Phylogénie moléculaire

Modéliser l'évolution moléculaire

Lors de la duplication de l'ADN, des processus biochimiques se mettent en place avec un nombre suffisant de garde-fous pour assurer le minimum d'erreur.
Néanmoins des changements peuvent survenir.
Ils ont de plusieurs ordres mais les mutations les plus fréquentes, introduites lors de la copie de séquences d'ADN, concernent les substitutions de bases. Il s'agit simplement du remplacement d'une base par une autre sur certains sites de la séquence. On rappelle que lorsque une purine remplace une purine (nucléotides A et G) ou une pyrimidine une pyrimidine, on parle de *transition*, tandis qu'un échange entre ces classes est appelé une *transvection*.
Les transitions sont les plus fréquentes, observées plus souvent que les transvections, sans doute parce que les structures chimiques sont plus proches.
D'autres mutations sont observées, qui incluent des suppressions de bases, des insertions d'une ou de plusieurs bases consécutives ou encore des inversions de sections au sein d'une séquence. Ces dernières sont beaucoup plus rares, conduisant à des effets dramatique sur l'encodage des protéines associées.
Pour notre part, nous négligerons ces dernières mutations, à la fois pour rendre le modèle plus clair et le rendre mathématiquement accessible.

En se focalisant uniquement sur les substitutions de bases, le premier problème à savoir résoudre consiste à déterminer le nombre de mutations qui ont pu avoir lieu au cours de la descendance d'une séquence d'ADN.
Par exemple, supposons qu'une espèce S_2 descend d'une espèce intermédiaire S_1 qui a son tour descend d'une espèce ancestrale S_0. Imaginons que, pour chacune d'entre elle, un certain gène inclus la séquence :

$$S_0 :' AC\mathbf{C}TGC\mathbf{G}CTA...'$$

$$S_1 :' AC\mathbf{G}TGC\mathbf{A}CTA...'$$

$$S_2 :' AC\mathbf{G}TGC\mathbf{G}CTA...'$$

Si nous ne considérons que les séquences S_0 et S_2, seule une substitution peut être décelée sur les 10 bases. Dès lors, une proportion $p = 1/10$ de mutations par sites semble une bonne approximation du nombre de mutations ayant eu lieu pour passer de S_0 à S_2. La connaissance de S_1 nous permet de constater que cette approximation sous-évalue le nombre de mutations réellement réalisées, en fait ici 3/10.
L'objectif est de mettre en place un modèle mathématique permettant d'estimer le nombre de mutations réelles à l'appui uniquement des séquences initiales et finales.

Pour chaque site d'une séquence, nous noterons X_n variable aléatoire égale à 0, 1, 2 ou 3 selon qu'il est occupé à la génération n par de l'adénine, de la guanine, de la cytosine ou de la thymine.
Le vecteur $p_0 = (\mathbb{P}(X_0 = 0), \mathbb{P}(X_0 = 1), \mathbb{P}(X_0 = 2), \mathbb{P}(X_0 = 3))$ décrit la distribution ancestrale des bases.

① **Le modèle de Jukes-Cantor** : On suppose que les probabilités conditionnelles décrivant chacune des substitutions de bases sont toutes identiques, les transitions et transvections ayant toutes la même chance de se produire.
On note alors $\alpha/3 = \mathbb{P}_{(X_n=j)}(X_{n+1} = i)$ pour tout $i \neq j$ [1].

a) Que vaut, pour tout $j \in [\![0, 3]\!]$, $\mathbb{P}_{(X_n=j)}(X_{n+1} = j)$?

b) Déterminer la matrice M telle que :

$$X_{n+1} = MX_n \text{ où } X_n = \mathcal{M}_{\mathcal{B}}(p_n) \text{ avec}$$
$$p_n = (\mathbb{P}(X_n = 0), \mathbb{P}(X_n = 1), \mathbb{P}(X_n = 2), \mathbb{P}(X_n = 3)).$$

Vérifier que M est une matrice stochastique.

c) Montrer que $\mathrm{Sp}(M) = \{1, 1 - \frac{4}{3}\alpha\}$ et déterminer une base des espaces vectoriels propres de telle manière que chacun des vecteurs de base ait sa première coordonnée égale à 1.

d) En déduire l'expression de M^n. Quelle est la probabilité qu'un site initialement occupé par une base A dans la séquence ancestrale, soit à l'issue de 100 générations, occupé par une base T ? Que se passe-t-il lorsque n tend vers l'infini ?

② **Le modèle de Kimura** : Il suppose quant à lui que les taux de substitutions diffèrent selon qu'il s'agisse de transitions et de transvections. Justifier que la matrice stochastique associée à ce modèle est de la forme :

$$M = \begin{pmatrix} \star & \beta & \gamma & \gamma \\ \beta & \star & \gamma & \gamma \\ \gamma & \gamma & \star & \beta \\ \gamma & \gamma & \beta & \star \end{pmatrix}$$

Exprimer $\star$ en fonction de β et γ.

③ **Cas pratique :** Supposons que nous disposions d'une séquence d'ADN ancestrale de 40 bases :

S0='ACTTGTCGGATGATCAGCGGTCCATGCACCTGACAACGGT'

et de la séquence correspondante d'un descendant :

S1='ACATGTTGCTTGACGACAGGTCCATGCGCCTGAGAACGGC'

Il est possible de déterminer les fréquences conjointes des couples $(S0 = i, S1 = j)$ où $i, j \in \{A, G, C, T\}$. On obtient alors :

S_1 / S_0	A	G	C	T
A	7	0	1	1
G	1	9	2	0
C	0	2	7	2
T	1	0	1	6

1. Plusieurs chercheurs on donné des estimations de α. Il est autour de 1.1×10^{-9} mutations par site et par année pour certaines sections de l'ADN du chloroplaste de maïs ou de l'orge et autour de 10^{-8} mutations par sites et par an pour l'ADN mitochondrial des mammifères.

a) Si `n=len(S0)`. Que retourne l'expression Python :

```
np.sum([S0[k]=='G' and S1[k]=='C' for k in range(n)])
```

b) En déduire une fonction python `Tf=frequence(S0,S1)` qui retourne le tableau des fréquences conjointes ci-dessus.

c) Que retourne `colsum=np.dot(np.array([1,1,1,1]),Tf)` ? En déduire le moyen d'estimer les probabilités conditionnelles de substitution à partir des séquences $S0$ et $S1$.

d) Tracer le graphe de l'évolution dans le temps des vecteurs distributions de probabilité à partir de la séquence ancestrale. Déterminer l'état d'équilibre pour ce modèle (assurez-vous de donner un vecteur dont la somme des coordonnées vaut 1).

e) Écrire une fonction `nbMutations(S0,S1)` permettant d'obtenir le nombre de bases ayant muté entre les deux séquences.

④ **Distance phylogénétique** : Il s'agit d'utiliser le modèle de Jukes-Cantor pour comprendre comment estimer, à partir du nombre de mutations observées entre une séquence ancestrale et une séquence de sa descendance, le nombre de mutations ayant réellement eu lieu.

a) Interpréter les coefficients de la diagonale de M^n en terme de probabilités conditionnelles.

b) En déduire que la proportion de sites différant de leur base initiale est donnée par la formule :

$$p(t) = \frac{3}{4} - \frac{3}{4}\left(1 - \frac{4\alpha}{3}\right)^t$$

où t désigne le temps en générations. Tracer le graphe de p pour $\alpha = 0.01$.

c) Pourquoi peut-on toujours s'attendre à au moins 1/4 de bases communes entre deux séquences dont l'une descend de l'autre ?

d) Imaginons qu'on dispose de deux séquences dont l'une est l'ancêtre de l'autre. En quoi la connaissance du nombre de bases distinctes permet d'exprimer le temps t qui les sépare en fonction de α ? Justifier pourquoi, α étant extrêmement faible, il est possible d'écrire : $d = t\alpha \approx -\frac{3}{4}\ln(1 - \frac{4}{3}p)$

e) Donner un sens à la variable d appelée *distance de Jukes-Cantor*. La calculer dans le cas de l'exemple de la question 3.