4. Variables inobservées

Génomique comparative. Principe de génomique comparative : sélection négative ralentit le taux d'évolution ⇒ plus de conservation de séquence à un élément fonctionnel.

Inférence par logarithme des chances à partir d'un alignement : hypothèse 0 = substitutions par $e^{\mu \mathbf{Q} t}$, hypothèse 1 = substitutions par $e^{\rho\mu \mathbf{Q} t}$, on connaît μt, $\mathbf{Q}$, $\rho < 1$. Alors, on calcule les vraisemblances L_0, L_1 selon les hypothèses et utilise $\mathsf{LLR} = \log \frac{L_1}{L_0}$ pour détecter l'**ombre** de sélection.

Avec plus de séquences, on peut détecter la conservation mieux... mais comment calculer L_i ?

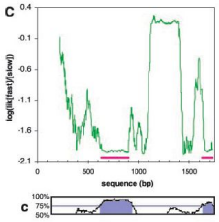


FIG. 1: **Phylogenetic shadowing** illustré par Boffelli & al. *Science* 299 :1391 (2003). En haut : LLR dans un fenêtre de 50 pb glissant dans l'alignement multiple de 5 primates, avec des exons indiqués. En bas : % identité humain-souris.

Vraisemblance sur une phylogénie

Modèle markovien pour étiquetage (v.a. $\xi[u] \in \mathcal{X}$ à tout nœud u) :
(1) distribution à la racine par $\pi_x = \mathbb{P}\{\xi[\mathsf{racine}] = x\}$;
(2) probabilités de changement d'étiquette sur arête uv

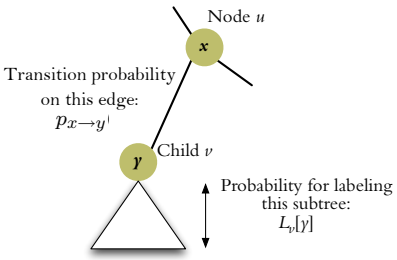
$$p_{x\to y}(uv) = \mathbb{P}\Big\{\xi[v] = y \;\Big|\; \xi[u] = x\Big\}$$

Vraisemblance pour un étiquetage x des nœuds terminaux — il faut considérer tous les étiquetages de tous les nœuds

$$L[x] = \sum_{\tilde{x}\text{: étiquettes possibles aux nœuds}} \pi_{\tilde{x}[\mathsf{racine}]} \prod_{\text{arêtes } uv} p_{\tilde{x}[u]\to\tilde{x}[v]}(uv).$$

Calcul efficace par programmation dynamique (algorithme de Felsenstein — *peeling*)

Idée : définir la vraisemblance conditionnelle $L_u[x]$ pour tout caractère $x \in \mathcal{X}$: probabilité pour étiquettes dans le sous-arbre de u quand $\xi[u] = x$.



$$L_u[x] = \{\xi[u] = x\} \qquad u \text{ terminal}$$
$$L_u[x] = \prod_{v\in\mathsf{enfants}(u)} \Big(\sum_{y\in\mathcal{X}} p_{x\to y}(uv) L_v[y]\Big) \qquad u \text{ ancestral}$$

Vraisemblance et modèles graphiques

On peut généraliser l'approche à des graphes orientés. Le modèle probabiliste comprend un ensemble de variables aléatoires inobservées $Y_1, \ldots, Y_m$ et l'obsérvation Y_0 même :

$$p(x|\theta) = \sum_{y_1,y_2,\ldots,y_m} \mathbb{P}\Big\{Y_0 = x \;\Big|\; Y_1 = y_1, \ldots, Y_m = y_m; \theta\Big\} \cdot \mathbb{P}\Big\{Y_1 = y_1, \ldots, Y_m = y_m \;\Big|\; \theta\Big\}.$$

Ici, θ représente les paramètres du modèle. La somme suggère une temps exponentiel en m, mais on peut souvent exploiter les indépendances conditionnelles entre les Y_i pour accélerer le calcul. Considérons le graphe G de

dépendances entres les variables aléatoires. Les nœuds de G correspondent aux variables $Y_i : i = 0, 1, \ldots, m$. Les arcs de G sont orientés. L'ensemble des arcs entrants au nœud j, dénoté par $\mathsf{pred}(j) = \{i : ij \in G\}$, définit l'indépendance conditionnelle :

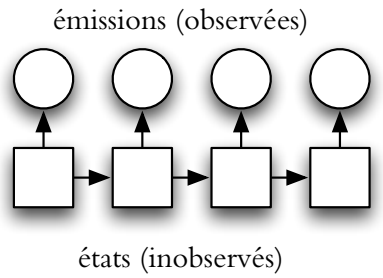
$$\mathbb{P}\Big\{Y_j \;\Big|\; Y_i : i \neq j\Big\} = \mathbb{P}\Big\{Y_j \;\Big|\; Y_i : \in \mathsf{pred}(j)\Big\}.$$

Un *modèle graphique* est un modèle probabiliste dont le graphe de dépendances ne contient aucun cycle.

Modèle de Markov caché

Un modèle de Markov caché (*hidden Markov model* — HMM) sur un alphabet $\mathcal{A}$ comprend les paramètres suivants.

- ⋆ ensemble d'états $\mathcal{Q}$ de taille $|\mathcal{Q}| = N$
- ⋆ probabilités de transition entre états : $\tau : \mathcal{Q} \times \mathcal{Q} \to [0, 1]$
- ⋆ probabilités d'émission à chaque état : $p : \mathcal{Q} \times \mathcal{A} \to [0, 1]$
- ⋆ probabilités d'état initial : $\pi : \mathcal{Q} \to [0, 1]$



Le modèle $M = (\mathcal{Q}, \tau, p, \pi)$ définit une distribution sur séquences d'une longueur fixe. En particulier, M produit une séquence $X[0..\ell-1]$ selon la procédure aléatoire suivante.

> G1 choisir l'état initial $q[0]$ au hasard selon les probabilités π
> G2 **for** $i \leftarrow 0, 1, \ldots, \ell - 1$
> G3 choisir $X[i]$ au hasard par les probabilités d'émission $p(q[i], \cdot)$
> G4 choisir prochain état $q[i+1]$ au hasard par les probabilités de transition $p(q[i], \cdot)$

Vraisemblance. La vraisemblance selon M se définit par

$$\begin{aligned} L(M) = \mathbb{P}\Big\{X = x \;\Big|\; M\Big\} = \\ \sum_{q[0..\ell-1] \in \mathcal{Q}^\ell} & \Big(\pi\big(q[0]\big) \cdot p\big(q[0], x[0]\big) && \textit{(1er état+symbole)} \\ & \times \tau\big(q[0], q[1]\big) \cdot p\big(q[1], x[1]\big) \times \cdots && \textit{(2e état+symbole)} \\ & \times \tau\big(q[\ell-2], q[\ell-1]\big) \cdot p\big(q[\ell-1], x[\ell-1]\big)\Big). \end{aligned} \tag{4.1}$$

Probabilité de préfixe. On définit la vraisemblance sur les préfixes (*forward probabilities*)

$$\alpha_k(i) = \sum_{q[0..k]; q[k]=i} \mathbb{P}\Big\{X[0..k] = x[0..k] \;\Big|\; M\Big\},$$

ce qu'on peut calculer par programmation dynamique.

$$\alpha_0(q) = \pi(q) \tag{4.2a}$$

$$\alpha_k(i) = \Big(\sum_{j \in \mathcal{Q}} \alpha_{k-1}(j) \cdot \tau(j, i)\Big) \cdot p\big(i, x[k]\big) \qquad \{k > 0\} \tag{4.2b}$$