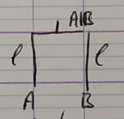
Correction AAGB TD2

AAGB - TD2

1) UPGMA

1. Unweighted Pair Group w/ Mean Arithmetic
2. Entrée : matrice de distance
   Méthode : clustering hiérarchique
   Sortie : Arbre racinié représentant entre la distance entre chaque élément.
3. Idée : A chaque étape, on considère les deux clusters les plus proches, et on les combine en un cluster de niveau supérieur
   Avantage : simple (le plus simple) pour la construction d'arbres.
   Inconvénient : Hypothèse de l'horloge moléculaire, ie. même vitesse d'évolution partout. Cela se traduit par le fait que l'algorithme construit de, à chaque étape, des branches entre les deux espèces à clusteriser qui sont de même longueur :



En réalité ce n'est pas du tout le cas, et UPGMA ne permet donc pas de reconstruire un arbre correspondant à l'histoire évolutive réelle. (voir ultramétricité).
Aussi, donne un arbre racinié

4. 1ère étape : trouver les deux clusters les plus proches, grâce à la matrice de distances donnée. Soient A et B les deux clusters tels que pour tout cluster les plus proches, et $\{(x_1, x_2, \dots x_n)$ $(y_1, x_2, \dots y_m)$ les espèces respectivement présentes dans les clusters A et B.

   2ème étape : On met à jour la matrice de distance en remplaçant les clusters A et B par le cluster $\{A,B\}$ et en calculant la distance des éléments restant à ce nouveau cluster

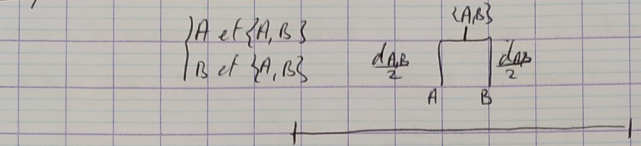
4. 1ère étape : Trouver les deux clusters $A = (x_1 \dots x_n)$ et $B = (y_1 \dots y_m)$ les plus proches.

   2ème étape : Mettre à jour la matrice de distance en calculant la distance du nouveau cluster $\{A,B\}$ à tous les autres clusters.

AAGB - TD2 (2)

Pour tout cluster C : $d_{\{A,B\},C} = \frac{|A| d_{A,C} + |B| d_{B,C}}{|A| + |B|}$, où $|A|$ et $|B|$ sont les cardinaux des clusters A et B, en l'occurrence n et m.

3e étape : construction de l'arbre : deux branches de longueur $\frac{d(A,B)}{2}$ entre $\{A,B\}$ et $\{A,B\}$



5) Types de données : peut être n'importe quoi : nom d'alignement, caractères, etc.

6. (voir 3.) Complexité : Si n espèces, (n-1) étapes, avec (n-1) mises à jour.

7. (voir feuille)

---

2) Neighbor Joining

1) Amélioration : permet d'obtenir un arbre non raciné.
N'est pas basé sur l'horloge moléculaire. La distance entre deux espèces et leur ancêtre commun n'est pas forcément égale.

2. Principe : à la différence de UPGMA, on ~~détermine~~ choisit le couple de clusters à réunir s'ils sont proches l'un de l'autre mais aussi loin des autres.

3. (voir feuille)

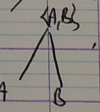
Note sur ultramétrique / additif

A) Ultramétrique

Une distance est dite ultramétrique si, pour tout (x, y, z), $d(x,z) \leq \max(d(x,y), d(y,z))$

. Cela implique que tous les triangles sont isocèle (sinon, il existe un côté plus grand, et donc un couple (x,z) tel que $d(x,z) > \max\{d(x,y), d(y,z)\}$)

Or UPGMA ne permet de construire que des triangles isocèles :



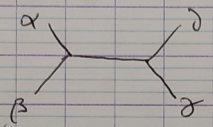
Donc si la condition d'ultramétricité n'est pas vérifiée pour tout triplet d'espèces,

alors UPGMA ne peut pas construire un arbre correspondant à la matrice de distances avoisinée. ▷

B) Additif

Soit une matrice de distances de taille (n×n) associé à n espèces (ou caractères, ...)
Cette matrice est additive si pour tout quadruplet $(\alpha, \beta, \delta, \gamma)$ d'espèces, la condition des quatre points est vérifiée :

$$d_{\alpha,\beta} + d_{\delta,\gamma} \leq \max \{ d_{\alpha,\delta} + d_{\beta,\gamma}, d_{\alpha,\gamma} + d_{\beta,\delta} \}$$



La condition d'additivité est nécessaire pour qu'un arbre fidèle à la matrice de distances puisse être construit.

2) Neighbor Joining principle

2) A chaque étape, pour une matrice de taille n×n :

a) On calcule $\begin{cases} \forall i \in [1,n] \quad u_i = \dfrac{\sum_k d_{i,k}}{(n-2)} \\ \forall (i,j) \in [1,n]^2 \quad Q_{i,j} = d_{i,j} - u_i - u_j \end{cases}$

b) On choisit le (ou un) couple (i,j) pour lequel $Q_{i,j}$ est min.

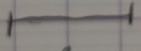
c) On met à jour la matrice de distance :

$$\forall k \quad d_{(i,j),k} = \frac{d_{i,k} + d_{j,k} - d_{i,j}}{2}$$

d) On détermine la longueur des deux branches reliant $\begin{cases} i \text{ à } (i,j) \\ j \text{ à } (i,j) \end{cases}$:

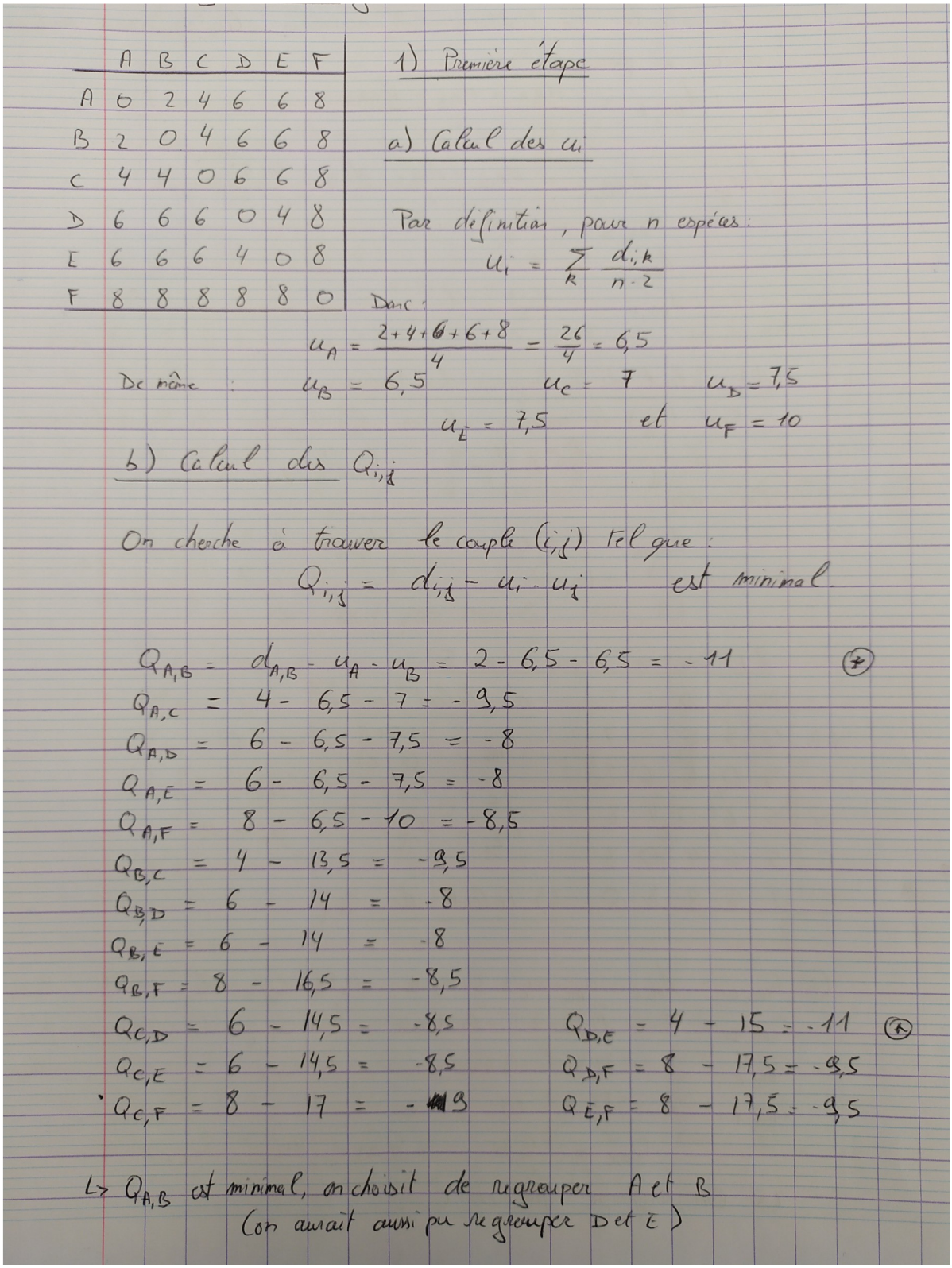
$$d_{i,(i,j)} = \frac{d_{i,j} + u_i - u_j}{2} \quad \text{et} \quad d_{j,(i,j)} = \frac{d_{i,j} + u_j - u_i}{2}$$

e) S'il ne reste plus que deux clusters α et β, on les relie par une branche de longueur $d_{\alpha,\beta}$



(S'il y a plusieurs valeurs minimales pour $Q_{i,j}$, le choix du couple (i,j) n'influe pas sur l'arbre final.)

3. Exercice correction



	A	B	C	D	E	F
A	0	2	4	6	6	8
B	2	0	4	6	6	8
C	4	4	0	6	6	8
D	6	6	6	0	4	8
E	6	6	6	4	0	8
F	8	8	8	8	8	0

1) Première étape

a) Calcul des u_i

Par définition, pour n espèces :

$$u_i = \sum_k \frac{d_{i,k}}{n-2}$$

Donc :

$$u_A = \frac{2+4+6+6+8}{4} = \frac{26}{4} = 6,5$$

De même : $u_B = 6,5$ $u_C = 7$ $u_D = 7,5$

$u_E = 7,5$ et $u_F = 10$

b) Calcul des $Q_{i,j}$

On cherche à trouver le couple (i,j) tel que :

$Q_{i,j} = d_{i,j} - u_i - u_j$ est minimal.

$Q_{A,B} = d_{A,B} - u_A - u_B = 2 - 6,5 - 6,5 = -11$ ⑦

$Q_{A,C} = 4 - 6,5 - 7 = -9,5$

$Q_{A,D} = 6 - 6,5 - 7,5 = -8$

$Q_{A,E} = 6 - 6,5 - 7,5 = -8$

$Q_{A,F} = 8 - 6,5 - 10 = -8,5$

$Q_{B,C} = 4 - 13,5 = -9,5$

$Q_{B,D} = 6 - 14 = -8$

$Q_{B,E} = 6 - 14 = -8$

$Q_{B,F} = 8 - 16,5 = -8,5$

$Q_{C,D} = 6 - 14,5 = -8,5$

$Q_{C,E} = 6 - 14,5 = -8,5$

$Q_{C,F} = 8 - 17 = -9$

$Q_{D,E} = 4 - 15 = -11$ ①

$Q_{D,F} = 8 - 17,5 = -9,5$

$Q_{E,F} = 8 - 17,5 = -9,5$

↳ $Q_{A,B}$ est minimal, on choisit de regrouper A et B

(On aurait aussi pu regrouper D et E)

c) Mise à jour des distances

On calcule la distance du nouveau cluster aux autres espèces k par :

$$d_{AB,k} = \frac{d_{A,k} + d_{B,k} - d_{AB}}{2}$$

donc : $d_{AB,C} = \frac{d_{A,C} + d_{B,C} - d_{A,B}}{2} = \frac{4+4-2}{2} = 3$

$d_{AB,D} = \frac{12-2}{2} = 5$

$d_{AB,E} = 5$

$d_{AB,F} = \frac{16-2}{2} = 7$

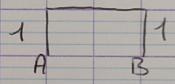
d) Calcul de la longueur des branches

Soit (ij) le nouveau cluster les longueurs des branches sont :

$$d_{i,ij} = \frac{d_{ij}}{2} + \frac{u_i - u_j}{2} \quad \text{et} \quad d_{j,ij} = \frac{d_{ij}}{2} + \frac{u_j - u_i}{2}$$

On a donc :

$d_{A,AB} = \frac{2}{2} + 0 = 1$ et $d_{B,AB} = 1$



2) Deuxième étape (n=5)

	AB	C	D	E	F
AB	0	3	5	5	7
C	3	0	6	6	8
D	5	6	0	4	8
E	5	6	4	0	8
F	7	8	8	8	0

a) Calcul des u_i

$u_{AB} = \frac{3+5+5+7}{(5-2)} = 6,67$

$u_C = 7,67$ $\quad u_D = 7,67$

$u_E = 7,67$ $\quad u_F = 10,33$

b) Calcul des $Q_{i,j}$

(4) $Q_{AB,C} = d_{AB,C} - u_{AB} - u_C = 3 - 14,33 = -11,33$

$Q_{AB,D} = 5 - 14,33 = -9,33$

$Q_{AB,E} = 5 - 14,33 = -9,33$

$Q_{AB,F} = 7 - 17 = -10$

$Q_{C,D} = 6 - 15,33 = -9,33$

$Q_{C,E} = 6 - 15,33 = -9,33$

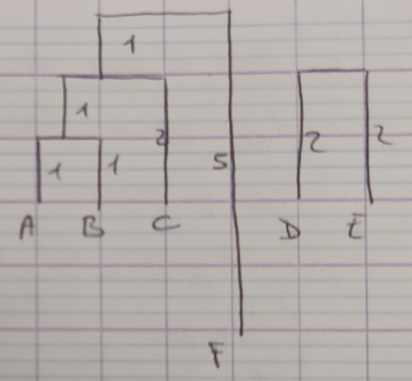
$Q_{CF} = 8 - 18 = -10$

$Q_{D,E} = -9,3$ $\quad Q_{D,F} = -10$ $\quad Q_{E,F} = -10$

↳ On regroupe (AB) et C.

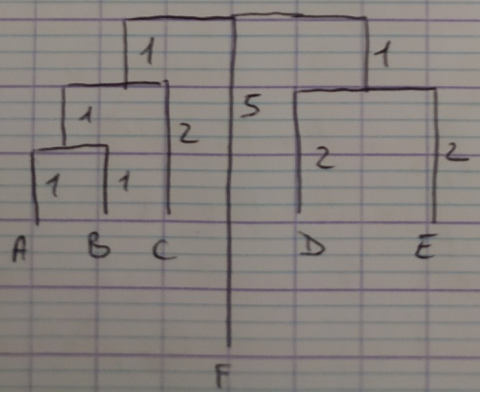
c) $d_{ABCF,DE} = \frac{3+3-4}{2} = 1$

d) $\begin{cases} d_{D,DE} = \frac{4}{2} + 0 = 2 \\ d_{i,DE} = 2 \end{cases}$



5) Dernière étape

~~$d_{ABCF,ABCDEF}$~~ ~~$d_{DE,ABCDEF}$~~ La résolution est finie, il suffit de connecter les deux clusters restant avec une branche de longueur $D_{ABCF,DE} = 1$



□

4. Additive Phylogeny