1 DM mars 2020 - Application des réduction d'endomorphismes.

Correction - Modéliser l'évolution moléculaire

Le problème consiste à mettre en place un modèle mathématique permettant d'estimer le nombre de mutations qui ont pu avoir lieu au cours de la descendance d'une séquence d'ADN.
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 $j \in [\![0, 3]\!]$, $\mathbb{P}_{(X_n=j)}(X_{n+1} = j)$?*

On utilise le fait que $\{(X_{n+1} = i), 0 \leq i \leq 3\}$ est un système complet d'événements.
d'après les propriétés des probabilités conditionnelles, on a donc, pour tout $j \in [\![0, 3]\!]$:

$$\sum_{i=0}^{3} \mathbb{P}_{(X_n=j)}(X_{n+1} = i) = \mathbb{P}_{(X_n=j)}(\Omega) = 1$$

Dès lors :

$$\boxed{\mathbb{P}_{(X_n=j)}(X_{n+1} = j)} = 1 - \sum_{i=0, i\neq j}^{3} \mathbb{P}_{(X_n=j)}(X_{n+1} = i) = 1 - 3\frac{\alpha}{3} = \boxed{1 - \alpha}$$

b. *Déterminons la matrice M telle que $X_{n+1} = MX_n$ où $X_n = \mathcal{M}_{\mathcal{B}}(p_n)$*

On utilise la formule des probabilités totales et pour ça on met en évidence le système complet d'événements : $\{(X_n = j), 0 \leq j \leq 3\}$. Alors :

$$\mathbb{P}(X_{n+1} = i) = \sum_{j=0}^{3} \mathbb{P}(X_{n+1} = i, X_n = j) = \sum_{j=0}^{3} \mathbb{P}_{(X_n=j)}(X_{n+1} = i)\mathbb{P}(X_n = j)$$

Pour $i = 0$, on obtient :

$$\begin{aligned}\mathbb{P}(X_{n+1} = 0) &= \sum_{j=0}^{3} \mathbb{P}_{(X_n=j)}(X_{n+1} = 0)\mathbb{P}(X_n = j) = \sum_{j=0}^{3} m_{0,j}\mathbb{P}(X_n = j) \\ &= (1 - \alpha)\mathbb{P}(X_n = 0) + \frac{\alpha}{3}\mathbb{P}(X_n = 1) + \frac{\alpha}{3}\mathbb{P}(X_n = 2) + \frac{\alpha}{3}\mathbb{P}(X_n = 3)\end{aligned}$$

De même, pour $i = 1$:

$$\begin{aligned}\mathbb{P}(X_{n+1} = 1) &= \sum_{j=0}^{3} \mathbb{P}_{(X_n=j)}(X_{n+1} = 1)\mathbb{P}(X_n = j) = \sum_{j=0}^{3} m_{1,j}\mathbb{P}(X_n = j) \\ &= \frac{\alpha}{3}\mathbb{P}(X_n = 0) + (1 - \alpha)\mathbb{P}(X_n = 1) + \frac{\alpha}{3}\mathbb{P}(X_n = 2) + \frac{\alpha}{3}\mathbb{P}(X_n = 3)\end{aligned}$$

De façon identique pour $\mathbb{P}(X_{n+1} = 2)$ et $\mathbb{P}(X_{n+1} = 3)$, on peut conclure :

$$X_{n+1} = MX_n \text{ avec } M = \begin{pmatrix} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{pmatrix}$$

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.

✐ *Remarque :* Dire que M est une matrice stochastique est immédiat puisque, par définition de ces matrices, leurs termes sont positifs et la somme de chaque colonne vaut 1.

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

– **Méthode 1 :** On nous a donné les valeurs propres. On peut se contenter de montrer que $\text{rg}(M - I_4) < 4$ et $\text{rg}(M - (1 - 4\alpha/3)I_4) < 4$, ce qui est très rapide car :

$$M - (1 - \tfrac{4\alpha}{3})I_4 = \frac{\alpha}{3}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

D'où $\text{rg}(M - (1 - \frac{4\alpha}{3})I_4) = 1 \Rightarrow \lambda = 1 - \frac{4\alpha}{3}$ est valeur propre de M et $\dim(E_{1-4\alpha/3}) = 4 - 1 = 3$.
Il est par ailleurs immédiat que $\boxed{E_{1-\frac{4\alpha}{3}} = \text{Vect}\{(1,-1,0,0),(1,0,-1,0),(1,0,0,-1)\}}$

De même,

$$M - I_4 = \frac{\alpha}{3}\begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}$$

On a alors $\text{rg}(M - I_4) \neq 4$ car on note que la somme des colonnes fait 1. Donc 1 est valeur propre de M.
✐ Il ne peut y avoir d'autre valeur propre car on sait que la somme des dimensions des espaces propres est inférieure ou égale à 4 et on rappelle que $\dim(E_{1-4\alpha/3}) = 3$. Et comme la dimension de E_1 vaut au moins 1, on peut assurer que $\boxed{\dim(E_1) = 1}$.
Il suffit de trouver 1 vecteur de base de E_1, ce qui est immédiat si on rappelle que la somme des colonnes vaut 1, soit

$$\boxed{E_1 = \text{Vect}\{(1,1,1,1)\}}$$

– **Méthode 2 :** (La méthode « classique »)

$\lambda \in \text{Sp}(M) \Leftrightarrow \exists X \neq 0 / (M - \lambda I_4)X = 0 \Leftrightarrow M - \lambda I_4$ non inversible $\Leftrightarrow \text{rg}(M - \lambda I_4) < 4$

$$\Leftrightarrow \text{rg}\begin{pmatrix} 1-\alpha-\lambda & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha-\lambda & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha-\lambda & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha-\lambda \end{pmatrix} < 4$$

$$\Leftrightarrow \text{rg}\begin{pmatrix} 1-\lambda & 1-\lambda & 1-\lambda & 1-\lambda \\ \alpha/3 & 1-\alpha-\lambda & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha-\lambda & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha-\lambda \end{pmatrix} < 4 \begin{matrix} L_1 \leftarrow L_1 + L_2 + L_3 + L_4 \\ L_2 \\ L_3 \\ L_4 \end{matrix}$$

$$\Leftrightarrow \text{rg}\begin{pmatrix} 1-\lambda & 0 & 0 & 0 \\ \alpha/3 & (1-4\alpha/3)-\lambda & 0 & 0 \\ \alpha/3 & 0 & (1-4\alpha/3)-\lambda & 0 \\ \alpha/3 & 0 & 0 & (1-4\alpha/3)-\lambda \end{pmatrix} < 4 \begin{matrix} C_2 \leftarrow C_2 - C_1 \\ C_3 \leftarrow C_3 - C_1 \\ C_4 \leftarrow C_4 - C_1 \end{matrix}$$

$\lambda \in \text{Sp}(M) \Leftrightarrow \lambda = 1$ ou $(1 - 4\alpha/3) - \lambda = 0$

Conclusion : $\boxed{\text{Sp}(M) = \{1, 1 - \frac{4}{3}\alpha\}}$

Déterminons les espaces propres associés :

Commençons par $\lambda_2 = 1 - \frac{4}{3}\alpha$: $X \in E_{\lambda_2} \Leftrightarrow (M - \lambda_2 I_2)X = 0 \Leftrightarrow U_{\lambda_2}X = 0$ où U_λ est la matrice obtenu précédemment par réduction de gauss, juste avant l'opération sur les colonnes. Alors :

$$X \in E_{\lambda_2} \Leftrightarrow \frac{\alpha}{3}\begin{pmatrix}4&4&4&4\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}\cdot\begin{pmatrix}x\\y\\z\\t\end{pmatrix}=0$$

On peut noter que U_{λ_2} est de rang 1 donc, d'après la formule du rang : $\dim(E_{\lambda_2}) = 4-1 = 3$.
Il est par ailleurs immédiat que :

$$X \in E_{\lambda_2} \Leftrightarrow x+y+z+t=0$$

Conclusion : $\boxed{E_{1-\frac{4}{3}\alpha}(M) = \{(x,y,z,t)\in\mathbb{R}^4 / x=-y-z-t\} = \text{Vect}\left\{\begin{pmatrix}1\\-1\\0\\0\end{pmatrix},\begin{pmatrix}1\\0\\-1\\0\end{pmatrix},\begin{pmatrix}1\\0\\0\\-1\end{pmatrix}\right\}}$

Comme la somme des dimensions des espaces propres d'une matrice de $\mathcal{M}_4(\mathbb{R})$ ne peut pas dépasser 4 et que $\dim(E_1) \geq 1$, on en déduit que $\dim(E_1) = 1$.
Comme par ailleurs $(1,1,1,1)^T$ est une solution évidente de $(M-I_4)X = 0$ (la somme de chaque ligne fait 1), on a immédiatement :

$$\boxed{E_1(M) = \text{Vect}\left\{\begin{pmatrix}1\\1\\1\\1\end{pmatrix}\right\}}$$

✎*Remarque :* Si on note que M est stochastique et que M et ${}^tM$ ont même valeur propre, alors il est possible de montrer sans calcul que 1 est valeur propre de M.

d. *Déduisons-en l'expression de M^n :*

On montre par une récurrence de cours que : $M^n = PD^nP^{-1}$ où $D = \begin{pmatrix}1&0&0&0\\0&1-\frac{4\alpha}{3}&0&0\\0&0&1-\frac{4\alpha}{3}&0\\0&0&0&1-\frac{4\alpha}{3}\end{pmatrix}$.

Par ailleurs, $P = \begin{pmatrix}1&1&1&1\\1&-1&0&0\\1&0&-1&0\\1&0&0&-1\end{pmatrix}$

Le calcul de P^{-1} est très rapide si on utilise l'équivalence : $X = PX' \Leftrightarrow P^{-1}X = X'$.
En effet :

$$PX' = X \Leftrightarrow \begin{cases}x+y+z+t &= x'\\ x-y &= y'\\ x-z &= z'\\ x-t &= t'\end{cases} \Leftrightarrow \begin{cases}x+(x-y')+(x-z')+(x-t') &= x'\\ y &= x-y'\\ z &= x-z'\\ t &= x-t'\end{cases}$$

Soit

$$P^{-1} = \frac{1}{4}\begin{pmatrix}1&1&1&1\\1&-3&1&1\\1&1&-3&1\\1&1&1&-3\end{pmatrix}$$

Dès lors, après calculs :

$$M^n = PD^nP^{-1} = \begin{pmatrix}1-a&b&b&b\\b&1-a&b&b\\b&b&1-a&b\\b&b&b&1-a\end{pmatrix} \text{ où } a = \frac{3}{4}-\frac{3}{4}\left(1-\frac{4\alpha}{3}\right)^n \text{ et } b = \frac{1}{4}-\frac{1}{4}\left(1-\frac{4\alpha}{3}\right)^n$$

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 ?*

On rappelle que $X_{n+1} = MX_n$ pour tout $n \in \mathbb{N}$.
Donc, par récurrence, $X_n = M^n X_0$, $\forall n \in \mathbb{N}$.

La matrice M^n étant une matrice stochastique, ses coefficients peuvent être lus à la lumière de la formule des probabilités totales. En particulier la ligne 4 de ce produit donne :

$$\mathbb{P}(X_{100} = 3) = \mathbb{P}_{(X_0=0)}(X_{100} = 3)\mathbb{P}(X_0 = 0) + \mathbb{P}_{(X_0=1)}(X_{100} = 3)\mathbb{P}(X_0 = 1) + \mathbb{P}_{(X_0=2)}(X_{100} = 3)\mathbb{P}(X_0 = 2) + \mathbb{P}_{(X_0=3)}(X_{100} = 3)\mathbb{P}(X_0 = 3)$$

Le coefficient $m_{3,0}$ de M^n permet donc de conclure : $\boxed{\mathbb{P}_{(X_0=0)}(X_{100} = 3) = b = \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4\alpha}{3}\right)^{100}}$

Que se passe-t-il lorsque n tend vers l'infini ? On note que $0 < \frac{4\alpha}{3} << 1$. Dès lors :

$\lim_{n\to\infty} \mathbb{P}_{(X_0=0)}(X_n = 3) = \frac{1}{4}$

C'est le cas de toutes les autres bases. Quelle que soit la répartition initiale, les chances de trouver l'un ou l'autre des nucléotides deviennent uniformes.

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. *Justifions que la matrice stochastique associée à ce modèle est de la forme :*

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

Rappelons que, comme dans le modèle de Jukes-Cantor :

$$\mathbb{P}(X_{n+1}=i) = \sum_{j=0}^{3} \mathbb{P}(X_{n+1}=i, X_n=j) = \sum_{j=0}^{3} \mathbb{P}_{(X_n=j)}(X_{n+1}=i)\mathbb{P}(X_n=j) = \sum_{j=0}^{3} m_{i,j}\mathbb{P}(X_n=j)$$

Supposons que les taux de mutation soit de β pour les transitions (A-G ou C-T) et de γ pour les transvections (A-C, A-T, C-G ou G-T), alors :

$$\mathbb{P}_{(X_n=j)}(X_n=i) = \begin{cases} \beta \text{ si } (i,j) \in \{(0,1),(1,0),(2,3),(3,2)\} \\ \gamma \text{ si } (i,j) \in \{(0,2),(2,0),(0,3),(3,0),(1,2),(2,1),(1,3),(3,1)\} \end{cases}$$

Si nous prenons garde à ce que les distributions p_n soient données dans l'ordre *adénine*, *guanine*, *cytosine* et *thymine*, alors la matrice de transition est bien celle indiquée.

*Exprimons $\star$ en fonction de β et γ :*La matrice de Kimura est une matrice stochastique puisque la somme de ses colonnes vérifie

$$\sum_{i=0}^{3} \mathbb{P}_{(X_n=j)}(X_{n+1}=i) = \mathbb{P}_{(X_n=j)}(\Omega) = 1 \text{ pour tout } 0 \leq j \leq 3$$

Donc $\star+\beta+2\gamma=1$ soit $\boxed{\star = 1-\beta-2\gamma}$

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

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

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

retourne le nombre de nucléotide qui valaient `'G'` dans la séquence ancestrale et ont muté en `'C'` dans la séquence S_1 (transvection). A la lecture du tableau, cette expression retourne 2.

b. *Déduisons-en une fonction python `Tf=frequence(S0,S1)` qui retourne le tableau des fréquences conjointes ci-dessus :*
L'idée est de créer un tableau des bases possibles `base=['A','G','C','T']` de telle façon que `base[0]='A'`, `base[1]='G'`, etc. et de l'utiliser pour dénombrer les cas où `S0[k]=base[i]` **et** `S1[k]=base[j]` pour i variant de 0 à 3 et j variant de 0 à 3.

```
def frequences(S0,S1):
    Tf=np.zeros((4,4))
       n=len(S0)
       base=['A','G','C','T'];
       for j in range(4):
          for i in range(4):
             Tf[i,j]=np.sum([S0[k]==base[j] and S1[k]==base[i] for k in range(n)])
    return Tf
```

c. *Que retourne* `colsum=np.dot(np.array([1,1,1,1]),Tf)` *?*

Il s'agit d'un produit matriciel.
Soit $U = \begin{pmatrix}1 & 1 & 1 & 1\end{pmatrix}$, alors colsum est la matrice **ligne** $C = U \cdot Tf$.

$$C_j = \sum_{k=1}^{4} U_k Tf_{k,j} = \sum_{k=1}^{4} Tf_{k,j}$$

Il s'agit bien de la somme de la colonne j du tableau des fréquences conjointes.
Autrement dit, `colsum` est la matrice ligne des fréquences marginales de A, G, C et T dans la séquence ancestrale `S0`. Soit `colsum = [9,11,11,9]`.

Conséquence : Estimons les probabilités conditionnelles de substitution à partir des séquences $S0$ et $S1$:
Par exemple, pour la probabilité conditionnelle qu'une place occupée par de l'adényne sur la séquence `S0` le soit encore sur la séquence `S1` :

$$\mathbb{P}_{(X_0=0)}(X_1 = 0) = \frac{\mathbb{P}(X_0 = 0, X_1 = 0)}{\mathbb{P}(X_0 = 0)}$$

Or ces probabilités peuvent être calculées d'un point de vue fréquentiste en divisant le nombre de cas favorable par le nombre de cas possible. A savoir :

$$\mathbb{P}(X_0 = 0, X_1 = 0) \approx \frac{f_{S0='A',S1='A'}}{40} = \frac{Tf_{0,0}}{40} \text{ et } \mathbb{P}(X_0 = 0) \approx \frac{f_{S0='A'}}{40} = \frac{colsum_0}{40}$$

Dès lors :

$$\mathbb{P}_{(X_0=0)}(X_1 = 0) = \frac{Tf_{0,0}}{colsum_0} = \frac{7}{9}$$

De même, $\mathbb{P}_{(X_0=0)}(X_1 = 1) = \dfrac{\mathbb{P}(X_0 = 0, X_1 = 1)}{\mathbb{P}(X_0 = 0)} \approx \dfrac{f_{S0='A',S1='G'}}{f_{S0='A'}} = \dfrac{Tf_{1,0}}{colsum_0} = \dfrac{1}{9}$
Etc.
Il suffit donc de diviser les valeurs du tableau des fréquences conjointes par les fréquences marginales obtenues au début de cette question.

Conclusion : Les probabilités conditionnelles $\mathbb{P}_{(X0=j)}(X_1 = i) = m_{i,j}$ s'estiment par $\dfrac{Tf_{i,j}}{colsum_j}$

Sous Python, c'est très facile à obtenir. Il suffit de taper `ProbaCond=Tf/colsum`

Concrètement, avec les séquences `S0` et `S1` ci dessus, on obtient pour les probabilités conditionnelles :

$\mathbb{P}_{(X_0=j)}(X_1 = i)$	0	1	2	3
0	7/9	0	1/11	1/9
1	1/9	9/11	2/11	0
2	0	2/11	7/11	2/9
3	1/9	0	1/11	6/9

=

$\mathbb{P}_{(X_0=j)}(X_1 = i)$	0	1	2	3
0	0.778	0	0.091	0.111
1	0.111	0.818	0.182	0
2	0	0.182	0.636	0.222
3	0.111	0	0.091	0.667

✐ Ce dernier tableau est une estimation de la matrice M de transition et peut-être utilisé comme tel dans la suite !

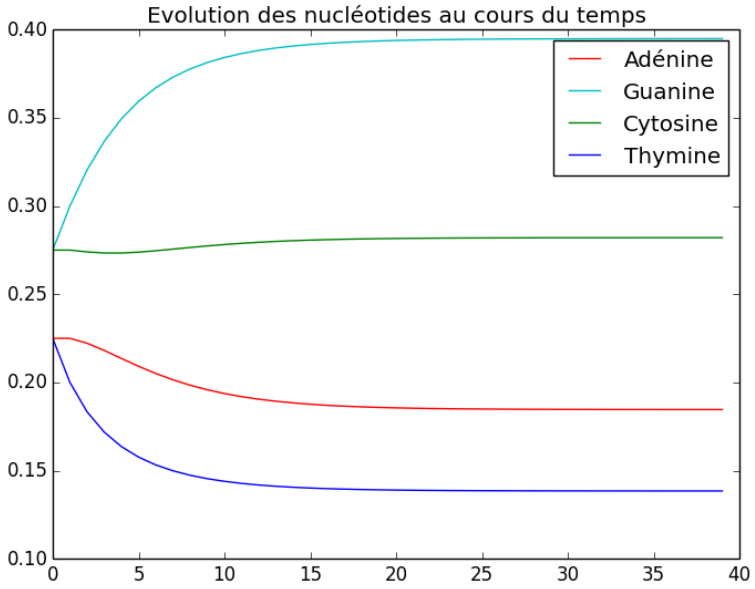
d. *Traçons le graphe de l'évolution dans le temps des vecteurs distributions de probabilité à partir de la séquence ancestrale :*
Il suffit pour ça de connaître la distribution ancestrale. Or, la séquence ancestrale nous permet d'obtenir que :

$$\mathbb{P}(X_0 = 0) = 9/40 = 0.225,\ \mathbb{P}(X_0 = 1) = 11/40 = 0.275,\ \mathbb{P}(X_0 = 2) = 11/40 = 0.275,$$
$$\mathbb{P}(X_0 = 3) = 9/40 = 0.225$$

Il suffit alors d'écrire une fonction qui calcule $X_{n+1} = M \cdot X_n$, $\forall n \in \mathbb{N}$.
A savoir :

```
def evolution(M,duree): # M est la matrice de transition obtenue à la question 3.c)
    X=np.array([[0.225],[0.275],[0.275],[0.225]])#matrice colonne
    A=[0.225];G=[0.275] # purines
    C=[0.275];T=[0.225] # pyrimidines
    for k in range(1,duree):
       Y=np.dot(M,X)
       A.append(Y[0]);G.append(Y[1])
       C.append(Y[2]);T.append(Y[3])
       X=Y
    return A,G,C,T
```

On obtient alors le graphe suivant :



Déterminons l'état d'équilibre pour ce modèle : Autrement dit, déterminons $\lim\limits_{n\to\infty} X_n$...
La matrice M étant une matrice stochastique, nous savons que 1 est valeur propre (car M et sa transposée ayant même valeurs propres, il est immédiat qu'en multipliant ${}^tM$ par le vecteur colonne formé uniquement de 1, on obtient ce même vecteur puisque la somme de chaque ligne vaut 1...). Pour autant, les autres valeurs propres sont difficiles à obtenir et les coefficients de la matrice sont suffisamment pénibles à manipuler pour qu'on fasse appel à Python pour calculer à notre place les valeurs propres et les espaces propres associés...

En tapant : `D,P=np.linalg.eig(M)` on obtient directement le spectre noté `D` et une famille `P` de vecteurs propres associés que nous appellerons (Y_0, Y_1, Y_2, Y_3). Or

$$\mathrm{Sp}(M) = \{1., 0.46901547, 0.63566181, 0.79431262\}$$

donc la matrice M est diagonalisable puisqu'elle possède quatre valeurs propres distinctes alors qu'elle est d'ordre 4. Dès lors, la famille (Y_0, Y_1, Y_2, Y_3) est une base de vecteur propre que nous noterons $\mathcal{B}'$ et P est

la matrice de passage de la base canonique dans cette base $\mathcal{B}'$.

Nous savons que $X_n = M^n X_0$ avec $X_0 = \begin{pmatrix} 0.225 \\ 0.275 \\ 0.275 \\ 0.225 \end{pmatrix}$.

Il est possible d'exprimer X_0 dans la base $\mathcal{B}'$ grâce aux formules de changement de bases. A savoir :

$$X_0 = P \cdot X_0' \Leftrightarrow X_0' = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = P^{-1} X_0$$

Sous Python : `Xprim0=np.dot(np.linalg.inv(P),X0)` permet d'écrire :

$$X_0 = c_0 Y_0 + c_1 Y_1 + c_2 Y_2 + c_3 Y_3 = -0.0537 Y_0 + 0.033 Y_1 + 0.099 Y_2 + 0.158 Y_3$$

Dès lors, par linéarité :

$$X_n = c_0 M^n Y_0 + c_1 M^n Y_1 + c_2 M^n Y_2 + c_3 M^n Y_3 \text{ avec } Y_k \in E_{\lambda_k}(M) \Rightarrow M^n Y_k = \lambda_k^n Y_k,\ \forall n \in \mathbb{N}.$$

où on a posé $\lambda_0 = 1$, $\lambda_1 = 0.47$, $\lambda_2 = 0.64$ et $\lambda_3 = 0.79$.

Par passage à la limite, toutes les valeurs propres à l'exception de la première étant strictement comprises entre 0 et 1, on obtient :

$$\boxed{\lim_{n\to\infty} X_n = c_0 Y_0 = -0.0537 \cdot Y_0 = \big(0.18, 0.39, 0.28, 0.14\big)}$$

On vient d'obtenir qu'à l'issue d'un nombre de générations suffisamment grand, la proportion d'adényne tend vers 18%, de guanine vers 39%, de cytosine vers 28% et de thymine vers 14%.

e. *Écrivons une fonction `nbMutations(S0,S1)` permettant d'obtenir le nombre de bases ayant muté entre les deux séquences :*
Il suffit pour ça d'exploiter le tableau des fréquences conjointes. En effet, tous les effectifs qui sont sur sa diagonales sont ceux des bases qui n'ont pas muté. Il suffit de connaître le nombre total de bases (`=len(S0)`) pour en déduire le nombre de mutations. Soit :

```
def nbMutations(Tf):
    return len(S0)-np.sum([Tf[i,i] for i in range(4)])
```

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. *Interprétons les coefficients de la diagonale de M^n en terme de probabilités conditionnelles :* Sachant que $X_n = M^n X_0$, on a par construction de la matrice M^n (dont les coefficients seront notés $m_{i,j}^n$) :

$$\begin{aligned} \mathbb{P}(X_n = i) &= \sum_{j=0}^{3} m_{i,j}^n \mathbb{P}(X_0 = j) \text{ expression du produit matriciel} \\ &= \sum_{j=0}^{3} \mathbb{P}_{(X_0=j)}(X_n = i)\mathbb{P}(X_0 = j) \text{ d'après la F.P.T.} \end{aligned}$$

Soit, par identification :
$\forall 0 \leq i, j \leq 3$, $m_{i,j}^n = \mathbb{P}_{(X_0=j)}(X_n = i)$ et donc $\boxed{m_{i,i}^n = \mathbb{P}_{(X_0=i)}(X_n = i)}$.

Conclusion : Les termes de la diagonale de M^n donnent la probabilité conditionnelle que la base considérée au temps $t = n$ soit la même qu'au temps $t = 0$.

b. *Déduisons-en la proportion de sites différant de leur base initiale :*
Il fallait comprendre : On cherche, pour chaque site, la probabilité que sa distribution au temps $t = n$ diffère de sa distribution ancestrale. Il s'agit donc d'estimer la probabilité d'au moins une mutation entre la séquence `S0` et la séquence `S1`, autrement dit $\mathbb{P}(X_n \neq X_0)$.

Pour faire ce calcul il est possible de passer par l'événement complémentaire, à savoir $(X_n = X_0)$.
Or, d'après la question précédente :

$$\begin{aligned}\mathbb{P}(X_0 = X_n) &= \sum_{i=0}^{3} \mathbb{P}(X_0 = i, X_n = i) = \sum_{i=0}^{3} \mathbb{P}_{(X_0=i)}(X_n = i)\mathbb{P}(X_0 = i) \\ &= \sum_{i=0}^{3} m_{i,i}^n \mathbb{P}(X_0 = i) = \sum_{i=0}^{3} (1-a)\mathbb{P}(X_0 = i) = (1-a)\sum_{i=0}^{3} \mathbb{P}(X_0 = i) \text{ d'après 1.d} \\ &= 1 - a = \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4\alpha}{3}\right)^n \quad \text{car } \{(X_0 = i), 0 \leq i \leq 3\} \text{ est un syst. complet d'év.}\end{aligned}$$

Conclusion : $\boxed{\mathbb{P}(X_n \neq X_0) = 1 - \mathbb{P}(X_0 = X_n) = 1 - (1-a) = a = \frac{3}{4} - \frac{3}{4}\left(1 - \frac{4\alpha}{3}\right)^n}$

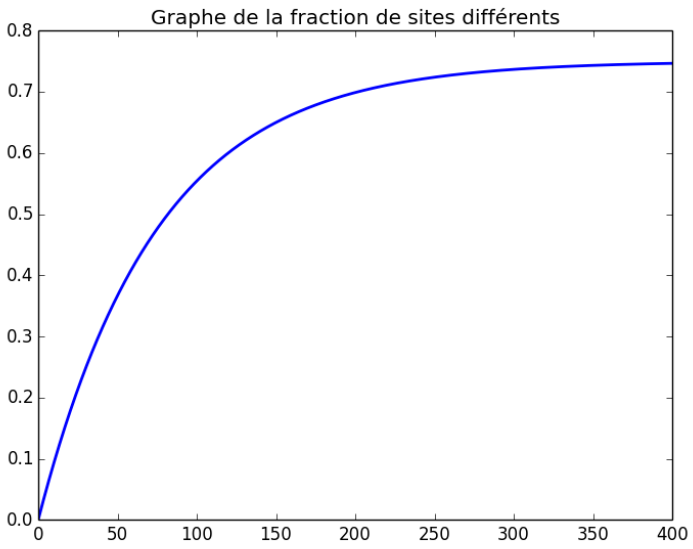
✎ *Remarque :* Si t désigne le temps en générations, on a bien :

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

Traçons le graphe de $p(t) = \frac{3}{4} - \frac{3}{4}\left(1 - \frac{4 \cdot 10^{-2}}{3}\right)^t$ *pour* $\alpha = 0.01$ *:*

Cette fonction est dérivable de dérivée égale à : $p'(t) = \alpha t \left(1 - \frac{4\alpha}{3}\right)^{t-1}$

Elle est donc continue et strictement croissante sur $\mathbb{R}_+$ et sa limite en l'infini vaut $\frac{3}{4}$.



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 ?* La réponse découle du théorème de la limite monotone et du graphe précédent...
Pour tout $t \geq 0$, $0 \leq p(t) \leq \frac{3}{4}$ et sa limite en l'infini vaut $\frac{3}{4}$.
Dès lors, la fraction de sites qui ont muté entre deux séquences, aussi éloignées soient-elles dans le temps, ne pourra excéder $\frac{3}{4}$.
Finalement, même si le nombre de mutations entre deux séquences est tel que rien ne semble plus les relier, on peut s'attendre à ce qu'au moins un quart des sites coïncident.

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 ?
A partir des deux séquences d'ADN il est possible d'estimer $p = p(t)$ comme nous l'avons fait en 3.e).
Dès lors :

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

Par ailleurs, si α est proche de zéro (ce qui est le cas d'après les données rappelées en note de bas de page 1) alors :

$$\ln\left(1 - \frac{4\alpha}{3}\right) \approx -\frac{4\alpha}{3}$$

D'où

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

Conclusion : $\boxed{d = t\alpha \approx -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)}$

e. *Donnons un sens à la variable d appelée distance de Jukes-Cantor :*
On doit noter que notre choix du pas de temps au sein du modèle a un impact à la fois sur le taux de mutation α et le nombre d'intervalles de temps qui séparent l'ancêtre de ses descendants. L'idée de la distance de Jukes-Cantor est de fournir une donnée moins dépendantes de ces paramètres en considérant un pas de temps suffisamment petit pour que α reste proche de zéro.
En effet :

$$\begin{aligned} d &= t\alpha \\ &= (\text{nombre de pas de temps}) \cdot (\text{taux de mutation}) \\ &= (\text{nombre de pas de temps})(\text{nombre de substitutions par site/pas de temps}) \\ &= \text{nombre estimé de substitutions par site durant l'intervalle de temps étudié} \end{aligned}$$

✐ Il faut noter que ce nombre estimé de substitutions inclus y-compris celle que nous ne pouvons pas observer parce que dissimulées par des substitutions successives.
La notion de distance est ici une notion abstraite qui indique à quel point deux séquences sont différentes au regard du nombre de mutations. Si on fait l'hypothèse d'une horloge moléculaire, la distance qui est ici calculée est proportionnelle au temps écoulé, la constante de proportionnalité étant le taux de mutation. Dès lors, la distance de Jukes-Cantor peut être vue comme une mesure du temps requis par une séquence pour muter en une autre.
Si il existe d'autres données (comme des données géologiques) suggérant le temps écoulé entre les deux séquences, alors le taux de mutation peut-être déduit de cette distance.

Dans l'exemple de 40 bases étudié en 3., d'après le tableau des fréquences conjointes, on a obtenu à la question e) que 11 sites avaient subi une substitution et donc que $p = 11/40 = 0.275$.
Dès lors :

$$d_{JC}(S_0, S_1) = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{11}{40}\right) \approx 0.3426$$

En conclusion, si nous avons observé une moyenne de 0.275 substitutions par sites, nous estimons qu'au cours de son évolution 0.3426 substitutions par sites ont effectivement eu lieu...