -9$$

$$q_{c,d} = 3-7-5 = -9$$

  - On choisit la valeur Q minimale : Ici la valeur Q pour ab,c.
  - On recalcule les distances au nouveau noeud ABC :

$$d_{abc,d} = \frac{3+4-2}{2} = 2.5$$

  - On calcule les distances des branches :

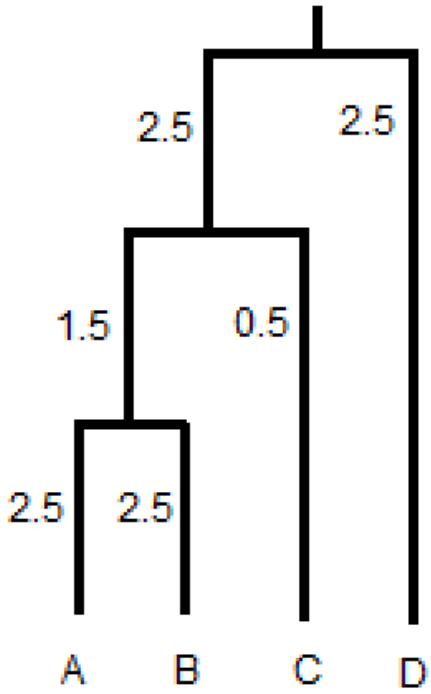
$$d_{ab,abc} = \frac{2}{2} + \frac{(6-5)}{2} = 0.5$$

$$d_{c,abc} = \frac{2}{2} + \frac{(5-6)}{2} = 1.5$$

– Cela nous donne le nouveau tableau :

	ABC	D
ABC	0	2.5
D	2.5	0

- On peut maintenant construire l'arbre de phylogénétique :



- Après vérification, on se rend compte que la matrice de distance n'était pas additive ni ultramétrique. Il est donc impossible d'obtenir les bonnes valeurs pour les branches de l'arbre de phylogénétique associé à NJ. Exemple : La distance entre A et D dans la matrice est de 7. Lorsque l'on regarde le chemin de A a D dans l'arbre on a 2.5+1.5+2.5+2.5 = 9. On ne retombe donc pas sur 7. On a donc une approximation de la taille des branches.