Modèles phylogénétiques stochastiques

Aimé Lachal

Avril 2001

Introduction

L'ADN est une macromolécule très longue constituée d'un grand nombre de désoxyribonucléotides, eux-mêmes composés d'une base, d'un ose et d'un groupe phosphate. La séquence précise des bases recèle l'information génétique, tandis que les oses et les groupes phosphates jouent un rôle structural. Les bases d'un ADN sont des dérivés de la purine et de la pyrimidine. Les purines des ADN sont l'adénine (A) et la guanine (G) et les pyrimidines sont la cytosine (C) et la thymine (T). Un ADN est donc caractérisé par la séquence de ses bases A, C, G, T. Lors de son évolution au cours du temps, l'ADN subit des mutations : une ou plusieurs bases peuvent se substituer à d'autres. Plusieurs modèles stochastiques ont été introduits pour décrire le processus de mutation d'une base occupant un certain site fixé sur l'ADN. On définit X_t le type de base présent au site considéré à l'instant t. $(X_t)_{t\geqslant 0}$ est un processus stochastique à valeurs dans l'alphabet $\mathcal{A} = \{A, C, G, T\}$.

Les modèles présentés ci-dessous sont des modèles pour lesquels $(X_t)_{t\geqslant 0}$ est une chaîne de Markov homogène. Cette chaîne est caractérisée par la donnée des probabilités suivantes, pour $a, b \in \mathcal{A}$ et $0 \leqslant s \leqslant t$,

$$\mathbb{P}(X_0 = a) = \pi_a(0),$$

$$\mathbb{P}(X_t = b \mid X_s = a \text{ et } (X_u)_{0\leqslant u\leqslant s}) = p_{ab}(t-s).$$

On introduit alors les vecteurs-ligne $\pi(0) = (\pi_a(0))_{a\in\mathcal{A}}$ (loi initiale) et $\pi(t) = (\mathbb{P}(X_t = a))_{a\in\mathcal{A}}$ (loi de X_t), ainsi que la matrice $P(t) = (p_{ab}(t))_{(a,b)\in\mathcal{A}\times\mathcal{A}}$. On a

$$\pi(t) = \pi(0)P(t).$$

La famille de matrices $(P(t))_{t\geqslant 0}$ constitue un semi-groupe. En effet,

$$\begin{aligned} p_{ab}(s+t) &= \mathbb{P}(X_{s+t} = b \mid X_0 = a) \\ &= \sum_{c\in\mathcal{A}} \mathbb{P}(X_s = c \mid X_0 = a)\mathbb{P}(X_{s+t} = b \mid X_s = c) \\ &= \sum_{c\in\mathcal{A}} p_{ac}(s)p_{cb}(t) \end{aligned}$$

(égalité de Chapman-Kolmogorov), soit :

$$P(s+t) = P(s)P(t).$$

On introduit alors le générateur infinitésimal de ce semi-groupe :

$$\boldsymbol{A} = \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}[P(\varepsilon) - I] = P'(0).$$

On a

$$\frac{1}{\varepsilon}[P(t+\varepsilon) - P(t)] = P(t)\,\frac{1}{\varepsilon}[P(\varepsilon) - I] = \frac{1}{\varepsilon}[P(\varepsilon) - I]P(t),$$

relation de laquelle on déduit les équations différentielles

$$\boxed{P'(t) = \boldsymbol{A}P(t) \quad \text{et} \quad P'(t) = P(t)\boldsymbol{A}.}$$

Ce sont les équations progressive et rétrograde de Kolmogorov (forward and backward equations) qui, complétées de la condition initiale $P(0) = I$, se résolvent facilement lorsque l'espace d'états est fini. Leur solution commune est

$$\boxed{P(t) = \exp(t\boldsymbol{A}).}$$

Signalons que la matrice $\boldsymbol{A} = (A_{ab})_{(a,b)\in\mathcal{A}\times\mathcal{A}}$ contient des informations sur les sauts de la chaîne $(X_t)_{t\geqslant 0}$. On a d'abord $\forall (a,b) \in \mathcal{A}\times\mathcal{A}, A_{ab} \geqslant 0$ et $\sum_{(a,b)\in\mathcal{A}\times\mathcal{A}} A_{ab} = 0$. Notons ensuite $T_1 = \inf\{t \geqslant 0 : X_t \neq X_0\}$ l'instant du premier saut et $\lambda_a = -A_{aa} \geqslant 0$. Alors la v.a. $(T_1 \mid X_0 = a)$ suit la loi de Poisson $\mathcal{P}(\lambda_a)$ *i.e.* sa densité est $\mathbb{P}(T_1 \in dt \mid X_0 = a)/dt = \lambda_a e^{-\lambda_a t}$ et la loi de probabilité de la v.a. $(X_{T_1} \mid X_0 = a)$ est donnée par $\mathbb{P}(X_{T_1} = b \mid X_0 = a) = \dfrac{A_{ab}}{\lambda_a}$.

1 Modèle à un paramètre (T.H. Jukes & C. Cantor, 1969)

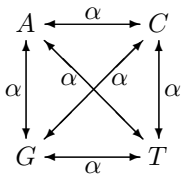
On suppose que les substitutions entre les bases se produisent au hasard de manière équiprobable. Ainsi, pendant un laps de temps $[t, t+\varepsilon]$ avec $\varepsilon \to 0^+$, pour tout $a \in \mathcal{A}$ et $b \in \mathcal{A}\setminus\{a\}$,

$$\mathbb{P}(X_{t+\varepsilon} = b \mid X_t = a) = \alpha\varepsilon + o(\varepsilon),$$

et donc

$$\mathbb{P}(X_{t+\varepsilon} = a \mid X_t = a) = 1 - 3\alpha\varepsilon + o(\varepsilon)$$

(voir Figure 1). Le paramètre 3α représente le taux de substitution d'une base.



	A	G	C	T
A	-3α	α	α	α
G	α	-3α	α	α
C	α	α	-3α	α
T	α	α	α	-3α

FIGURE 1 – Modèle à un paramètre

1.1 Détermination du semi-groupe

On va déterminer la probabilité $p_{ab}(t) = \mathbb{P}(X_t = b \mid X_0 = a)$ pour toutes bases $a, b \in \mathcal{A}$. Les données s'écrivent sous forme matricielle

$$P(\varepsilon) = I - \boldsymbol{A}\varepsilon + o(\varepsilon)$$

où $\boldsymbol{A}$ est le générateur infinitésimal de la chaîne de Markov $(X_t)_{t\geqslant 0}$:

$$\boxed{\boldsymbol{A} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}.}$$

Calculons explicitement $\exp(t\boldsymbol{A})$. Écrivons $\boldsymbol{A} = 4\alpha(J - I)$ avec $J = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$. On a $J^2 = J$ et alors

$$\exp(4\alpha t J) = \sum_{n=0}^{\infty} \frac{(4\alpha t)^n}{n\,!}\, J^n = I + (e^{4\alpha t} - 1)J$$

puis

$$P(t) = e^{-4\alpha t}\exp(4\alpha t J) = e^{-4\alpha t} I + (1 - e^{-4\alpha t})J.$$

Ainsi $P(t)$ est de la forme

$$\boxed{P(t) = \begin{pmatrix} a(t) & b(t) & b(t) & b(t) \\ b(t) & a(t) & b(t) & b(t) \\ b(t) & b(t) & a(t) & b(t) \\ b(t) & b(t) & b(t) & a(t) \end{pmatrix}}$$

où $a(t)$ et $b(t)$ sont respectivement associés aux probabilités de conservation et de substitution d'une base à l'instant t selon

$$\boxed{\begin{aligned} a(t) &= \mathbb{P}(X_t = X_0) = \tfrac{1}{4}\,[1 + 3e^{-4\alpha t}], \\ b(t) &= \tfrac{1}{3}\,\mathbb{P}(X_t \neq X_0) = \tfrac{1}{4}\,[1 - e^{-4\alpha t}]. \end{aligned}}$$

On voit que

$$\boxed{P(t) \underset{t\to+\infty}{\longrightarrow} \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}}$$

prouvant ainsi que la chaîne de Markov $(X_t)_{t\geqslant 0}$ est ergodique et que sa loi stationnaire (état d'équilibre) est la probabilité $\boldsymbol{\pi} = (1/4, 1/4, 1/4, 1/4)$.

1.2 Probabilité de mutation

Pour estimer le taux α à partir des observations ou la date d'origine de l'espèce étudiée lorsqu'elle est très éloignée des temps d'observation, il est judicieux de suivre l'évolution de deux descendances issues du même ancêtre et de les comparer à chaque instant sur l'échelle des temps d'observation.

Considérons donc deux descendances $(X_t)_{t\geqslant 0}$ et $(Y_t)_{t\geqslant 0}$ issues du même ancêtre $X_0 = Y_0$. On suppose que ces deux descendances sont, conditionnellement à l'égalité $X_0 = Y_0$, indépendantes et de même loi et l'on va déterminer à l'aide d'une équation différentielle la probabilité d'observer à l'instant t une identité : $I(t) = \mathbb{P}(X_t = Y_t)$. On a

$$I(t+\varepsilon) = \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t = Y_t) I(t) + \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t \neq Y_t)(1 - I(t))$$

Or

$$\begin{aligned} \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t = Y_t) &= \frac{\sum_{a,b\in\mathcal{A}} \mathbb{P}(X_t = Y_t = a, X_{t+\varepsilon} = Y_{t+\varepsilon} = b)}{\sum_{a\in\mathcal{A}} \mathbb{P}(X_t = Y_t = a)} \\ &= \frac{\sum_{a,b\in\mathcal{A}} \mathbb{P}(X_t = a, X_{t+\varepsilon} = b)\mathbb{P}(Y_t = a, Y_{t+\varepsilon} = b)}{\sum_{a\in\mathcal{A}} \mathbb{P}(X_t = a)\mathbb{P}(Y_t = a)} \\ &= \frac{\sum_{a,b\in\mathcal{A}} \pi_a(t)^2 p_{ab}(\varepsilon)^2}{\sum_{a\in\mathcal{A}} \pi_a(t)^2}. \end{aligned}$$

Cette expression se simplifie en remarquant que par définition

$$p_{ab}(\varepsilon) = \begin{cases} \alpha\varepsilon + o(\varepsilon) & \text{si } a \neq b, \\ 1 - 3\alpha\varepsilon + o(\varepsilon) & \text{si } a = b, \end{cases}$$

et donc

$$\sum_{b\in\mathcal{A}} p_{ab}(\varepsilon)^2 = 1 - 6\alpha\varepsilon + o(\varepsilon).$$

On en déduit

$$\mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t = Y_t) = 1 - 6\alpha\varepsilon + o(\varepsilon).$$

De même,

$$\begin{aligned} \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t \neq Y_t) &= \frac{\sum_{a,b,c\in\mathcal{A},a\neq b} \mathbb{P}(X_t = a, Y_t = b, X_{t+\varepsilon} = Y_{t+\varepsilon} = c)}{\sum_{a,b\in\mathcal{A},a\neq b} \mathbb{P}(X_t = a, Y_t = b)} \\ &= \frac{\sum_{a,b,c\in\mathcal{A},a\neq b} \mathbb{P}(X_t = a, X_{t+\varepsilon} = c)\mathbb{P}(Y_t = b, Y_{t+\varepsilon} = c)}{\sum_{a,b\in\mathcal{A},a\neq b} \mathbb{P}(X_t = a)\mathbb{P}(Y_t = b)} \\ &= \frac{\sum_{a,b,c\in\mathcal{A},a\neq b} \pi_a(t)\pi_b(t) p_{ac}(\varepsilon) p_{bc}(\varepsilon)}{\sum_{a,b\in\mathcal{A},a\neq b} \pi_a(t)\pi_b(t)}. \end{aligned}$$

On voit facilement que pour $a \neq b$,

$$p_{ac}(\varepsilon)p_{bc}(\varepsilon) = \begin{cases} (\alpha\varepsilon + o(\varepsilon))^2 & \text{si } a, b \neq c, \\ (1 - 3\alpha\varepsilon + o(\varepsilon))(\alpha\varepsilon + o(\varepsilon)) & \text{si } a = c \text{ ou } b = c, \end{cases}$$

et donc

$$\sum_{c \in \mathcal{A}} p_{ac}(\varepsilon)p_{bc}(\varepsilon) = 2\alpha\varepsilon + o(\varepsilon).$$

On obtient alors

$$\mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t \neq Y_t) = 2\alpha\varepsilon + o(\varepsilon).$$

De ces deux probabilités conditionnelles, on tire la relation

$$I(t + \varepsilon) = (1 - 6\alpha\varepsilon)I(t) + 2\alpha\varepsilon(1 - I(t)) + o(\varepsilon),$$

ou encore

$$\frac{1}{\varepsilon}[I(t + \varepsilon) - I(t)] = -8\alpha I(t) + 2\alpha + o(1)$$

ce qui conduit à l'équation différentielle

$$I'(t) = -8\alpha I(t) + 2\alpha \quad \text{avec} \quad I(0) = 1.$$

Cette dernière a pour solution

$$\boxed{I(t) = \mathbb{P}(X_t = Y_t) = \tfrac{1}{4}\left[1 + 3e^{-8\alpha t}\right].}$$

On remarque que

$$\boxed{I(t) = a(2t).}$$

Cette coïncidence sera expliquée ultérieurement. Le nombre moyen de substitutions entre les deux lignées pendant le laps de temps $[0, t]$ (instants pour lesquels $X_s \neq Y_s$ avec $X_s \neq X_{s-}$ ou $Y_s \neq Y_{s-}$) vaut $D_t = 2 \times 3\alpha t = 6\alpha t$; cette quantité s'exprime à l'aide de $I(t)$ selon

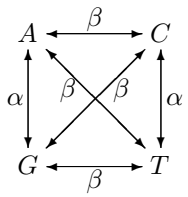
$$\boxed{D_t = -\tfrac{3}{4} \ln\left[\tfrac{1}{3}\left(4I(t) - 1\right)\right].}$$

2 Modèle à deux paramètres (M. Kimura, 1980)

Le modèle de Jukes-Cantor suppose les substitutions entre bases équiprobables, ce qui n'est pas très réaliste. Le modèle de Kimura à deux paramètres distingue deux types de substitutions entre les bases (voir Figure 2) :

- les transitions (ts) : substitutions internes entre purines $A \longleftrightarrow G$ et substitutions internes entre pyrimidines $C \longleftrightarrow T$ (taux α) ;
- les transversions (tv) : substitutions externes entre purines et pyrimidines $\{A, G\} \longleftrightarrow \{C, T\}$ (taux β).

Généralement, les transitions sont plus fréquentes que les transversions.



	A	G	C	T
A	$-\alpha - 2\beta$	α	β	β
G	α	$-\alpha - 2\beta$	β	β
C	β	β	$-\alpha - 2\beta$	α
T	β	β	α	$-\alpha - 2\beta$

FIGURE 2 – Modèle à deux paramètres

2.1 Détermination du semi-groupe

Les équations de Kolmogorov sont encore utilisables ici et le calcul de $P(t)$, plus compliqué, est encore faisable. Posons $J = \frac{1}{2}\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$, $K = \frac{1}{2}\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$ et $L = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$. Le générateur infinitésimal de la chaîne de Markov $(X_t)_{t\geqslant 0}$ s'exprime selon

$$\boxed{\boldsymbol{A} = \begin{pmatrix} -\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha-2\beta & \beta & \beta \\ \beta & \beta & -\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha-2\beta \end{pmatrix} = 2[-\alpha J + \beta(K-I)].}$$

Remarquons que $J^2 = J$, $K^2 = L^2 = L$, $JK = KJ = JL = LJ = 0$, $KL = LK = K$ et alors

$$\exp(-2\alpha t J) = I - (1 - e^{-2\alpha t})J,$$

$$\exp(2\beta t K) = I + \left[\sum_{n=1}^{\infty} \frac{(2\beta t)^{2n}}{(2n)!}\right] L + \left[\sum_{n=1}^{\infty} \frac{(2\beta t)^{2n+1}}{(2n+1)!}\right] K = I + (\cosh(2\beta t) - 1)L + \sinh(2\beta t)K$$

puis

$$\begin{aligned} P(t) &= e^{-2\beta t}\exp(-2\alpha t J)\exp(2\beta t K) \\ &= e^{-2\beta t}\Big[I - (1-e^{-2\alpha t})J + \sinh(2\beta t)K + (\cosh(2\beta t) - 1)L\Big] \\ &= e^{-2\beta t} I - (e^{-2\beta t} - e^{-2(\alpha+\beta)t})J + \frac{1}{2}(1 - e^{-4\beta t})K + \frac{1}{2}(1 - 2e^{-2\beta t} + e^{-4\beta t})L. \end{aligned}$$

Ainsi $P(t)$ est de la forme

$$\boxed{P(t) = \begin{pmatrix} a(t) & b(t) & c(t) & c(t) \\ b(t) & a(t) & c(t) & c(t) \\ c(t) & c(t) & a(t) & b(t) \\ c(t) & c(t) & b(t) & a(t) \end{pmatrix}}$$

où $a(t)$, $b(t)$ et $c(t)$ sont respectivement associés aux probabilités de conservation, de transition et de transversion d'une base à l'instant t comme suit :

$$\boxed{\begin{aligned} a(t) &= \mathbb{P}(X_t = X_0) = \tfrac{1}{4}\left[1 + 2e^{-2(\alpha+\beta)t} + e^{-4\beta t}\right], \\ b(t) &= \mathbb{P}(X_t \longleftrightarrow X_0 \text{ ts}) = \tfrac{1}{4}\left[1 - 2e^{-2(\alpha+\beta)t} + e^{-4\beta t}\right], \\ c(t) &= \tfrac{1}{2}\,\mathbb{P}(X_t \longleftrightarrow X_0 \text{ tv}) = \tfrac{1}{4}\left[1 - e^{-4\beta t}\right]. \end{aligned}}$$

On voit que à l'aide de ces expressions que

$$\boxed{P(t) \underset{t\to+\infty}{\longrightarrow} \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}}$$

ce qui montre que la chaîne de Markov $(X_t)_{t\geqslant 0}$ est ergodique de loi stationnaire la probabilité $\boldsymbol{\pi} = (1/4, 1/4, 1/4, 1/4)$.

2.2 Probabilité de mutation

Pour estimer les taux α et β à partir des observations on va suivre l'évolution de deux descendances $(X_t)_{t\geqslant 0}$ et $(Y_t)_{t\geqslant 0}$ issues du même ancêtre, indépendantes et de même loi, et de les comparer à chaque instant.

On introduit les probabilités suivantes :

$$\begin{aligned} I(t) &= \mathbb{P}(X_t = Y_t), \\ TS(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ ts}) = \mathbb{P}((X_t, Y_t) \in \{(A,G),(G,A),(C,T),(T,C)\}), \\ TV(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ tv}) = \mathbb{P}((X_t, Y_t) \in \{(A,C),(A,T),(G,C),(G,T),(C,A),(C,G),(T,A),(T,G)\}). \end{aligned}$$

On a

$$\begin{aligned} I(t+\varepsilon) = {} & \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t = Y_t) I(t) + \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t \longleftrightarrow Y_t \text{ ts}) TS(t) \\ & + \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t \longleftrightarrow Y_t \text{ tv}) TV(t). \end{aligned}$$

En énumérant toutes les possibilités, on trouve

$$\begin{aligned} \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t = Y_t) &= \frac{\sum_{a,b \in \mathcal{A}} \pi_a(t)^2 p_{ab}(\varepsilon)^2}{\sum_{a \in \mathcal{A}} \pi_a(t)^2} \\ &= 1 - 2(\alpha + 2\beta)\varepsilon + o(\varepsilon), \\ \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t \longleftrightarrow Y_t \text{ ts}) &= \frac{\sum_{a,b,c \in \mathcal{A}, a \leftrightarrow b \text{ ts}} \pi_a(t) \pi_b(t) p_{ac}(\varepsilon) p_{bc}(\varepsilon)}{\sum_{a,b \in \mathcal{A}, a \leftrightarrow b \text{ ts}} \pi_a(t) \pi_b(t)} \\ &= 2\alpha\varepsilon + o(\varepsilon), \\ \mathbb{P}(X_{t+\varepsilon} = Y_{t+\varepsilon} \mid X_t \longleftrightarrow Y_t \text{ tv}) &= \frac{\sum_{a,b,c \in \mathcal{A}, a \leftrightarrow b \text{ tv}} \pi_a(t) \pi_b(t) p_{ac}(\varepsilon) p_{bc}(\varepsilon)}{\sum_{a,b \in \mathcal{A}, a \leftrightarrow b \text{ tv}} \pi_a(t) \pi_b(t)} \\ &= 2\beta\varepsilon + o(\varepsilon). \end{aligned}$$

De ces trois probabilités conditionnelles, on tire la relation

$$I(t+\varepsilon) = [1 - 2(\alpha + 2\beta)\varepsilon] I(t) + 2\alpha\varepsilon TS(t) + 2\beta\varepsilon TV(t) + o(\varepsilon),$$

de laquelle on déduit l'équation différentielle

$$I'(t) = -2(\alpha + 2\beta) I(t) + 2\alpha TS(t) + 2\beta TV(t).$$

De même pour $TS(t+\varepsilon)$:

$$\begin{aligned} TS(t+\varepsilon) = {} & \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ ts} \mid X_t = Y_t) I(t) + \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ ts} \mid X_t \longleftrightarrow Y_t \text{ ts}) TS(t) \\ & + \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ ts} \mid X_t \longleftrightarrow Y_t \text{ tv}) TV(t). \end{aligned}$$

On a, après calculs,

$$\begin{aligned} \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ ts} \mid X_t = Y_t) &= \frac{\sum_{a,b,c \in \mathcal{A}, b \leftrightarrow c \text{ ts}} \pi_a(t)^2 p_{ab}(\varepsilon) p_{ac}(\varepsilon)}{\sum_{a \in \mathcal{A}} \pi_a(t)^2} \\ &= 2\alpha\varepsilon + o(\varepsilon), \\ \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ ts} \mid X_t \longleftrightarrow Y_t \text{ ts}) &= \frac{\sum_{a,b,c,d \in \mathcal{A}, a \leftrightarrow b \text{ ts}, c \leftrightarrow d \text{ ts}} \pi_a(t) \pi_b(t) p_{ac}(\varepsilon) p_{bd}(\varepsilon)}{\sum_{a,b \in \mathcal{A}, a \leftrightarrow b \text{ ts}} \pi_a(t) \pi_b(t)} \\ &= 1 - 2(\alpha + 2\beta)\varepsilon + o(\varepsilon), \\ \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ ts} \mid X_t \longleftrightarrow Y_t \text{ tv}) &= \frac{\sum_{a,b,c,d \in \mathcal{A}, a \leftrightarrow b \text{ tv}, c \leftrightarrow d \text{ ts}} \pi_a(t) \pi_b(t) p_{ac}(\varepsilon) p_{bd}(\varepsilon)}{\sum_{a,b \in \mathcal{A}, a \leftrightarrow b \text{ tv}} \pi_a(t) \pi_b(t)} \\ &= 2\beta\varepsilon + o(\varepsilon). \end{aligned}$$

On en tire la relation

$$TS(t+\varepsilon) = 2\alpha\varepsilon I(t) + [1 - 2(\alpha + 2\beta)\varepsilon] TS(t) + 2\beta\varepsilon TV(t) + o(\varepsilon),$$

de laquelle on déduit l'équation différentielle

$$TS'(t) = 2\alpha I(t) - 2(\alpha + 2\beta) TS(t) + 2\beta TV(t).$$

De même pour $TV(t+\varepsilon)$:

$$\begin{aligned} TV(t+\varepsilon) = {} & \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ tv} \mid X_t = Y_t) I(t) + \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ tv} \mid X_t \longleftrightarrow Y_t \text{ ts}) TS(t) \\ & + \mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ tv} \mid X_t \longleftrightarrow Y_t \text{ tv}) TV(t). \end{aligned}$$

On a, après calculs,

$$\begin{aligned}
\mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ tv} \mid X_t = Y_t) &= \frac{\sum_{a,b,c\in\mathcal{A}, b\leftrightarrow c \text{ tv}} \pi_a(t)^2 p_{ab}(\varepsilon) p_{ac}(\varepsilon)}{\sum_{a\in\mathcal{A}} \pi_a(t)^2} \\
&= 4\beta\varepsilon + o(\varepsilon), \\
\mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ tv} \mid X_t \longleftrightarrow Y_t \text{ ts}) &= \frac{\sum_{a,b,c,d\in\mathcal{A}, a\leftrightarrow b \text{ ts}, c\leftrightarrow d \text{ tv}} \pi_a(t)\pi_b(t) p_{ac}(\varepsilon) p_{bd}(\varepsilon)}{\sum_{a,b\in\mathcal{A}, a\leftrightarrow b \text{ ts}} \pi_a(t)\pi_b(t)} \\
&= 4\beta\varepsilon + o(\varepsilon), \\
\mathbb{P}(X_{t+\varepsilon} \longleftrightarrow Y_{t+\varepsilon} \text{ tv} \mid X_t \longleftrightarrow Y_t \text{ tv}) &= \frac{\sum_{a,b,c,d\in\mathcal{A}, a\leftrightarrow b \text{ tv}, c\leftrightarrow d \text{ tv}} \pi_a(t)\pi_b(t) p_{ac}(\varepsilon) p_{bd}(\varepsilon)}{\sum_{a,b\in\mathcal{A}, a\leftrightarrow b \text{ tv}} \pi_a(t)\pi_b(t)} \\
&= 1 - 4\beta\varepsilon + o(\varepsilon).
\end{aligned}$$

On obtient la relation

$$TV(t+\varepsilon) = 2\alpha\varepsilon I(t) + [1 - 2(\alpha+2\beta)\varepsilon] TS(t) + 2\beta\varepsilon TV(t) + o(\varepsilon),$$

de laquelle on déduit

$$TV'(t) = 4\beta I(t) - 4\beta TS(t) - 4\beta TV(t).$$

Remarquons que l'identité $I'+TS'+TV' = 0$ est satisfaite, ce qui était prévisible puisque $I+TS+TV = 1$. Cette remarque nous conduit à un système différentiel à deux inconnues :

$$\begin{cases} I'(t) = 2\beta - (2\alpha+6\beta) I(t) + 2(\alpha-\beta) TS(t) \\ TS'(t) = 2\beta + 2(\alpha-\beta) I(t) - (2\alpha+6\beta) TS(t) \end{cases}$$

avec les conditions initiales $I(0) = 1$ et $TS(0) = 0$. En introduisant la matrice associée à ce système $B = \begin{pmatrix} -(2\alpha+6\beta) & 2(\alpha-\beta) \\ 2(\alpha-\beta) & -(2\alpha+6\beta) \end{pmatrix}$, la solution s'exprime selon

$$\begin{pmatrix} I'(t) \\ TS'(t) \end{pmatrix} = \exp(tB) \begin{pmatrix} 1 \\ 0 \end{pmatrix} + [\exp(tB) - 1] B^{-1} \begin{pmatrix} 2\beta \\ 2\beta \end{pmatrix}.$$

La matrice B est de la forme $\begin{pmatrix} a & b \\ b & a \end{pmatrix}$, cette dernière est diagonalisable et se décompose sous la forme

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a+b & 0 \\ 0 & a-b \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$

d'où

$$\exp\left[t \begin{pmatrix} a & b \\ b & a \end{pmatrix} \right] = e^{at} \begin{pmatrix} \cosh(bt) & \sinh(bt) \\ \sinh(bt) & \cosh(bt) \end{pmatrix}$$

soit encore

$$\exp(tB) = e^{-(2\alpha+6\beta)t} \begin{pmatrix} \cosh(2(\alpha-\beta)t) & \sinh(2(\alpha-\beta)t) \\ \sinh(2(\alpha-\beta)t) & \cosh(2(\alpha-\beta)t) \end{pmatrix}.$$

On obtient finalement

$$\boxed{\begin{aligned}
I(t) &= \mathbb{P}(X_t = Y_t) = \tfrac{1}{4}\left[1 + 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}\right], \\
TS(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ ts}) = \tfrac{1}{4}\left[1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}\right], \\
TV(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ tv}) = \tfrac{1}{2}\left[1 - e^{-8\beta t}\right].
\end{aligned}}$$

On observe de nouveau que

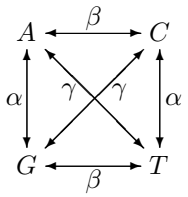
$$\boxed{I(t) = a(2t), \quad TS(t) = b(2t), \quad TV(t) = 2c(2t).}$$

Le nombre moyen de substitutions entre les deux lignées pendant le laps de temps $[0,t]$ vaut $D_t = 2 \times (\alpha+2\beta)t$; cette quantité s'exprime à l'aide de $I(t), TS(t)$ et $TV(t)$ selon

$$\boxed{D_t = -\tfrac{1}{2} \ln\left[(I(t) - TS(t))\sqrt{1 - 2TV(t)}\right].}$$

3 Modèle à trois paramètres (M. Kimura, 1981)

On reprend le modèle précédent en différenciant de plus les transversions $A \longleftrightarrow C$ et $G \longleftrightarrow T$ des transversions $A \longleftrightarrow T$ et $G \longleftrightarrow C$. On distingue ainsi trois types de substitutions (voir Figure 3) :
- les transitions (ts) $A \longleftrightarrow G$ et $C \longleftrightarrow T$ (taux α) ;
- les transversions (tv1) $A \longleftrightarrow C$ et $G \longleftrightarrow T$ (taux β) ;
- les transversions (tv2) $A \longleftrightarrow T$ et $G \longleftrightarrow C$ (taux γ).



	A	G	C	T
A	$-\alpha-\beta-\gamma$	α	β	γ
G	α	$-\alpha-\beta-\gamma$	γ	β
C	β	γ	$-\alpha-\beta-\gamma$	α
T	γ	β	α	$-\alpha-\beta-\gamma$

FIGURE 3 – Modèle à trois paramètres

3.1 Détermination du semi-groupe

Le générateur infinitésimal de la chaîne de Markov $(X_t)_{t\geqslant 0}$ est donné par

$$\boxed{\boldsymbol{A} = \begin{pmatrix} \sigma & \alpha & \beta & \gamma \\ \alpha & \sigma & \gamma & \beta \\ \beta & \gamma & \sigma & \alpha \\ \gamma & \beta & \alpha & \sigma \end{pmatrix}},$$

où l'on a posé $\sigma = -\alpha-\beta-\gamma$. Cette matrice est diagonalisable et se décompose selon $\boldsymbol{A} = QDQ^{-1}$ avec

$$Q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(\beta+\gamma) & 0 & 0 \\ 0 & 0 & -2(\alpha+\gamma) & 0 \\ 0 & 0 & 0 & -2(\alpha+\beta) \end{pmatrix}, \quad Q^{-1} = \frac{1}{4}\,Q.$$

Le semi-groupe a donc pour expression

$$\boxed{P(t) = \begin{pmatrix} a(t) & b(t) & c(t) & d(t) \\ b(t) & a(t) & d(t) & c(t) \\ c(t) & d(t) & a(t) & b(t) \\ d(t) & c(t) & b(t) & a(t) \end{pmatrix}}$$

où $a(t)$, $b(t)$, $c(t)$, $d(t)$ représentent respectivement les probabilités suivantes

$$\boxed{\begin{aligned} a(t) &= \mathbb{P}(X_t = X_0) = \tfrac{1}{4}\,[1 + e^{-2(\alpha+\beta)t} + e^{-2(\alpha+\gamma)t} + e^{-2(\beta+\gamma)t}], \\ b(t) &= \mathbb{P}(X_t \longleftrightarrow X_0 \text{ ts}) = \tfrac{1}{4}\,[1 - e^{-2(\alpha+\beta)t} - e^{-2(\alpha+\gamma)t} + e^{-2(\beta+\gamma)t}], \\ c(t) &= \tfrac{1}{2}\,\mathbb{P}(X_t \longleftrightarrow X_0 \text{ tv1}) = \tfrac{1}{4}\,[1 - e^{-2(\alpha+\beta)t} + e^{-2(\alpha+\gamma)t} - e^{-2(\beta+\gamma)t}], \\ d(t) &= \tfrac{1}{2}\,\mathbb{P}(X_t \longleftrightarrow X_0 \text{ tv2}) = \tfrac{1}{4}\,[1 + e^{-2(\alpha+\beta)t} - e^{-2(\alpha+\gamma)t} - e^{-2(\beta+\gamma)t}]. \end{aligned}}$$

En particulier,

$$\mathbb{P}(X_t \longleftrightarrow X_0 \text{ tv}) = c(t) + d(t) = \frac{1}{2}\,[1 - e^{-2(\beta+\gamma)t}].$$

On observe directement à l'aide des expressions précédentes que

$$\boxed{P(t) \underset{t\to+\infty}{\longrightarrow} \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}}$$

ce qui montre de nouveau que la chaîne de Markov $(X_t)_{t\geqslant 0}$ est ergodique de loi stationnaire la probabilité $\boldsymbol{\pi} = (1/4, 1/4, 1/4, 1/4)$.

3.2 Probabilité de mutation

On peut estimer les taux α, β et γ à partir des observations comme précédemment en suivant l'évolution de deux descendances $(X_t)_{t\geqslant 0}$ et $(Y_t)_{t\geqslant 0}$ issues du même ancêtre, indépendantes et de même loi.

On montre à l'aide d'un système différentiel que les probabilités

$$\begin{aligned}
I(t) &= \mathbb{P}(X_t = Y_t),\\
TS(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ ts}) = \mathbb{P}((X_t, Y_t) \in \{(A,G),(G,A),(C,T),(T,C)\}),\\
TV1(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ tv1}) = \mathbb{P}((X_t, Y_t) \in \{(A,C),(G,T),(C,A),(T,G)\}),\\
TV2(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ tv2}) = \mathbb{P}((X_t, Y_t) \in \{(A,T),(G,C),(C,G),(T,A)\})
\end{aligned}$$

coïncident respectivement avec $a(2t)$, $b(2t)$, $c(2t)$, $d(2t)$:

$$\boxed{\begin{aligned}
I(t) &= \mathbb{P}(X_t = Y_t) = a(2t),\\
TS(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ ts}) = b(2t),\\
TV1(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ tv1}) = c(2t),\\
TV2(t) &= \mathbb{P}(X_t \longleftrightarrow Y_t \text{ tv2}) = d(2t).
\end{aligned}}$$

Désignons par N_t le nombre de substitutions pendant l'intervalle de temps $[0,t]$ et par $T_0 = 0 < T_1 < T_2 < \ldots < T_n < \ldots$ la suite des instants successifs de substitutions. On a $N_t = \max\{n \in \mathbb{N} : T_n \leqslant t\}$. Dans le cas d'une chaîne de Markov générale $(X_t)_{t\geqslant 0}$, la suite des laps de temps entre deux mutations consécutives $T_1, T_2 - T_1, \ldots, T_n - T_{n-1}$, conditionnellement à la nature de chaque base aux instants de substitutions, disons $X_{T_0} = a_0, X_{T_1} = a_1, \ldots, X_{T_{n-1}} = a_{n-1}$, est constituée de v.a. indépendantes suivant les lois exponentielles de paramètres respectifs $\lambda_{a_0}, \lambda_{a_1}, \ldots, \lambda_{a_{n-1}}$. Dans le cas des modèles à un, deux et trois paramètres considérés, il se trouve que les termes diagonaux $(A_{aa} = -\lambda_a)$ du générateur $\boldsymbol{A}$ sont identiques ; les λ_a, $a \in \mathcal{A}$, ne dépendent pas de la nature de la base suivie prouvant ainsi que les v.a. $T_1, T_2 - T_1, \ldots, T_n - T_{n-1}, \ldots$ sont indépendantes de loi exponentielle $\mathcal{E}(\lambda)$, λ étant la valeur commune des λ_a, $a \in \mathcal{A}$. Dans ces conditions, T_n suit la loi d'Erlang $E(n,\lambda)$ de densité

$$\mathbb{P}(T_n \in dt)/dt = \frac{\lambda^n t^{n-1}}{(n-1)!}\, e^{-\lambda t}$$

et $(N_t)_{t\geqslant 0}$ est alors un processus de Poisson d'intensité λ. En particulier, N_t suit la loi de Poisson $\mathcal{P}(\lambda t)$:

$$\mathbb{P}(N_t = n) = \frac{(\lambda t)^n}{n\,!}\, e^{-\lambda t}.$$

Son espérance vaut $\mathbb{E}(N_t) = \lambda t$.

Par exemple, dans le cas du modèle à trois paramètres, $\lambda = \alpha + \beta + \gamma$ et le nombre moyen de substitutions dans une lignée durant l'intervalle de temps $[0,t]$ vaut $(\alpha+\beta+\gamma)t$. Dans le cas de deux lignées indépendantes modélisées par le couplage $Z_t = (X_t, Y_t)$, $t \geqslant 0$, le premier instant de saut T_1^Z est relié aux premiers instants de saut T_1^X et T_1^Y par $T_1^Z = \min(T_1^X, T_1^Y)$ et par conséquent T_1^Z suit la loi exponentielle dont le paramètre est la somme des paramètres des lois de T_1^X et T_1^Y, soit ici $2(\alpha+\beta+\gamma)$. De même, La suite des différences successives entre les deux descendances parallèles $T_1^Z, T_2^Z - T_1^Z, \ldots, T_n^Z - T_{n-1}^Z, \ldots$ est constituée de v.a. indépendantes suivant la loi $\mathcal{E}(2(\alpha+\beta+\gamma))$. Le nombre moyen de substitutions entre les deux lignées pendant le laps de temps $[0,t]$ vaut alors $D_t = 2(\alpha+\beta+\gamma)t$ et s'exprime à l'aide de $I(t), TS(t)$ $TV1(t)$ et $TV2(t)$ selon

$$\boxed{\begin{aligned}
D_t = -\tfrac{1}{4} \ln \big[&(I(t) - TS(t) - TV1(t) + TV2(t))\\
&\times (I(t) - TS(t) + TV1(t) - TV2(t)) \times (I(t) + TS(t) - TV1(t) - TV2(t))\big].
\end{aligned}}$$

Naturellement, ce modèle regroupe en particulier les deux modèles précédents. Il est possible de calculer de manière beaucoup plus simple les probabilités de mutation en faisant appel à la notion de réversibilité.

La chaîne de Markov $(X_t)_{t\geqslant 0}$ est réversible si elle vérifie la propriété de symétrie

$$\forall (a,b) \in \mathcal{A}\times\mathcal{A},\ \forall s,t,\ 0 \leqslant s \leqslant t, \quad \mathbb{P}(X_s = a, X_t = b) = \mathbb{P}(X_s = b, X_t = a),$$

ce qui est équivalent, en posant $\pi = \pi(0)$, à

$$\forall (a,b) \in \mathcal{A} \times \mathcal{A},\ \forall t \geqslant 0, \quad \pi_a p_{ab}(t) = \pi_b p_{ba}(t),$$

ou encore sous forme matricielle, à l'aide du générateur $\boldsymbol{A}$,

$$D_\pi \boldsymbol{A} = {}^t\boldsymbol{A} D_\pi,$$

D_π désignant la matrice diagonale dont les termes diagonaux sont les $\pi_a, a \in \mathcal{A}$. Sous cette condition, la chaîne $(X_t)_{t\geqslant 0}$ peut être prolongée en une chaîne stationnaire indexée par $\mathbb{R}$ en posant $X_t = Y_{-t}$ pour $t \leqslant 0$, $(Y_t)_{t\geqslant 0}$ étant une copie de $(X_t)_{t\geqslant 0}$ démarrant de $Y_0 = X_0$ et indépendante de celle-ci. En introduisant le couplage $Z_t = (X_t, Y_t)$ et sa loi écrite sous forme d'une matrice carrée $S(t) = (s_{ab}(t))_{(a,b)\in\mathcal{A}\times\mathcal{A}}$ avec $s_{ab}(t) = \mathbb{P}(Z_t = (a,b))$, on a

$$s_{ab}(t) = \sum_{c\in\mathcal{A}} \mathbb{P}(X_0 = c)\mathbb{P}(X_t = a \mid X_0 = c)\mathbb{P}(Y_t = b \mid Y_0 = c) = \sum_{c\in\mathcal{A}} \pi_c p_{ca}(t) p_{cb}(t),$$

soit encore

$$S(t) = {}^tP(t) D_\pi P(t) = D_\pi P(t)^2 = D_\pi P(2t).$$

En particulier, les probabilités d'observer respectivement une mutation entre X_0 et X_t $(X_t \neq X_0)$ et une mutation entre les deux lignées à l'instant t $(X_t \neq Y_t)$ s'obtiennent de la manière suivante :

$$\begin{aligned} r(t) \equiv \mathbb{P}(X_t \neq X_0) &= 1 - \mathbb{P}(X_t = X_0) \\ &= 1 - \sum_{a\in\mathcal{A}} \pi_a p_{aa}(t) \\ \rho(t) \equiv \mathbb{P}(X_t \neq Y_t) &= 1 - \mathbb{P}(X_t = Y_t) \\ &= 1 - \sum_{a\in\mathcal{A}} s_{aa}(t) \\ &= 1 - \sum_{a\in\mathcal{A}} \pi_a p_{aa}(2t), \end{aligned}$$

et l'on remarque que ces deux probabilités sont reliées par

$$\boxed{\rho(t) = r(2t).}$$

Cette relation explique les coïncidences rencontrées dans les modèles précédents entre les diverses probabilités de mutations et les probabilités du semi-groupe $(P_t)_{t\geqslant 0}$ via le changement de temps $t \mapsto 2t$. Le taux de mutation défini par

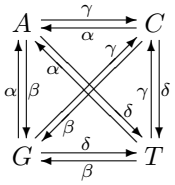
$$\kappa = \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\,\mathbb{P}(X_{t+\varepsilon} \neq X_t) = \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\,\mathbb{P}(X_\varepsilon \neq X_0) = r'(0) = -\sum_{a\in\mathcal{A}} \pi_a A_{aa}$$

est utilisé pour définir un paramètre d'évolution :

$$\boxed{D_t = 2\kappa t.}$$

4 Modèle à quatre paramètres (Felsenstein, 1981 et F. Tajima & M. Nei, 1982)

Ce modèle suppose que les substitutions sur le laps de temps $[t, t+\varepsilon]$ ne dépendent pas de la nature de la base à l'instant t, ce qui se traduit par le fait que le taux $\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon} p_{ab}(\varepsilon)$ ne dépend pas de a pour tout $a \in \mathcal{A} \setminus \{b\}$ (voir Figure 4).



	A	G	C	T
A	$-\beta-\gamma-\delta$	β	γ	δ
G	α	$-\alpha-\gamma-\delta$	γ	δ
C	α	β	$-\alpha-\beta-\delta$	δ
T	α	β	γ	$-\alpha-\beta-\gamma$

FIGURE 4 – Modèle à quatre paramètres

4.1 Détermination du semi-groupe

Posons $\sigma = \alpha+\beta+\gamma+\delta$ et $J = \frac{1}{\sigma}\begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \end{pmatrix}$. Le générateur infinitésimal de la chaîne de Markov $(X_t)_{t\geqslant 0}$ a pour expression

$$\boxed{\boldsymbol{A} = \begin{pmatrix} \alpha-\sigma & \beta & \gamma & \delta \\ \alpha & \beta-\sigma & \gamma & \delta \\ \alpha & \beta & \gamma-\sigma & \delta \\ \alpha & \beta & \gamma & \delta-\sigma \end{pmatrix} = \sigma(J-I).}$$

Puisque $J^2 = J$, on trouve facilement que le semi-groupe recherché est donné par

$$P(t) = \exp(t\boldsymbol{A}) = e^{-\sigma t}I + \frac{1}{\sigma}\,(1-e^{-\sigma t})J.$$

On en déduit immédiatement que

$$\boxed{P(t) \underset{t\to+\infty}{\longrightarrow} \tfrac{1}{\sigma}\,J = \tfrac{1}{\sigma}\begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \end{pmatrix}}$$

montrant ainsi que la chaîne de Markov $(X_t)_{t\geqslant 0}$ est ergodique de loi stationnaire $\boldsymbol{\pi} = \frac{1}{\sigma}(\alpha,\beta,\gamma,\delta)$.

4.2 Probabilité de mutation

La chaîne $(X_t)_{t\geqslant 0}$ est stationnaire lorsque l'on choisit comme loi initiale $\pi = \frac{1}{\sigma}(\alpha,\beta,\gamma,\delta)$. On peut donc appliquer la théorie développée dans le paragraphe précédent. On a tout d'abord

$$\begin{aligned} p_{aa}(t) &= e^{-\sigma t} + \pi_a(1-e^{-\sigma t}) = \pi_a + (1-\pi_a)e^{-\sigma t} \\ A_{aa} &= -\sigma(1-\pi_a). \end{aligned}$$

Avec les mêmes notations que dans la section précédente, les probabilités et taux de mutation ont respectivement pour expression

$$\begin{aligned} r(t) &= 1 - \sum_{a\in\mathcal{A}} \pi_a[\pi_a + (1-\pi_a)e^{-\sigma t}] = (1-\sum_{a\in\mathcal{A}}\pi_a^2)(1-e^{-\sigma t}), \\ \rho(t) &= r(2t), \\ \kappa &= r'(0) = \sigma(1-\sum_{a\in\mathcal{A}}\pi_a^2). \end{aligned}$$

Le paramètre d'évolution est donc donné par

$$\boxed{D_t = 2\kappa t = -\frac{\kappa}{\sigma}\,\ln\Big[1-\frac{\sigma}{\kappa}\,\rho(t)\Big].}$$

4.3 Statistique

On étudie n paires de sites homologues en parallèle dans la même espèce à l'instant t : $(X_1(t), Y_1(t)), \ldots, (X_n(t), Y_n(t))$ indépendantes de même loi que Z_t. Ces hypothèses d'indépendance et d'identique distribution ne sont pas toujours réalistes, notamment lorsque l'on travaille avec des codons (succession de trois bases) traduisant une protéine du fait de la dégénérescence du code génétique ; en effet, plusieurs codons distincts peuvent coder pour le même acide aminé (codons synonymes).

Soit, pour $(a, b) \in \mathcal{A} \times \mathcal{A}$,

$$N_{ab} = \text{card}\{l \in \{1, \ldots, n\} : X_l(t) = a, Y_l(t) = b\}$$

et

$$D = \sum_{(a,b)\in\mathcal{A}\times\mathcal{A}, a\neq b} N_{ab}$$

le nombre de paires ayant des bases distinctes. Le vecteur aléatoire $(N_{ab})_{(a,b)\in\mathcal{A}\times\mathcal{A}}$ suit la loi multinomiale $\mathcal{M}(n, (s_{ab}(t))_{(a,b)\in\mathcal{A}\times\mathcal{A}})$ où $s_{ab}(t) = \mathbb{P}(X_t = a, Y_t = b) = \pi_a p_{ab}(2t)$, et alors la v.a. D suit la loi binomiale $\mathcal{B}(n, \sum_{(a,b)\in\mathcal{A}\times\mathcal{A}, a\neq b} s_{ab}(t))$. On a donc

$$\mathbb{E}(D) = n \sum_{(a,b)\in\mathcal{A}\times\mathcal{A}, a\neq b} s_{ab}(t) = n\mathbb{P}(X_t \neq Y_t) = n\rho(t)$$

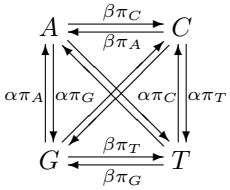
ce qui signifie que $\frac{1}{n} D$ est un estimateur sans biais de $\rho(t)$. On pourrait ainsi produire un estimateur du paramètre d'évolution D_t.

5 Modèle à six paramètres (M. Hasegawa, H. Kishino & T. Yano, 1985)

Par opposition aux modèles à un, deux ou trois paramètres, on considère à présent que les substitutions ne sont plus nécessairement symétriques (voir Figure 5) :

$$\mathbb{P}(X_{t+\varepsilon} = b \mid X_t = a) = \begin{cases} \alpha\pi_b\varepsilon + o(\varepsilon) \text{ si } a \longleftrightarrow b \text{ est une transition} \\ \beta\pi_b\varepsilon + o(\varepsilon) \text{ si } a \longleftrightarrow b \text{ est une transversion} \end{cases}.$$

On pose $\pi_R = \pi_A + \pi_G$ et $\pi_Y = \pi_C + \pi_T$ avec $\pi_R + \pi_Y = 1$.



	A	G	C	T
A	$-\alpha\pi_A - \beta\pi_Y$	$\alpha\pi_G$	$\beta\pi_C$	$\beta\pi_T$
G	$\alpha\pi_A$	$-\alpha\pi_G - \beta\pi_Y$	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	$-\alpha\pi_C - \beta\pi_R$	$\alpha\pi_T$
T	$\beta\pi_A$	$\beta\pi_G$	$\alpha\pi_C$	$-\alpha\pi_T - \beta\pi_R$

FIGURE 5 – Modèle à six paramètres

5.1 Détermination du semi-groupe

Le générateur infinitésimal de la chaîne de Markov $(X_t)_{t\geqslant 0}$ est donné par

$$\boldsymbol{A} = \begin{pmatrix} -\alpha\pi_G - \beta\pi_Y & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & -\alpha\pi_A - \beta\pi_Y & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & -\alpha\pi_T - \beta\pi_R & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & -\alpha\pi_C - \beta\pi_R \end{pmatrix}.$$

Afin de déterminer le semi-groupe correspondant, on va diagonaliser la matrice $\boldsymbol{A}$. Calculons d'abord son polynôme caractéristique :

$$\det(\boldsymbol{A}-\lambda I) = \begin{vmatrix} -\alpha\pi_G-\beta\pi_Y-\lambda & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & -\alpha\pi_A-\beta\pi_Y-\lambda & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & -\alpha\pi_T-\beta\pi_R-\lambda & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & -\alpha\pi_C-\beta\pi_R-\lambda \end{vmatrix}$$

$$= -\lambda \begin{vmatrix} 1 & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ 1 & -\alpha\pi_A-\beta\pi_Y-\lambda & \beta\pi_C & \beta\pi_T \\ 1 & \beta\pi_G & -\alpha\pi_T-\beta\pi_R-\lambda & \alpha\pi_T \\ 1 & \beta\pi_G & \alpha\pi_C & -\alpha\pi_C-\beta\pi_R-\lambda \end{vmatrix}$$

$$= -\lambda \begin{vmatrix} 0 & \alpha\pi_R+\beta\pi_Y+\lambda & 0 & 0 \\ 1 & -\alpha\pi_A-\beta\pi_Y-\lambda & \beta\pi_C & \beta\pi_T \\ 0 & 0 & -\alpha\pi_Y-\beta\pi_R-\lambda & \alpha\pi_Y+\beta\pi_R+\lambda \\ 1 & \beta\pi_G & \alpha\pi_C & -\alpha\pi_C-\beta\pi_R-\lambda \end{vmatrix}$$

$$= \lambda(\lambda+\alpha\pi_R+\beta\pi_Y) \begin{vmatrix} 1 & \beta\pi_C & \beta\pi_T \\ 0 & -\alpha\pi_Y-\beta\pi_R-\lambda & \alpha\pi_Y+\beta\pi_R+\lambda \\ 1 & \alpha\pi_C & -\alpha\pi_C-\beta\pi_R-\lambda \end{vmatrix}$$

$$= \lambda(\lambda+\alpha\pi_R+\beta\pi_Y)(\lambda+\alpha\pi_Y+\beta\pi_R) \begin{vmatrix} 1 & \beta\pi_C & \beta\pi_T \\ 0 & -1 & 1 \\ 1 & \alpha\pi_C & -\alpha\pi_C-\beta\pi_R-\lambda \end{vmatrix}$$

$$= \lambda(\lambda+\alpha\pi_R+\beta\pi_Y)(\lambda+\alpha\pi_Y+\beta\pi_R)(\lambda+\beta).$$

Les valeurs propres de $\boldsymbol{A}$ sont donc $0, \beta, -\alpha\pi_R-\beta\pi_Y, -\alpha\pi_Y-\beta\pi_R$. Déterminons les sous-espaces propres à droite associés :

1. pour la valeur propre 0, le sous-espace propre associé est donné par le système
$$\begin{cases} (-\alpha\pi_G-\beta\pi_Y)\,x + & \alpha\pi_G\,y + & \beta\pi_C\,z + & \beta\pi_T\,t = 0 \\ \alpha\pi_A\,x + (-\alpha\pi_A-\beta\pi_Y)\,y + & & \beta\pi_C\,z + & \beta\pi_T\,t = 0 \\ \beta\pi_A\,x + & \beta\pi_G\,y + (-\alpha\pi_T-\beta\pi_R)\,z + & & \alpha\pi_T\,t = 0 \\ \beta\pi_A\,x + & \beta\pi_G\,y + & \alpha\pi_C\,z + (-\alpha\pi_C-\beta\pi_R)\,t & = 0 \end{cases}$$
duquel on tire $x=y$ et $z=t$ puis $-\beta\pi_Y\,x+\beta\pi_Y\,z=0$, soit encore $x=y=z=t$; d'où la droite engendrée par $V_1 = {}^t(1,1,1,1)$.
2. pour la valeur propre $-\beta$, le sous-espace propre associé est donné par le système
$$\begin{cases} (-\alpha\pi_G+\beta\pi_R)\,x + & \alpha\pi_G\,y + & \beta\pi_C\,z + & \beta\pi_T\,t = 0 \\ \alpha\pi_A\,x + (-\alpha\pi_A+\beta\pi_R)\,y + & & \beta\pi_C\,z + & \beta\pi_T\,t = 0 \\ \beta\pi_A\,x + & \beta\pi_G\,y + (-\alpha\pi_T+\beta\pi_Y)\,z + & & \alpha\pi_T\,t = 0 \\ \beta\pi_A\,x + & \beta\pi_G\,y + & \alpha\pi_C\,z + (-\alpha\pi_C+\beta\pi_Y)\,t & = 0 \end{cases}$$
duquel on tire $x=y$ et $z=t$ puis $\beta\pi_R\,x+\beta\pi_Y\,z=0$, soit encore $z=-\frac{\pi_R}{\pi_Y}x$; d'où la droite engendrée par $V_2 = {}^t(\pi_Y,\pi_Y,-\pi_R,-\pi_R)$.
3. pour la valeur propre $-\alpha\pi_R-\beta\pi_Y$, le sous-espace propre associé est donné par le système
$$\begin{cases} \alpha\pi_A\,x + \alpha\pi_G\,y + & \beta\pi_C\,z + & \beta\pi_T\,t = 0 \\ \alpha\pi_A\,x + \alpha\pi_G\,y + & \beta\pi_C\,z + & \beta\pi_T\,t = 0 \\ \beta\pi_A\,x + \beta\pi_G\,y + [\alpha(\pi_R-\pi_T)-\beta(\pi_R-\pi_Y)]\,z + & & \alpha\pi_T\,t = 0 \\ \beta\pi_A\,x + \beta\pi_G\,y + & \alpha\pi_C\,z + [\alpha(\pi_R-\pi_C)-\beta(\pi_R-\pi_Y)]\,t & = 0 \end{cases}$$
duquel on tire $z=t$ puis $\alpha\pi_A\,x+\alpha\pi_G\,y+\beta\pi_Y\,z=0$ et $\beta\pi_A\,x+\beta\pi_G\,y+[\alpha\pi_R-\beta(\pi_R-\pi_Y)]\,z=0$, soit encore $z=0$ et $y=-\frac{\pi_A}{\pi_G}x$; d'où la droite engendrée par $V_3 = {}^t(\pi_G,-\pi_A,0,0)$.
4. pour la valeur propre $-\alpha\pi_Y-\beta\pi_R$, le sous-espace propre associé est donné par le système
$$\begin{cases} [\alpha(\pi_Y-\pi_G)-\beta(\pi_Y-\pi_R)]\,x + & \alpha\pi_G\,y + \beta\pi_C\,z + \beta\pi_T\,t = 0 \\ \alpha\pi_A\,x + [\alpha(\pi_Y-\pi_A)-\beta(\pi_Y-\pi_R)]\,y + & \beta\pi_C\,z + \beta\pi_T\,t = 0 \\ \beta\pi_A\,x + & \beta\pi_G\,y + \alpha\pi_C\,z + \alpha\pi_T\,t = 0 \\ \beta\pi_A\,x + & \beta\pi_G\,y + \alpha\pi_C\,z + \alpha\pi_T\,t = 0 \end{cases}$$
duquel on tire $x=y$ puis $[\alpha\pi_Y-\beta(\pi_Y-\pi_R)]\,x+\beta\pi_C\,z+\beta\pi_T\,t=0$ et $\beta\pi_R\,x+\alpha\pi_C\,z+\alpha\pi_T\,t=0$, soit encore $x=0$ et $t=-\frac{\pi_C}{\pi_T}z$; d'où la droite engendrée par $V_4 = {}^t(0,0,\pi_T,-\pi_C)$.

La matrice $\boldsymbol{A}$ se décompose alors selon $\boldsymbol{A} = QDQ^{-1}$ avec

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \beta & 0 & 0 \\ 0 & 0 & -\alpha\pi_R - \beta\pi_Y & 0 \\ 0 & 0 & 0 & -\alpha\pi_Y - \beta\pi_R \end{pmatrix}$$

et

$$Q = \begin{pmatrix} 1 & \pi_Y & \pi_G & 0 \\ 1 & \pi_Y & -\pi_A & 0 \\ 1 & -\pi_R & 0 & \pi_T \\ 1 & -\pi_R & 0 & -\pi_C \end{pmatrix}, \quad Q^{-1} = \begin{pmatrix} \pi_A & \pi_G & \pi_C & \pi_T \\ \pi_A/\pi_R & \pi_G/\pi_R & -\pi_C/\pi_Y & -\pi_T/\pi_Y \\ 1/\pi_R & -1/\pi_R & 0 & 0 \\ 0 & 0 & 1/\pi_Y & -1/\pi_Y \end{pmatrix}.$$

Le semi-groupe est donc donné par $P(t) = \exp(t\boldsymbol{A}) = Q\exp(tD)Q^{-1}$ et a pour expression

$$\boxed{P(t) = \begin{pmatrix} \pi_A\, a_1(t) + c_1(t) & \pi_G\, a_1(t) - c_1(t) & \pi_C\, b(t) & \pi_T\, b(t) \\ \pi_A\, a_1(t) - c_2(t) & \pi_G\, a_1(t) + c_2(t) & \pi_C\, b(t) & \pi_T\, b(t) \\ \pi_A\, b(t) & \pi_G\, b(t) & \pi_C\, a_2(t) + c_3(t) & \pi_T\, a_2(t) - c_3(t) \\ \pi_A\, b(t) & \pi_G\, b(t) & \pi_C\, a_2(t) - c_4(t) & \pi_T\, a_2(t) + c_4(t) \end{pmatrix}}$$

où

$$\boxed{\begin{array}{c} a_1(t) = 1 + \frac{\pi_Y}{\pi_R}\, e^{-\beta t}, \quad a_2(t) = 1 + \frac{\pi_R}{\pi_Y}\, e^{-\beta t}, \quad b(t) = 1 - e^{-\beta t}, \\ c_1(t) = \frac{\pi_G}{\pi_R}\, e^{-(\alpha\pi_R + \beta\pi_Y)t}, \quad c_2(t) = \frac{\pi_A}{\pi_R}\, e^{-(\alpha\pi_R + \beta\pi_Y)t}, \\ c_3(t) = \frac{\pi_T}{\pi_Y}\, e^{-(\alpha\pi_Y + \beta\pi_R)t}, \quad c_4(t) = \frac{\pi_C}{\pi_Y}\, e^{-(\alpha\pi_Y + \beta\pi_R)t}. \end{array}}$$

On constate que la chaîne de Markov $(X_t)_{t\geqslant 0}$ est ergodique de loi stationnaire $\boldsymbol{\pi} = (\pi_A, \pi_G, \pi_C, \pi_T)$:

$$\boxed{P(t) \underset{t\to+\infty}{\longrightarrow} \begin{pmatrix} \pi_A & \pi_G & \pi_C & \pi_T \\ \pi_A & \pi_G & \pi_C & \pi_T \\ \pi_A & \pi_G & \pi_C & \pi_T \\ \pi_A & \pi_G & \pi_C & \pi_T \end{pmatrix}.}$$

5.2 Probabilité de mutation

En fait la loi stationnaire π est réversible. En effet, en posant

$$D_\pi = \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_G & 0 & 0 \\ 0 & 0 & \pi_C & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix},$$

on vérifie que $D_\pi A D_\pi^{-1} = {}^tA$.

Suivons maintenant l'évolution de deux descendances $(X_t)_{t\geqslant 0}$ et $(Y_t)_{t\geqslant 0}$ issues du même ancêtre, indépendantes et de même loi et de loi initiale π. Considérons les probabilités

$$TS(t) = \mathbb{P}(X_t \longleftrightarrow Y_t \text{ ts}) \quad \text{et} \quad TV(t) = \mathbb{P}(X_t \longleftrightarrow Y_t \text{ tv}).$$

En introduisant le couplage $Z_t = (X_t, Y_t)$ et sa loi écrite sous forme d'une matrice carrée $S(t) = (s_{ab}(t))_{(a,b)\in\mathcal{A}\times\mathcal{A}}$ avec $s_{ab}(t) = \mathbb{P}(Z_t = (a,b))$, on a vu que $S(t) = D_\pi P(2t)$. Il vient de cette relation

$$\begin{aligned} TS(t) &= \sum_{a,b\in\mathcal{A}, a\leftrightarrow b \text{ ts}} s_{ab}(t) \\ &= \sum_{a,b\in\mathcal{A}, a\leftrightarrow b \text{ ts}} \pi_a p_{ab}(2t) \\ &= \pi_A p_{AG}(2t) + \pi_G p_{GA}(2t) + \pi_C p_{CT}(2t) + \pi_T p_{TC}(2t) \\ &= 2[\pi_A(\pi_G a_1(2t) - c_1(2t)) + \pi_C(\pi_T a_2(2t) - c_3(2t))], \end{aligned}$$

soit

$$\boxed{TS(t) = 2[\pi_A\pi_G + \pi_C\pi_T + (\pi_A\pi_G\tfrac{\pi_Y}{\pi_R} + \pi_C\pi_T\tfrac{\pi_R}{\pi_Y})e^{-2\beta t} - \tfrac{\pi_A\pi_G}{\pi_R}\, e^{-2(\alpha\pi_R+\beta\pi_Y)t} - \tfrac{\pi_C\pi_T}{\pi_Y}\, e^{-2(\alpha\pi_Y+\beta\pi_R)t}]}$$

ainsi que

$$\begin{aligned}
TV(t) &= \sum_{a,b\in\mathcal{A},a\leftrightarrow b \text{ tv}} s_{ab}(t) \\
&= \sum_{a,b\in\mathcal{A},a\leftrightarrow b \text{ tv}} \pi_a p_{ab}(2t) \\
&= \pi_A[p_{AC}(2t)+p_{AT}(2t)] + \pi_G[p_{GC}(2t)+p_{GT}(2t)] + \pi_C[p_{CA}(2t)+p_{CG}(2t)] + \pi_T[p_{TA}(2t)+p_{TG}(2t)] \\
&= 2\pi_A[p_{AC}(2t)+p_{AT}(2t)] + 2\pi_G[p_{GC}(2t)+p_{GT}(2t)] \\
&= 2[\pi_A(\pi_C+\pi_T)b(2t) + \pi_G(\pi_C+\pi_T)b(2t)],
\end{aligned}$$

soit

$$\boxed{TS(t) = 2\pi_R\pi_Y(1-e^{-2\beta t}).}$$

Ces expressions de $TS(t)$ et $TV(t)$ permettent de construire des estimateurs des paramètres α et β.