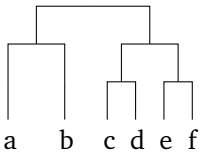
5 Inférence d'arbres phylogénétiques

5.1 Exercices

Exercice 24 : Implémenter l'algorithme de Fitch pour les données suivantes :

locus\indiv.	a	b	c	d	e	f
1	C	C	A	T	G	A
2	G	T	G	A	C	C
3	A	T	G	C	A	T



Combien de mutations a-t-on concernant chacun des sites, suivant le principe de parcimonie ?

Exercice 25 : Appliquer l'algorithme *unweighted pair group method with arithmetic mean* (UPGMA) pour reconstruire une phylogénie sur les données de distances génétiques suivantes :

	chien	ours	raton-laveur	phoque	chat
chien	0	32	48	50	98
ours	32	0	26	29	84
raton-laveur	48	26	0	44	86
phoque	50	29	44	0	89
chat	98	84	86	89	0

Exercice 26 : Un peu d'algèbre linéaire. Soit $n \in \mathbb{N}$. On note $I \in \mathcal{M}_n$ la matrice identité, et on définit $A \in \mathcal{M}_n$ par la matrice dont toutes les entrées valent $a_{ij} = 1/n$.

a) Que vaut A^k, pour tout $k \in \mathbb{N}$? *Attention à distinguer* $k = 0$.

b) En déduire la valeur de e^{sA}, puis de $e^{s(A-I)}$, pour tout $s \in \mathbb{R}$.

c) On rappelle que le modèle de Jukes-Cantor est le modèle d'évolution de l'ADN le plus simple (et assez inexact, rappelons-le), où l'on fait l'hypothèse que chaque locus, à partir de chaque état A, T, G ou C, peut muter indépendamment des autres à un taux $3\mu/4$, puis saute dans l'un des trois autres états possibles. Ainsi l'état d'un locus suit la chaîne de Markov à temps continu de matrice de transition

$$Q = \frac{\mu}{4}\begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}$$

En considérant $n = 4$, exprimer Q en fonction de A et I.

d) Si J_t désigne la chaîne de Markov de Jukes-Cantor, de matrice de transition Q, on rappelle que pour chaque paire d'états $i, j \in \{A, T, G, C\}$ et pour tout temps $t \geqslant 0$, on a

$$\mathbb{P}(J_t = j \mid J_0 = i) = (e^{tQ})_{ij}.$$

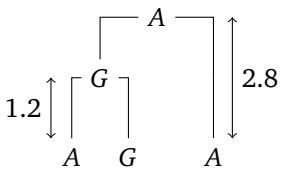
On rappelle la notation du cours $\eta_0(t) = \mathbb{P}(J_t = A \mid J_0 = A)$ et $\eta_1(t) = \mathbb{P}(J_t = T \mid J_0 = A)$. Déduire de ce qui précède que

$$\eta_0(t) = \frac{1}{4} + \frac{3}{4}e^{-\mu t} \quad \text{et} \quad \eta_1(t) = \frac{1}{4} - \frac{1}{4}e^{-\mu t}.$$

Exercice 27 : On dispose d'une généalogie et de données génétiques sur un locus, pour trois espèces.

Selon le modèle de Jukes–Cantor, on a calculé les probabilités suivantes :

t	1.2	1.6	2.8
$\eta_0(t)$	0.773	0.714	0.575
$\eta_1(t)$	0.076	0.095	0.142



Calculer la vraisemblance de l'arbre, selon le modèle "Yule/Jukes–Cantor".

Sans faire de calcul, quel serait une disposition plus vraisemblable des séquences ancestrales ?

5.2 Programmation

Exercice 28 : Écrire une fonction `fitch(tree)` qui implémente l'algorithme de Fitch (pour un seul locus). L'argument `tree` doit être un arbre dont les feuilles disposent d'une valeur représentant la base présente au locus considéré.

Exercice 29 : Écrire une fonction `UPGMA(distance_matrix)` qui implémente la méthode *unweighted pair group method with arithmetic mean* (UPGMA) à partir d'une matrice de distances.

Tester sur les données suivantes (exemple de Felsenstein, 2004) :

```
D = np.array([[  0,  32,  48,  51,  50,  48,  98, 148],
              [ 32,   0,  26,  34,  29,  33,  84, 136],
              [ 48,  26,   0,  42,  44,  44,  92, 152],
              [ 51,  34,  42,   0,  44,  38,  86, 142],
              [ 50,  29,  44,  44,   0,  24,  89, 142],
              [ 48,  33,  44,  38,  24,   0,  90, 142],
              [ 98,  84,  92,  86,  89,  90,   0, 148],
              [148, 136, 152, 142, 142, 142, 148,   0]])
# Pour avoir le nom des espèces correspondantes:
species = ['dog', 'bear', 'raccoon', 'weasel', 'seal', 'sea lion', 'cat', 'monkey']
```

Exercice 30 : Écrire une fonction `build_phylogeny(data)` qui utilise les deux exercices ci-dessus pour reconstruire une phylogénie (naïve), à partir de données génétiques encodées par une liste de tuple.

Exercice 31 :

a) Écrire une fonction `eta(mu, diff, t)` qui renvoie $\eta_0(t)$ si `diff` vaut `False`, et renvoie $\eta_1(t)$ si `diff` vaut `True`.

b) Écrire une fonction `likelihood_YJC(T, mu)` qui, étant donné un arbre muni de longueurs de branches et de séquences génétiques sur les nœuds, calcule la log-vraisemblance de l'arbre T selon le modèle Yule/Jukes–Cantor.