2 Modèles de séquences

Modèle iid. Modèle probabiliste pour une séquence : caractères aléatoires iid (indépendants et identiquement distribués) : $\mathbb{P}\{S[i] = a\} = \pi_a$. Exemple : $\pi_{\mathtt{A}} = \pi_{\mathtt{T}} = 32\%$, $\pi_{\mathtt{C}} = \pi_{\mathtt{G}} = 18\%$, ou taux de $(\mathtt{G} + \mathtt{C})$ à 36%. Par conséquence, le nombre d'occurrences $\mathtt{A}, \mathtt{C}, \mathtt{GT}$ détermine la probabilité d'une séquence, sans égard à l'ordre des nucléotides. On a alors $\mathbb{P}\{S = \mathtt{ACAT}\} = \mathbb{P}\{S = \mathtt{GTTT}\} = 0.32^3 \times 0.18 = 0.00589\cdots$

2.1 Pondération de substitutions

Supposons qu'on a deux séquences ADN $s[0..\ell-1]$ et $t[0..\ell-1]$, aui correspondent à une région sans trous dans un alignement. On veut décider s'il y a une dépendence 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 obsève s et t : les modèles définissents $\mathbb{P}\{S = s; T = t\}$.

Hypothèse 0 : S et T sont indépendantes. On a

$$L_0 = \mathbb{P}\Big\{S = s; T = t \;\Big|\; H_0\Big\} = \prod_{i=0}^{\ell-1} \pi_{s[i]} \cdot \pi_{t[i]}. \tag{2.1a}$$

Hypothèse 1 : S et T sont dépendantes selon une distribution jointe connue. Alors

$$L_1 = \mathbb{P}\Big\{S = s; T = t \;\Big|\; H_1\Big\} = \prod_{i=0}^{\ell-1} \pi_{s[i]} \cdot \mathbb{P}\Big\{S[i] = s[i] \;\Big|\; T[i] = t[i]\Big\}. \tag{2.1b}$$

Les quantités L_i mesurent le support aux données par des hypothèses incompatibles : en général, $\mathbb{P}\Big\{\text{données} \;\Big|\; \text{hypothèse}\Big\}$ s'appelle la **vraisemblance**.

Matrice de substitution. La **matrice de substitution** $\mathbf{M}$ se définit par la distribution conditionnelle W(en)

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

(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 (2.1) 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 : W(fr)

$$\mathsf{LODS} = \log \frac{L_1}{L_0}. \tag{2.2}$$

REMARQUE. On mesure l'évidence en *bits* par $\log_2$, en *nats* par $\log_e$, et en *ban* par $\log_{10}$.

Par (2.1a) et (2.1b), 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

$$C_{a,b} = \log_\alpha \frac{\mathbf{M}_{a,b}}{\pi_b}, \tag{2.3}$$

sur une échelle logarithmique quelconque $\alpha > 1$. En conséquence, l'évidence pour hypothèse 1 est la somme de poids de résidus alignés :

$$\mathsf{LODS} = \sum_{i=0}^{\ell-1} \mathbf{C}_{s[i],t[i]}. \tag{2.4}$$

.

Exemple 2.1. On considère la distribution $\pi_{\mathtt{A}} = \pi_{\mathtt{T}} = 32\%$, $\pi_{\mathtt{C}} = \pi_{\mathtt{G}} = 18\%$ avec la matrice de transition

$$\mathbf{M} = \begin{bmatrix} 0.914 & 0.018 & 0.036 & 0.032 \\ 0.032 & 0.886 & 0.018 & 0.064 \\ 0.064 & 0.018 & 0.886 & 0.064 \\ 0.032 & 0.036 & 0.018 & 0.914 \end{bmatrix} \qquad \textit{(ordre : } \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\textit{)} \tag{2.5}$$

On choisit une pondération par l'échelle $\log_\phi$ avec $\phi = 10^{0.1} = 1.258\cdots$. (Ainsi, on mesure en *decibans* : $\log_\phi = 10 \log_{10}$.) Après arrondi entier, on a les poids

$$\mathbf{C} = \begin{bmatrix} 5 & -10 & -7 & -10 \\ -10 & 7 & -10 & -7 \\ -7 & -10 & 7 & -10 \\ -10 & -7 & -10 & 5 \end{bmatrix} \tag{2.6}$$

Dans d'autres mots, la distribution jointe $(\pi, \mathbf{M})$ assumée par hypothèse 1, correspond à la pondération de (2.6). Et *vice versa* : la pondération de (2.6) assume implicitement la distribution $(\pi, \mathbf{M})$ de (2.5). ♠

2.2 Construction de matrices de pondération

Équation (2.3) reste valide avec n'importe quel alphabet, montrant comment choisir la pondération qui correspond à une distribution jointe connue. Mais d'ou'vient la distribution connue ? On les estime à partir d'un *échantillon* d'alignements. Les matrices de pondération populaires employées par des logiciels d'alignement, telles que les matrices BLOSUMxx et PAMxxx ont été construites à partir des alignements de haute qualité entre protéines homologues.

Matrice à partir d'un alignement. On considère un alignement de deux séquences $s[0..\ell-1]$ et $t[0..\ell-1]$ sans trous. Pour estimer π et $\mathbf{M}$ d'une distribution jointe de résidus alignés, on calcule les fréquences rélatives

$$f_{a,b} = \frac{\sum_{i=0}^{\ell-1} \{s[i] = a, t[i] = b\}}{\ell}$$

pour toute paire $a, b \in \mathcal{A}$ de résidus. Dans le cas de séquences protéiques, on a un alphabet à 20 lettres

$$\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{K}, \mathtt{L}, \mathtt{M}, \mathtt{N}, \mathtt{P}, \mathtt{Q}, \mathtt{R}, \mathtt{S}, \mathtt{T}, \mathtt{V}, \mathtt{W}, \mathtt{Y}\}.$$

On définit les probabilités empiriques

$$p_b = \frac{\sum_{c \in \mathcal{A}} f_{c,b}}{\sum_{c,d \in \mathcal{A}} f_{c,d}} \qquad q_{a,b} = \frac{f_{a,b}}{\sum_{d \in \mathcal{A}} f_{a,d}}. \tag{2.7}$$

Ici, p_b estime π_b. $q_{a,b}$ est notre estimation de $\mathbf{M}_{a,b}$. La pondération adéquate selon (2.3) est

$$\mathbf{C}_{a,b} = \log \frac{q_{a,b}}{p_b} = \log \frac{f_{a,b} \times \sum_{c,d} f_{c,d}}{\left(\sum_c f_{c,b}\right) \times \left(\sum_d f_{a,d}\right)}.$$

W(en)

Matrices BLOSUM. Steven et Jorja Henikoff [«Amino acid substitution matrices from protein blocks», *PNAS*, **89** :10915–10919, 1991] ont calculé la pondération LODS à partir des alignements de haute qualité comprenant des blocs sans trous. Ils arguent qu'une distribution jointe devrait être estimée à partir des paires de séquences de similarité comparable. Dans le but de filtrer la contribution de paralogues proches, on choisit un seuil de similarité β (pourcentage de résidus identiques comme $\beta = 85\%$) qui partage les séquences de chaque bloc en groupes. Un bloc comprend les séquences $s_1[0..\ell-1], s_2[0..\ell-1], \ldots, s_n[0..\ell-1]$. On met s_i et s_j dans le même groupe si $\rho(i,j) = \sum_{k=0}^{\ell-1}\{s_i[k] = s_j[k]\} \geq \beta\ell$ (identitité en au moins β pourcentage des positions). Ainsi, une partition en G groupes est définie par une vecteur $\phi_\beta[1..n]$: séquence i appartient à groupe $\phi_\beta[i] \in \{1, 2, \ldots, G\}$. Notez que $\phi_\beta[i] = \phi_\beta[j]$ est possible même si $\rho(i,j) < \alpha\ell$: par exemple, s'il existe k tel que $\rho(i,k) \geq \beta\ell$ et $\rho(j,k) \geq \beta\ell$. Par contre, si $\phi_\beta[i] \neq \phi_\beta[j]$, on est certain que les séquences s_i et s_j diffèrent en au moins $(1-\beta)\ell$ positions. Comme avant, on calcule les fréquences rélatives de paires observés entre séquences différentes $f_{a,b}(i,j) = \frac{1}{\ell}\sum_{k=0}^{\ell-1}\{s_i[k] = a; s_j[k] = b\}$, mais cela ne nous intéresse que si les séquences appartiennent à groupes différents ($\phi_\beta[i] \neq \phi_\beta[j]$). Pour toute paire de groupes $1 \leq g < g' \leq G$, on considère la fréquence symétrisée $\ddot{f}$ de résidus, après normalisation par la taille de groupes $n_g = \sum_{i=1}^{n}\{\phi_\beta[i] = g\}$:

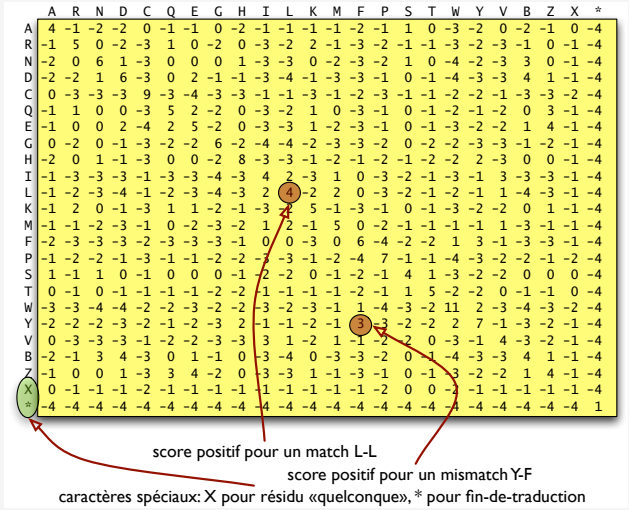
$$\ddot{f}_{a,b}(g,g') = \frac{\sum_i\{\phi_\beta[i] = g\}\sum_j\{\phi_\beta[j] = g'\}\big(f_{a,b}(i,j) + \{a \neq b\}f_{b,a}(i,j)\big)}{n_g \cdot n'_g}.$$

(On compte les paires non-ordonnées $a \neq b$, et identités $a = b$.) L'estimateur de probabilités de transition $\mathbf{M}$ considère toutes les paires de groupes :

$$q_{a,b} = \frac{\sum_{g=1}^{G-1}\sum_{g'=g+1}^{G}\ddot{f}_{a,b}(g,g')}{G(G-1)/2}.$$

On calcule aussi la probabilité empirique de toute résidu $a \in \mathcal{A}$:

$$p_a = \frac{\sum_{i=1}^{n}\sum_{k=0}^{\ell-1}\{s_i[k] = a\}}{n\ell}.$$



Les choix $\beta = 45\%, 62\%, 80\%$ correspondent aux matrices BLOSUM45, BLOSUM62 (illustrée à la gauche), BLOSUM80, dans cet ordre. La pondération est en mi-bits ($2\log_2 x = \log_{\sqrt{2}} x$) oui tiers-bits ($3\lg x$), arrondi à l'entier :

$$\mathbf{C}_{a,b} = \mathbf{C}_{b,a} = c\lg\frac{q_{a,b}}{2p_ap_b} \qquad \{a \neq b\}$$
$$\mathbf{C}_{a,a} = c\lg\frac{q_{a,a}}{p_a^2},$$

où $c = 2$ pour $\beta = 0.8, 0.62$ et $c = 3$ pour $\beta = 0.45$.