# MODÈLES D'ÉVOLUTION DE SÉQUENCES : EXERCICES

DIDIER PIAU – ALÉA 7-12 mars 2011

## Distances dans le modèle de Jukes-Cantor

On considère le modèle de Jukes-Cantor de paramètre $\alpha$ sur l'alphabet des nucléotides

$$\mathcal{A} = \{A, C, G, T\}.$$

**1.** Exprimer la Q-matrice $Q$ de ce modèle en fonction des deux matrices de transition suivantes : l'identité $I$ et la matrice $J$ dont tous les coefficients sont égaux à $\frac{1}{4}$.

**2.** Montrer que $\mathrm{Vect}(I, J)$ est une sous-algèbre de l'algèbre des matrices $4 \times 4$ à coefficients réels. On pourra commencer par dresser la table de multiplication de $\{I, J\}$.

**3.** Pour tout nombre réel $t \geqslant 0$, exprimer la matrice $M(t) = \mathrm{e}^{tQ}$ comme une combinaison linéaire $c(t)I + a(t)J$. Indication : on pourra considérer la matrice dérivée $M'(t)$ et en déduire un système différentiel linéaire dont $(a(t), c(t))$ est solution.

**4.** On considère deux séquences nucléotidiques $X$ et $X'$ de même longueur $n$ et alignées. Exprimer la log-vraisemblance $L$ de l'évènement : « $X'$ est le résultat d'une évolution selon la dynamique de Jukes-Cantor de paramètre $\alpha$ appliquée à $X$ pendant le temps $t$. » On écrira le résultat comme une fonction de $n$, $\alpha$, $t$ et de la proportion $D$ de sites ne coïncidant pas dans $X$ et $X'$ (on dit que $D$ est une statistique suffisante de $X$ et $X'$).

**5.** Calculer la valeur de $t$ qui maximise $L$.

**6.** Préciser combien de substitutions par unité de temps un site donné d'une séquence à l'équilibre subit en moyenne.

**7.** Déduire de tout ceci l'estimateur du maximum de vraisemblance $T$ du temps écoulé entre $X$ et $X'$ rapporté à l'échelle temporelle correspondant à 1 substitution en moyenne par unité de temps. Indication : $T = \tau(D)$ pour une certaine fonction $\tau$.

**8.** (Pour les enragés) Montrer que le comportement asymptotique de $T$ quand $n \to \infty$ est de la forme

$$T \approx \tau(D) + \frac{1}{\sqrt{n}}\sigma(D)\mathcal{N}, \qquad \sigma(D) = \frac{3D(1-D)}{3-4D}$$

où $\mathcal{N}$ désigne une variable aléatoire gaussienne centrée réduite. Commenter.

**Distances dans le modèle de Kimura**

On considère le modèle de Kimura à deux paramètres. On note $\alpha$ le taux de substitution pour les transitions $A \leftrightarrow G$ and $C \leftrightarrow T$ et $\beta$ le taux de substitution pour les transversions ($A$ ou $G$) $\leftrightarrow$ ($C$ ou $T$).

**1.** Exprimer la Q-matrice $Q$ de ce modèle en fonction des trois matrices de transition suivantes : l'identité $I$, la matrice $H$ dont les seuls coefficients non nuls sont $H(A,G)$, $H(G,A)$, $H(C,T)$ et $H(T,C)$, et la matrice $F$ dont tous les coefficients $F(x,y)$ pour $x$ dans $\{A,G\}$ et $y$ dans $\{C,T\}$ et vice versa sont égaux à $\frac{1}{2}$. On pourra commencer par écrire $H$ et $F$ de façon plus explicite que ci-dessus.

**2.** Montrer que $\mathrm{Vect}(I,H,F)$ est une sous-algèbre de l'algèbre des matrices $4 \times 4$ à coefficients réels. On pourra commencer par dresser la table de multiplication de $\{I,H,F\}$.

**3.** Pour tout nombre réel $t \geqslant 0$, exprimer la matrice $M(t) = \mathrm{e}^{tQ}$ comme une combinaison linéaire $c(t)I + a(t)H + b(t)F$. Indication : on pourra considérer la matrice dérivée $M'(t)$ et en déduire un système différentiel linéaire dont $(a(t), b(t), c(t))$ est solution.

**4.** On considère deux séquences nucléotidiques $X$ et $X'$ de même longueur $n$ et alignées. Exprimer la log-vraisemblance $L$ de l'évènement : « $X'$ est le résultat d'une évolution selon la dynamique de Kimura de paramètres $\alpha$ et $\beta$ appliquée à $X$ pendant le temps $t$. » On écrira le résultat comme une fonction de $n$, $\alpha$, $\beta$, $t$ et des proportions $D_T$ et $D_V$ de sites correspondant à une transition, respectivement à une transversion, quand on passe de $X$ à $X'$ (on dit que $(D_T, D_V)$ est une statistique suffisante de $X$ et $X'$).

**5.** Calculer la valeur de $(\alpha t, \beta t)$ qui maximise $L$.

**6.** Préciser combien de substitutions par unité de temps un site donné d'une séquence à l'équilibre subit en moyenne.

**7.** Déduire de tout ceci l'estimateur du maximum de vraisemblance $T_K$ du temps écoulé entre $X$ et $X'$ rapporté à l'échelle temporelle correspondant à 1 substitution en moyenne par unité de temps. Indication : $T_K = \tau_K(D_T, D_V)$ pour une certaine fonction $\tau_K$.

On vérifiera que $\tau_K(\delta, 2\delta) = \tau(3\delta)$ pour tout $\delta$ entre 0 et $\frac{1}{4}$ et on expliquera pourquoi.

**8.** (Pour les enragés) Préciser si les estimateurs $T$ et $T_K$ sont biaisés ou non, et s'ils le sont, dans quel sens.