Quelques modèles probabilistes en biologie moléculaire

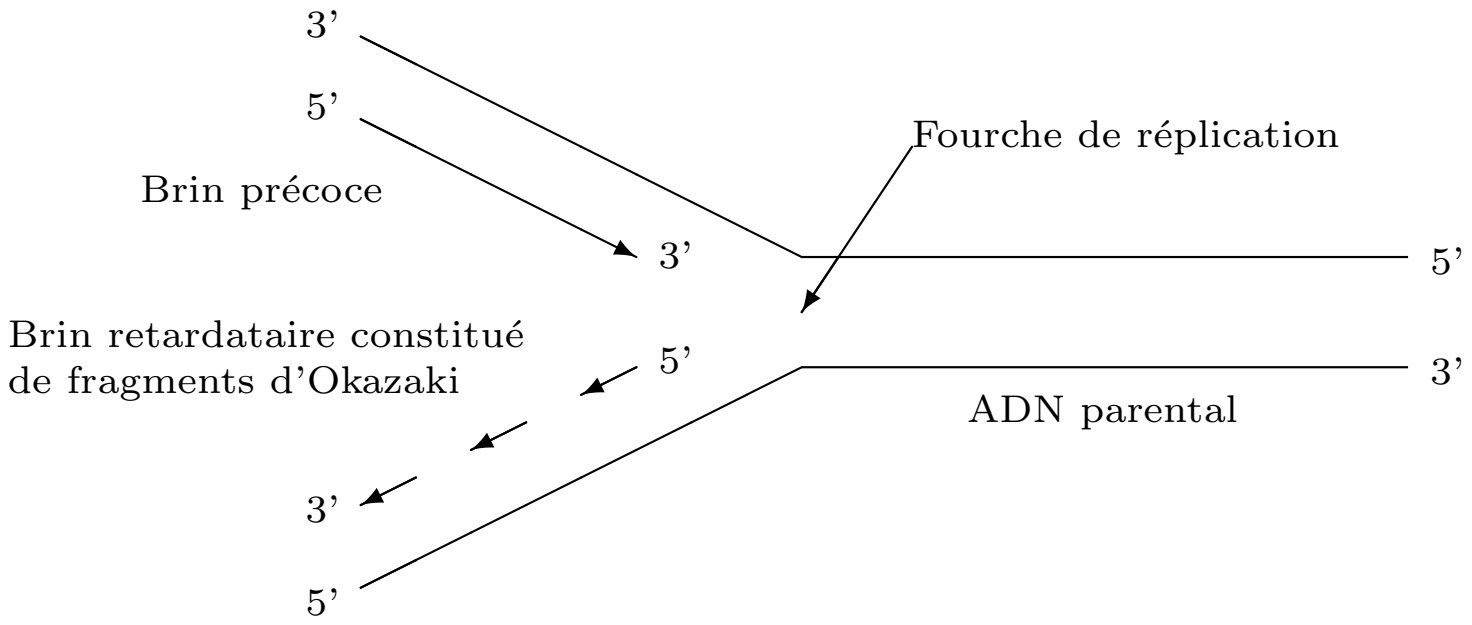
Aimé Lachal

1. **Réplication des ADN**
2. **Phylogénie**
3. **Séquences biologiques**

1 Réplication d'un ADN

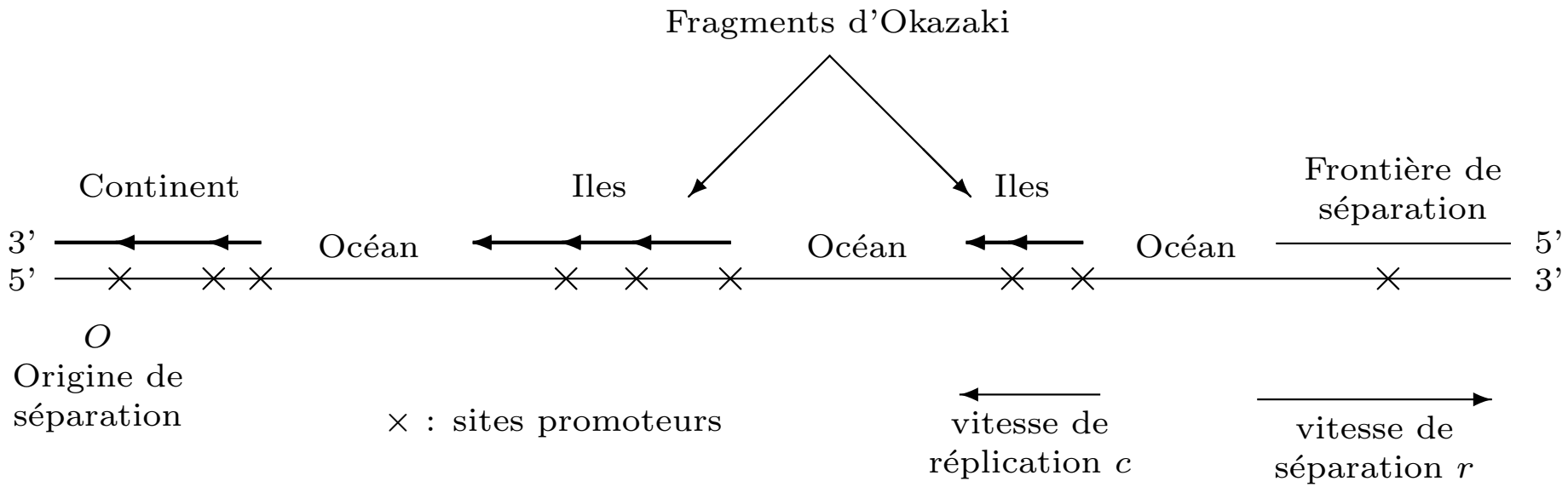
Les deux brins sont répliqués à l'aide d'une ADN polymérase.

- brin précoce : réplication continue dans le sens de la séparation ;
- brin retardataire : réplication discontinue dans le sens contraire de la séparation → fragments d'Okazaki (1968).



Constitution du brin retardataire

- un brin (le continent) obtenu par concaténation depuis l'origine de tous les fragments qui sont liés à leurs voisins de gauche ;
- les segments restants qui ne sont pas encore reliés au continent, ce sont les fragments d'Okazaki (les îles). Les interstices entre ces fragments constituent l'océan.



Modèle probabiliste de Cowan & Chiu

Les sites promoteurs sont répartis selon un processus de Poisson spatial d'intensité μ, ou encore un processus de Poisson temporel d'intensité $\lambda = r\mu$.

Quelques v.a. intéressantes

- N_t : nombre de fragments d'Okazaki existant à l'instant t ;
- L_t : longueur totale des interstices entre les fragments d'Okazaki à l'instant t ;
- D_t : distance entre la frontière de séparation et l'extrémité droite du continent à l'instant t.

Quelques résultats

1994 Cowan & Chiu (J. Appl. Probab.) : $\mathbb{E}(N_\infty) = \dfrac{r}{c}, \mathbb{E}(L_\infty) = \dfrac{r(r+c)}{\lambda c}$.

2000 Piau (J. Appl. Probab.) : $\mathbb{E}(D_\infty) = \dfrac{r+c}{\lambda} \displaystyle\sum_{n=1}^{\infty} \sigma(n) \Big(\frac{r}{r+c}\Big)^n$ où $\sigma(n)$ est le nombre de diviseurs de n.

2004 Lachal (Ann. Appl. Probab. et CRAS):

$$\begin{cases} \mathbb{E}(e^{-pN_\infty}) &= \displaystyle\prod_{n=1}^{\infty} \Big[1 - (1-e^{-p})\Big(\frac{r}{r+c}\Big)^n\Big], \\ \mathbb{E}(e^{-pL_\infty}) &= \displaystyle\prod_{n=0}^{\infty} \Big[1 + \frac{rp}{\lambda}\Big(\frac{r}{r+c}\Big)^n\Big]^{-1}, \\ \mathbb{E}(e^{-pD_\infty}) &= \displaystyle\prod_{n=1}^{\infty} \Big[1 + \frac{rp}{\lambda}\,\frac{(\frac{r}{r+c})^{n-1}}{1-(\frac{r}{r+c})^n}\Big]^{-1}. \end{cases}$$

2 Phylogénie

Évolution temporelle d'un nucléotide fixé sur un brin d'ADN ou d'ARN :

Modèle de chaîne de Markov $(X_t)_{t\geqslant 0}$ sur l'espace d'états $\mathcal{A} = \{A, G, C, T\}$ pour les ADN, $\mathcal{A} = \{A, G, C, U\}$ pour les ARN.

A = adénine, C = cytosine, G = guanine, T = thymine, U = uracile

La chaîne $(X_t)_{t\geqslant 0}$ est caractérisée par

- ses probabilités de transition (matrice de transition) :

$$\forall a, b \in \mathcal{A},\ P_{ab}(t) = \mathbb{P}(X_{s+t} = b | X_s = a) \quad \longrightarrow \quad \boldsymbol{P}(t) = (P_{ab}(t))_{a,b\in\mathcal{A}}$$

- sa loi initiale (vecteur probabilité, matrice-ligne) :

$$\forall a \in \mathcal{A},\ \pi_a(0) = \mathbb{P}(X_0 = a) \quad \longrightarrow \quad \boldsymbol{\pi}(0) = (\pi_A(0), \pi_G(0), \pi_C(0), \pi_T(0))$$

et alors, en posant

$$\forall a \in \mathcal{A},\ \pi_a(t) = \mathbb{P}(X_t = a) \quad \longrightarrow \quad \boldsymbol{\pi}(t) = (\pi_A(t), \pi_G(t), \pi_C(t), \pi_T(t))$$

on a $\boxed{\forall s, t,\ 0 \leqslant s \leqslant t,\ \boldsymbol{\pi}(t) = \boldsymbol{\pi}(s)P(t-s)}$

Propriété de Markov :

$$\forall s,t \geqslant 0,\ \boldsymbol{P}(s)\boldsymbol{P}(t) = \boldsymbol{P}(s+t) \Longrightarrow (\boldsymbol{P}(t))_{t\geqslant 0} \text{ semi-groupe}$$

Générateur infinitésimal :

$$\boldsymbol{A} = \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}\left(\boldsymbol{P}(\varepsilon) - I\right) = \boldsymbol{P}'(0) \Longrightarrow \boldsymbol{P}(t) = \exp(t\boldsymbol{A})$$

Pour les modèles phylogénétiques :

$$P_{ab}(\varepsilon) = \mathbb{P}(X_{t+\varepsilon} = b | X_t = a) = \begin{cases} \pi_{ab}\,\varepsilon + o(\varepsilon) & \text{si } a \neq b \\ 1 - \sum_{c\in\mathcal{A}\setminus\{a\}} \pi_{ac}\,\varepsilon + o(\varepsilon) & \text{si } a = b \end{cases}$$

$\longrightarrow$ paramètres phylogénétiques (taux de transition) :

$$\boldsymbol{A} = \begin{pmatrix} \dots & \pi_{AG} & \pi_{AC} & \pi_{AT} \\ \pi_{GA} & \dots & \pi_{GC} & \pi_{GT} \\ \pi_{CA} & \pi_{CG} & \dots & \pi_{CT} \\ \pi_{TA} & \pi_{TG} & \pi_{TC} & \dots \end{pmatrix} \begin{matrix} A \\ G \\ C \\ T \end{matrix}$$

$\dots$: somme sur chaque ligne $= 0$

Problèmes du probabiliste

- détermination explicite du semi-groupe $\boldsymbol{P}(t) = \exp(t\boldsymbol{A})$
- recherche d'une loi limite $\boldsymbol{\pi}_\infty$ (ergodicité) : $\boxed{\boldsymbol{\pi}(t) \underset{t\to+\infty}{\longrightarrow} \boldsymbol{\pi}_\infty}$
- recherche d'une loi invariante $\boldsymbol{\pi}_{\text{inv}}$ (régime stationnaire) :

$$\forall t \geqslant 0,\ \boldsymbol{\pi}(t) = \boldsymbol{\pi}_{\text{inv}}\boldsymbol{P}(t) = \boldsymbol{\pi}_{\text{inv}} \Longleftrightarrow \boxed{\boldsymbol{\pi}_{\text{inv}}\boldsymbol{A} = 0}$$

- recherche d'une loi réversible $\boldsymbol{\pi}_{\text{rév}}$ (retournement du temps) :

$$\forall a, b \in \mathcal{A},\ \forall s, t,\ 0 \leqslant s \leqslant t, \quad \mathbb{P}(X_s = a, X_t = b) = \mathbb{P}(X_s = b, X_t = a)$$

$$\Longleftrightarrow \boxed{\boldsymbol{D}_{\boldsymbol{\pi}_{\text{rév}}}\boldsymbol{A} = {}^t\boldsymbol{A}\boldsymbol{D}_{\boldsymbol{\pi}_{\text{rév}}}}$$

$\boldsymbol{D}_\pi$: matrice diagonale de termes diagonaux $\pi_a, a \in \mathcal{A}$

Problèmes du biologiste

- datation de l'origine de l'espèce :
  - probabilité de mutation $r(t) = \mathbb{P}(X_t \neq X_0)$
  - taux de mutation $\kappa = \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \mathbb{P}(X_{t+\varepsilon} \neq X_t) = r'(0)$
- datation de la divergence des espèces :
  - probabilité de mutation $\rho(t) = \mathbb{P}(X_t \neq Y_t)$ à partir de deux lignées $(X_t)_{t \geqslant 0}$ et $(Y_t)_{t \geqslant 0}$ issues d'une même souche $(X_0 = Y_0)$
  - si la chaîne est réversible : $\rho(t) = r(2t)$
  - paramètre d'évolution : $D(t) = 2\kappa t$
- estimation de paramètres (statistiques)

Modèle de Jukes-Cantor (1969)

- taux de transitions constants (!) $\longrightarrow$ générateur infinitésimal :

$$A_{ab} = \alpha \text{ si } a \neq b \quad \longrightarrow \quad \boldsymbol{A} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

- semi-groupe :

$$P(t) = \frac{1}{4} \begin{pmatrix} 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} \end{pmatrix}$$

- lois limite, invariante, réversible : $\boldsymbol{\pi}_\infty = \boldsymbol{\pi}_{\text{inv}} = \boldsymbol{\pi}_{\text{rév}} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$
- probabilité de mutation : $\rho(t) = \frac{3}{4}\left(1 - e^{-8\alpha t}\right)$
- paramètre d'évolution : $D(t) = 2 \times (3\alpha t) = -\frac{3}{4} \ln\left(1 - \frac{4}{3}\,\rho(t)\right)$

Modèles de Kimura (1980)

- distinction de deux types de substitutions entre les bases :
  - les transitions : substitutions internes entre purines $A \longleftrightarrow G$ et substitutions internes entre pyrimidines $C \longleftrightarrow T$ (taux α) ;
  - les transversions : substitutions externes entre purines et pyrimidines $\{A, G\} \longleftrightarrow \{C, T\}$ (taux β et γ).

  En général, les transitions sont plus fréquentes que les transversions.
- générateur infinitésimal :

$$A_{ab} = \begin{cases} \alpha & \text{si } a \longleftrightarrow b \text{ est une transition } A \longleftrightarrow G \text{ et } C \longleftrightarrow T \\ \beta & \text{si } a \longleftrightarrow b \text{ est une transversion } A \longleftrightarrow C \text{ et } G \longleftrightarrow T \\ \gamma & \text{si } a \longleftrightarrow b \text{ est une transversion } A \longleftrightarrow T \text{ et } G \longleftrightarrow C \end{cases}$$

$$\boldsymbol{A} = \begin{pmatrix} -\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & -\alpha-\beta-\gamma & \gamma & \beta \\ \beta & \gamma & -\alpha-\beta-\gamma & \alpha \\ \gamma & \beta & \alpha & -\alpha-\beta-\gamma \end{pmatrix}$$

Modèle de Hasegawa-Kishino-Yano (1985)

- générateur infinitésimal :

$$\boldsymbol{A} = \begin{pmatrix} -\alpha\pi_G - \beta\pi_Y & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & -\alpha\pi_A - \beta\pi_Y & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & -\alpha\pi_T - \beta\pi_R & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & -\alpha\pi_C - \beta\pi_R \end{pmatrix}$$

avec

$$\pi_R = \pi_A + \pi_G, \qquad \pi_Y = \pi_C + \pi_T$$

- loi limite (et réversible) :

$$\boldsymbol{\pi} = (\pi_A, \pi_G, \pi_C, \pi_T)$$

3 Séquences biologiques

Alignement de séquences biologiques, recherche de motifs, de région codante (exon), non-codante (intron), hydrophobe, etc.

- Séquence $\boldsymbol{A}$: suite de v.a. $(A_n)_{n\geqslant 1}$ à valeurs dans un alphabet $\mathcal{A}$.

$\mathcal{A} = \{A, C, G, T\}$ pour les ADN

$\mathcal{A} = \{A, C, G, U\}$ pour les ARN

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

- Fonction score $s : \mathcal{A} \longrightarrow \mathbb{R}$, $\begin{cases} S(\boldsymbol{A}) = \displaystyle\sum_{k=1}^{n} s(A_k) & \text{score global} \\ H(\boldsymbol{A}) = \displaystyle\max_{1\leqslant i\leqslant j\leqslant n} \sum_{k=i}^{j} s(A_k) & \text{score local} \end{cases}$

- **Problème :** loi de probabilité des v.a. $S(\boldsymbol{A})$ et $H(\boldsymbol{A})$?

0n pose $X_n = s(A_n)$ et $H_n = \max\limits_{0 \leqslant i \leqslant j \leqslant n} \sum\limits_{k=i}^{j} X_k$.

1987 Waterman, Gordon & Arratia :

$$\begin{cases} \text{si } \mathbb{E}(X) > 0, & \dfrac{H_n}{n} \xrightarrow[n\to\infty]{\text{p.s.}} \mathbb{E}(X), & \\ \text{si } \mathbb{E}(X) = 0, & \dfrac{H_n}{\sqrt{n}} \xrightarrow[n\to\infty]{\text{p.s.}} C & \text{pour un certain } C > 0, \\ \text{si } \mathbb{E}(X) < 0, & \dfrac{H_n}{\ln n} \xrightarrow[n\to\infty]{\text{p.s.}} \dfrac{1}{\lambda}, & \lambda > 0 \text{ tel que } \mathbb{E}[e^{\lambda X}] = 1. \end{cases}$$

1990 Karlin & Altschul :

$$H_n - \frac{\ln n}{\lambda} \xrightarrow{\text{loi}} \text{ loi de Gumbel, i.e. } \mathbb{P}(H_n \leqslant \frac{\ln n}{\lambda} + x) \underset{n\to\infty}{\longrightarrow} e^{-K_H e^{-\lambda x}};$$

$\longrightarrow$ résultat implémenté dans les logiciels d'alignement BLAST et FASTA