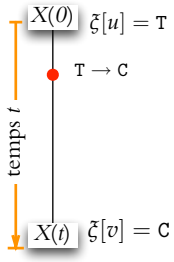
5. Modèles de dépendences



Procéssus de Markov à temps continu

Le PMTC est un procéssus $X(t)$ qui décrit les variables aléatoires au temps $t \in [0,\infty)$. Dans le contexte d'ADN, l'évolution entre le nœud parent u et l'enfant v suit un PMTC $X(t) \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$: on lance le procéssus avec $X(0) = \xi[u]$ (l'étiquette de u), et on met $\xi[v] = X(t)$ où t dénote la longueur de l'arête uv. On a alors la matrice de probabilités conditionnelles

$$\begin{aligned}\mathbf{M}(t) &= \left[\mathbb{P}\left\{\xi[v]=j \;\middle|\; \xi[u]=i\right\} : i,j=\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\right] && \textit{(substitutions } i\to j \textit{ sur arête } uv\textit{)}\\ &= \begin{bmatrix} p_{\mathtt{A}\to\mathtt{A}}(t) & p_{\mathtt{A}\to\mathtt{C}}(t) & p_{\mathtt{A}\to\mathtt{C}}(t) & p_{\mathtt{A}\to\mathtt{G}}(t)\\ p_{\mathtt{C}\to\mathtt{A}}(t) & p_{\mathtt{C}\to\mathtt{C}}(t) & p_{\mathtt{C}\to\mathtt{C}}(t) & p_{\mathtt{C}\to\mathtt{G}}(t)\\ p_{\mathtt{G}\to\mathtt{A}}(t) & p_{\mathtt{G}\to\mathtt{C}}(t) & p_{\mathtt{G}\to\mathtt{C}}(t) & p_{\mathtt{G}\to\mathtt{G}}(t)\\ p_{\mathtt{T}\to\mathtt{A}}(t) & p_{\mathtt{T}\to\mathtt{C}}(t) & p_{\mathtt{T}\to\mathtt{C}}(t) & p_{\mathtt{T}\to\mathtt{G}}(t)\end{bmatrix}\\ &= \exp(\mathbf{Q}t) && \textit{(}\mathbf{Q}\textit{ est la matrice de taux instantané)}\end{aligned}$$

où l'exponentiation[1] matricielle dénote la somme infinie

$$\exp(\mathbf{Q}t) = \begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} + \mathbf{Q}t + \frac{(\mathbf{Q}t)^2}{2} + \frac{(\mathbf{Q}t)^3}{3!} + \cdots = \sum_{k=0}^{\infty}\frac{(\mathbf{Q}t)^k}{k!}.$$

[1] à l'analogie de la série Taylor $e^{qt} = \sum_{k=0}^{\infty}\frac{(qt)^k}{k!}$

Modèle de Jukes-Cantor. La matrice de taux

$$\mathbf{Q} = \begin{bmatrix}-1&1/3&1/3&1/3\\1/3&-1&1/3&1/3\\1/3&1/3&-1&1/3\\1/3&1/3&1/3&-1\end{bmatrix} \quad \text{donne} \quad p_{i\to j}(t) = \begin{cases}\frac{1}{4}+\frac{3}{4}e^{-4t/3} & \{i=j\}\\ \frac{1}{4}-\frac{1}{4}e^{-4t/3} & \{i\neq j\}\end{cases}$$

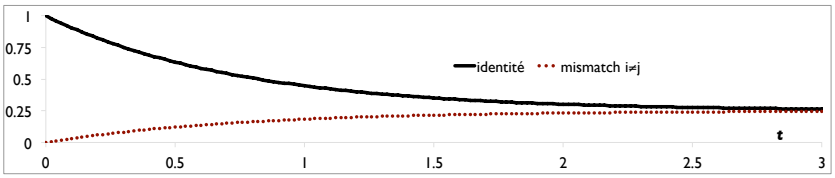


FIG. 1: Probabilités $p_{i\to j}(t)$ dans le modèle Jukes-Cantor.

Noter qu'à $t = 0$ (arête de longueur nulle), on a

$$\mathbb{P}\left\{X(0)=i \;\middle|\; X(0)=i\right\} = 1, \text{et}$$

$$\lim_{t\to\infty} p_{i\to j}(t) = \lim_{t\to\infty}\mathbb{P}\left\{X(t)=j \;\middle|\; X(0)=i\right\} = 1/4 \quad \text{pour tout } i,j=\mathtt{A},\mathtt{C},\mathtt{GT}.$$

Modèle F84/HKY. Le modèle de Felsenstein (1984) ou Hasegawa-Kishino-Yano (1985) s'écrit avec une matrice de taux

$$\mathbf{Q} = \begin{bmatrix} \cdot & \pi_{\mathtt{C}}\alpha & \pi_{\mathtt{G}}\beta & \pi_{\mathtt{T}}\alpha \\ \pi_{\mathtt{A}}\alpha & \cdot & \pi_{\mathtt{G}}\alpha & \pi_{\mathtt{T}}\beta \\ \pi_{\mathtt{A}}\beta & \pi_{\mathtt{C}}\alpha & \cdot & \pi_{\mathtt{T}}\alpha \\ \pi_{\mathtt{A}}\alpha & \pi_{\mathtt{C}}\beta & \pi_{\mathtt{G}}\alpha & \cdot \end{bmatrix} \qquad \textit{(somme est 0 dans toute rangée)}$$

définit mène à la composition stationnaire $\lim_{t\to\infty} \mathbb{P}\{X(t) = i\} = \pi_i$ et un rapport de taux transition :transversion $\kappa = \frac{\beta}{\alpha}$.

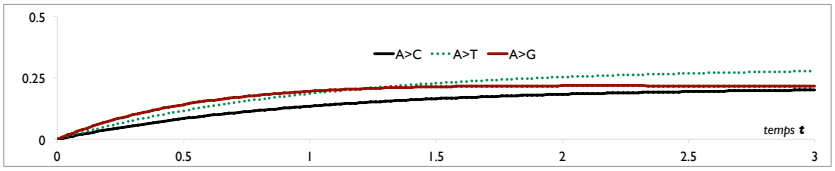


FIG. 3: Probabilités $p_{i\to j}(t)$ dans le modèle F84 pour GC% $= 42\%$ ($\pi_{\mathtt{A}} = pi_{\mathtt{T}} = 0.29$, $\pi_{\mathtt{C}} = \pi_{\mathtt{G}} = 0.21$) et $\kappa = 2$. On a $p_{\mathtt{A}\to\mathtt{C}} = 0.21\big(1 - e^{-t/0.9744}\big)$, $p_{\mathtt{A}\to\mathtt{T}} = 0.29\big(1 - e^{-t/0.9744}\big)$, et $p_{\mathtt{A}\to\mathtt{G}} = 0.21\big(1 + e^{-t/0.9744} - 2e^{-t/0.6496}\big)$.

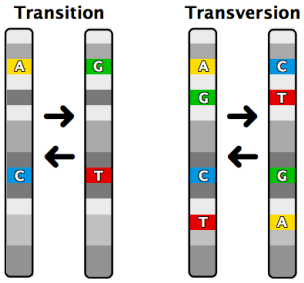


FIG. 2: Transitions et transversions

Multiplication. On a que $\mathbf{M}(t) \cdot \mathbf{M}(s) = \mathbf{M}(t+s)$ pour tout $s, t \geq 0$, et que $\boldsymbol{\pi} \cdot \mathbf{M}(t) = \boldsymbol{\pi}$ avec la distribution stationnaire $\boldsymbol{\pi}$.

Segmentation

Segmentation par blocs maximaux. On a une séquence de *scores* $a[1..\ell]$ avec $a[j] \in \mathbb{R}$ (peut être négatif ou positif). On pénalise un bloc $j..j+k-1$ par la somme $a[j] + a[j+1] + \cdots + a[j+k-1]$. Bloc de score maximal par programmation dynamique : $S \leftarrow 0$, et faire $S \leftarrow \max\{0, S + a[j]\}$ avec $j = 1, 2, \ldots, \ell$.

p.e., $a[k] =$ LODS pour deux hypothèses.

Modèle de Markov caché. Chaque colonne $k = 1, \ldots, \ell$ correspond à un état caché $q[k] \in \mathcal{Q}$: par exemple, $\mathcal{Q} = \{\text{neutre}, \text{conservé}\}$. Émission = résidues alignées $x[k] = (x_1[k] \cdots x_n[k])$ dans la colonne. Le modèle phylogénétique définit la probabilité $p(i, x[k])$ pour la colonne est émise dans l'état i. La séquence d'états $q[1..\ell]$ est au hasard avec $\tau(i,j) = \mathbb{P}\Big\{q[k+1] = j \;\Big|\; q[j] = i\Big\}$. Quel est la *séquence* d'états la plus probable ou le **chemin Viterbi** ? Pour cela on définit la vraisemblance du meilleur préfixe, ce qu'on peut calculer par programmation dynamique :

$$\delta_1(i) = \pi(i) \cdot p\big(i, x[1]\big) \tag{5.1a}$$

$$\delta_k(i) = \Big(\max_{j\in\mathcal{Q}} \delta_{k-1}(j) \cdot \tau(j,i)\Big) \cdot p\big(i, x[k]\big) \qquad \{k > 0\} \tag{5.1b}$$

On reconstruit le *chemin Viterbi* qui correspond à $\max_i \delta_\ell(i)$ par *backtracking* : il faut stocker le meilleur état qui précède j, ainsi que $\delta_k(i)$ dans la récurrence de (5.1b).