3. Substitutions

Modèle iid. Modèle probabiliste pour une séquence : caractères aléatoires iid[1] : $\mathbb{P}\{S[i] = a\} = \pi_a$. Par conséquence, le nombre d'occurrences $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ détermine la probabilité d'une séquence, sans égard à l'ordre des nucléotides[2]

[1] iid=indépendants et identiquement distribués

[2] Exemple : $\pi_{\mathtt{A}} = \pi_{\mathtt{T}} = 32\%$, $\pi_{\mathtt{C}} = \pi_{\mathtt{G}} = 18\%$, ou taux de $(\mathtt{G} + \mathtt{C})$ à 36%. On a alors $\mathbb{P}\{S = \mathtt{ACAT}\} = \mathbb{P}\{S = \mathtt{CTAA}\} = 0.32^3 \times 0.18 = 0.00589\cdots$

Pondération de substitutions

Supposons qu'on a deux séquences ADN $s[0..\ell-1]$ et $t[0..\ell-1]$, qui correspondent à une région sans trous dans un alignement. On veut décider s'il y a une dépendance entre les résidus alignés. On compare deux modèles paramétriques, correspondant à hypothèses distincts. Les modèles déterminent la distribution de deux séquences aléatoires $S[0..\ell-1]$ et $T[0..\ell-1]$, comprenants des paires iid $(S[i], T[i])$. Ici, S et T sont des variables aléatoires. On obsèrve s et t : les modèles définissents $\mathbb{P}\{S = s; T = t\}$.

Hypothèse 0 : S et T sont indépendantes. On définit

$$L_0 = \mathbb{P}\Big\{S = s; T = t \;\Big|\; H_0\Big\} = \prod_{i=0}^{\ell-1} \pi_{s[i]} \cdot \pi_{t[i]}. \tag{3.1a}$$

Hypothèse 1 : S et T sont dépendantes :

$$L_1 = \mathbb{P}\Big\{S = s; T = t \;\Big|\; H_1\Big\} = \prod_{i=0}^{\ell-1} \pi_{s[i]} \cdot \mathbb{P}\Big\{T[i] = t[i] \;\Big|\; S[i] = s[i]\Big\}. \tag{3.1b}$$

Les quantités L_0, L_1 mesurent le support aux données par les hypothèses. En général, $\mathbb{P}\Big\{\text{données} \;\Big|\; \text{hypothèse}\Big\}$ s'appelle la **vraisemblance**[3]

[3] Noter que c'est tout à fait différent de $\mathbb{P}\Big\{\text{hypothèse} \;\Big|\; \text{données}\Big\}$.

Matrice de substitution. La **matrice de substitution M** se définit par la distribution conditionnelle

$$\mathbf{M}_{a,b} = \mathbb{P}\Big\{T[i] = b \;\Big|\; S[i] = a\Big\}$$

(à n'importe quelle position i selon la supposition iid). C'est une **matrice stochastique** de taille 4×4 : $0 \le \mathbf{M}_{a,b} \le 1$ pour tout $a, b = \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$, ainsi que $\sum_b \mathbf{M}_{a,b} = 1$ pour tout a.

Rapport des chances. Le rapport des vraisemblances de (3.1) s'appelle le **rapport des chances** (*odds ratio*). Le log-*odds-ratio* quantifie l'évidence pour hypothèse 1 par rapport à hypothèse 0 sur une échelle logarithmique[4]

$$\mathsf{LODS} = \log \frac{L_1}{L_0}. \tag{3.2}$$

[4] On mesure l'évidence en *bits* par $\lg = \log_2$, en *nats* par $\ln = \log_e$, et en *ban* par $\log_{10}$. L'échelle $10 \log_{10} = \log_{1.2589\ldots}$ («Phred scale» ou deciban) est utilisé en séquençage.

Par (3.1a) et (3.1b), on a

$$\mathsf{LODS} = \log \frac{L_1}{L_0} = \log \frac{\prod_{i=0}^{\ell-1} \pi_{s[i]} \cdot \mathbf{M}_{s[i],t[i]}}{\prod_{i=0}^{\ell-1} \pi_{s[i]} \cdot \pi_{t[i]}} = \sum_{i=0}^{\ell-1} \log \frac{\mathbf{M}_{s[i],t[i]}}{\pi_{t[i]}}.$$

Matrice de pondération. On définit la pondération de paires (a,b) alignées par la valeur arrondie

$$C_{a,b} = \left\lceil \log_\alpha \frac{\mathbf{M}_{a,b}}{\pi_b} \right\rceil,$$

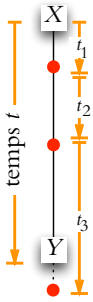
sur une échelle logarithmique quelconque $\alpha > 1$. L'évidence pour hypothèse 1 est la somme de poids : $\mathsf{LODS} = \sum_{i=0}^{\ell-1} \mathbf{C}_{s[i],t[i]}$.

Substitutions multiples. Évolution de caractères homologues est sans mémoire : états successifs forment une chaîne de Markov. Probabilités de substitution selon la matrice $\mathbf{M}_{X\to Y}$ où $\mathbf{M}_{X\to Y}[a,b] = \mathbb{P}\Big\{Y = b \;\Big|\; X = a\Big\}$ avec valeurs $a,b \in \Sigma = \{1,\dots,r\}$. (Exemple : $\Sigma = \{1,2,3,4\}$ encodant ADN)

$$\begin{aligned}
\mathbb{P}\Big\{Z = z \;\Big|\; X = x\Big\} &= \sum_{y=1}^{r} \mathbb{P}\Big\{Z = z, Y = y \;\Big|\; X = x\Big\} \\
&= \sum_{y=1}^{r} \mathbb{P}\Big\{Z = z \;\Big|\; X = x, Y = y\Big\}\mathbb{P}\Big\{Y = y \;\Big|\; X = x\Big\} \\
&= \sum_{y=1}^{r} \mathbb{P}\Big\{Z = z \;\Big|\; Y = y\Big\}\mathbb{P}\Big\{Y = y \;\Big|\; X = x\Big\} \qquad \textit{(propriété de Markov)}
\end{aligned}$$

Donc, on peut écrire que $\mathbf{M}_{X\to Z} = \mathbf{M}_{X\to Y} \cdot \mathbf{M}_{Y\to Z}$

Temps continu. À chaque mutation, le changement d'état est determiné par la matrice $\mathbf{M}$. Si k événements en temps t, alors $\mathbf{M}_{X\to Y} = \mathbf{M}^k$. Donc la matrice pour temps t devrait être $\mathbf{M}^{N(t)}$ où $N(t)$ est le nombre de mutations pendant temps t. Problème : $N(t)$ est aléatoire...



☞ **[«orderly»]** La probabilité qu'il y ait plus d'une occurrence dans un petit intervalle de temps est négligeable

$$\lim_{\delta\to 0} \mathbb{P}\Big\{N(t+\delta) > 1 \;\Big|\; N(t+\delta) \geq 1\Big\} = 0$$

☞ **[«memoryless»]** Le nombre d'occurrences pendant un intervalle ne dépend pas des événements précédents : $N(t+s) - N(t)$ est indépendant de $N(t)$

Procéssus de Poisson. Si le procéssus est *orderly* et *memoryless*, alors c'est un **procéssus de Poisson** avec les propriétés suivantes.

- ⋆ nombre d'arrivées pendant temps t est une v.a. Poisson : $\mathbb{P}\{N(t+s) - N(s) = k\} = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$. Le paramètre λ est l'intensité ou le **taux** du procéssus.
- ⋆ temps d'attente est une v.a. exponentielle : $\forall t,s$: $\mathbb{P}\{N(s+t) = N(s)\} = e^{-\lambda t}$

Procéssus de Markov. En retournant à ce qui se passe pendant temps t :

$$\sum_{k=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!}\mathbf{M}^k = \exp\Big(\lambda t(\mathbf{M} - \mathbf{I})\Big) = \exp\Big(\lambda t\mathbf{Q}\Big)$$

où $\mathbf{I}$ est la matrice d'unité. $\mathbf{Q} = \mathbf{M} - \mathbf{I}$ est la *matrice instantanée de taux de substitutions*. Calcul de $\exp(\lambda\mathbf{Q}t)$: décomposition de la matrice $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$, et faire $e^{\lambda\mathbf{Q}t} = \mathbf{U}e^{\lambda t\mathbf{\Lambda}}\mathbf{V}$.