BIN702 - Série d'exercices sur la phylogénétique

Manuel Lafond

1 Méthodes par distance

Exercice 1: Considérez la matrice de distances suivante :

	A	B	C	D	E
A	0	20	20	20	8
B		0	16	16	20
C			0	10	20
D				0	20
E					0

a. Est-ce que cette distance est ultra-métrique ? Si oui, reconstruisez l'arbre ultramétrique et sinon, dites pourquoi.

b. Est-ce que cette distance est additive ?

Exercice 2: Montrez comment implémenter UPGMA en temps $O(n^2 \log n)$.

Indice 1 : il faut pouvoir mettre à jour les distances après fusion en temps $O(1)$ et utiliser une structure de données qui donne un accès rapide au minimum.

Indice 2 : quand on combine $A \cup B$ et qu'on doit mettre à jour la distance $D_{A\cup B,C}$, au lieu de recalculer la somme des paires de distances possibles pour en faire la moyenne, on peut utiliser $D_{(A\cup B),C} = (|A|D_{A,C} + |C|D_{B,C})/(|A| + |B|)$. Vous pouvez utiliser ce fait en boîte noire. Si vous voulez un défi, montrez que c'est correct.

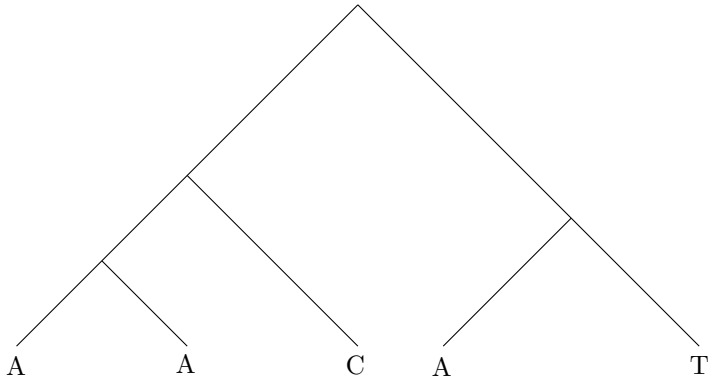
Exercice 3: Soit D une distance ultra-métrique et soit T l'arbre reconstruit avec UPGMA. Dans la version présentée en classe, nous n'avons pas montré en détail comment étiqueter les arêtes de façon à respecter les distances, et de façon à ce que chaque distance feuille-racine soit la même.

Donnez un algorithme qui trouve un étiquetage de T satisfaisant les conditions d'ultramétricité.

Exercice 4: Quelle est la complexité de l'algorithme Neighbor-Joining ? On peut argumenter qu'il s'implémente en temps $O(n^3)$, où n est le nombre de taxa qui étaient présent au départ.

2 Méthodes par caractères

Exercice 5: Soit l'arbre suivant :



a. Exécutez l'algorithme permettant de minimiser la distance Hamming, i.e. les changements sur cet arbre. Donnez les caractères candidats aux ancêtres ainsi qu'un étiquetage optimal.

b. Exécutez l'algorithme de programmation dynamique permettant de maximiser le score des branches sur cet arbre. Utilisez comme matrice de score

	A	C	G	T
A	2	-1	-2	-1
C	-1	2	-1	-1
G	-2	-1	1	-2
T	-1	-1	-2	3

Exercice 6: Supposez qu'il n'existe que les caractères A et C. Soient les séquences

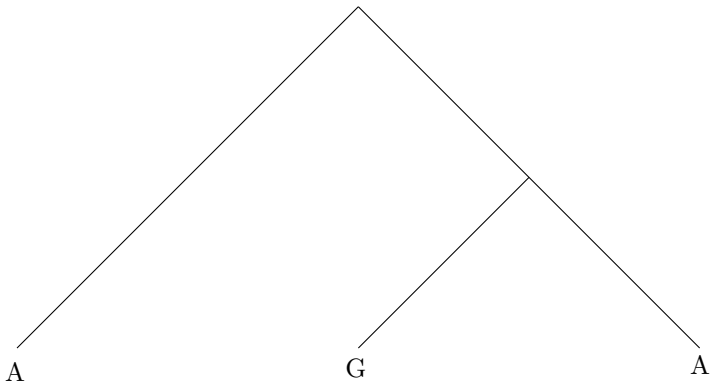
$$S_1 = ACAA$$
$$S_2 = AACA$$
$$S_3 = CAAC$$

Quel est l'arbre le plus parcimonieux, selon la distance Hamming ? C'est-à-dire, quel arbre minimise la somme des nombres de changements sur les branches pour l'ensemble des positions ?

Exercice 7: Dans l'algorithme de Sankoff-Rousseau, qui calcule le score maximum sur les branches d'un arbre, nous avons supposé que les noeuds avaient deux enfants. Montrez comment modifier l'algorithme pour trouver un étiquetage optimal des noeuds internes, et ce même si l'arbre n'est pas binaire (donc chaque noeud peut avoir un nombre arbitraire d'enfants).

Exercice 8: En classe, nous avons mentionné le nombre d'arbres binaires **enracinés** avec n feuilles étiquetées, qui est de $(2n-3)!! = (2n-3)(2n-5)(2n-7)\dots(3)(1)$. Démontrez cette identité.
Indice : vous pouvez utiliser le fait qu'un arbre binaire enraciné avec n feuilles a toujours $2n-2$ arêtes.
Par ailleurs, pour obtenir un arbre avec n feuilles, on peut démarrer d'un arbre avec $n-1$ feuilles puis "greffer" une feuille sur une des branches.

Exercice 9: Supposez qu'il n'existe que deux caractères possibles A et G. Soit l'arbre suivant :



Considérez que sur chaque branche, la probabilité qu'un caractère reste le même est 2/3 et la probabilité de changer est 1/3. On suppose que chacun des caractères a une probabilité 1/2 d'être à la racine. Quel est le score de vraisemblance de cet arbre ?

Exercice 10: (Défi complètement optionnel que vous devrez regarder seulement si vous êtes très enthousiaste)
Il existe une caractérisation des matrices ultramétriques qui ne regarde que trois points à la fois. Elle s'appelle la condition des trois points et dit qu'une distance D est ultra-métrique si et seulement si pour chaque trois taxa, on

peut les nommer x, y, z de façon à ce que $D_{xy} = D_{xz} \geq D_{yz}$. Vous pouvez vérifier ce fait sur les exemples vu en classe.
Démontrez la condition des trois points.