Quelques modèles de reproduction stochastiques

Aimé Lachal

Mars 2001

1 Modèle de reproduction I (R.A. Fisher & S. Wright, 1930)

1.1 Modèle de reproduction « haploïde simple »

On considère une population de N paires de chromosomes, et l'on étudie l'évolution d'une paire de gènes situé en un certain locus de chaque paire de chromosomes. Chaque paire de gènes est constituée de deux allèles, chacun pouvant se présenter sous deux formes : A ou a (génotype). La population ainsi considérée comporte $2N$ allèles A et a. Le mode de reproduction de cette population s'effectue au hasard selon le modèle binomial (recombinaison génétique) : si la n^{e} génération contient i allèles A et $2N-i$ allèles a, alors on construit la $(n+1)^{\text{e}}$ en prélevant au hasard avec remise $2N$ allèles de la n^{e} génération. La probabilité de tirer un allèle a est $\frac{i}{2N}$, celle de tirer un allèle a est $1-\frac{i}{2N}$. Ainsi, le nombre d'allèles A présents dans la n^{e} génération est donné par

$$X_n=\sum_{k=1}^{2N}\xi_{k,n} \quad \text{où} \quad \xi_{k,n}=\begin{cases}1 & \text{si le } k^{\text{e}} \text{ allèle prélevé de la génération précédente est } A,\\ 0 & \text{si le } k^{\text{e}} \text{ allèle prélevé de la génération précédente est } a.\end{cases}$$

Pour chaque $n\in\mathbb{N}^*$, la suite $(\xi_{k,n})_{1\leqslant k\leqslant 2N}$ est i.i.d. de loi de Bernoulli $\mathcal{B}(1,p_i)$ avec $p_i=\mathbb{P}(\xi_{k,n}=1)=\frac{i}{2N}$ et X_{n+1} sachant $X_n=i$ suit donc la loi binomiale $\mathcal{B}(2N,p_i)$:

$$p_{ij}=\mathbb{P}(X_{n+1}=j\mid X_n=i)=C_{2N}^j\Big(\frac{i}{2N}\Big)^j\Big(1-\frac{i}{2N}\Big)^{2N-j},\quad 0\leqslant i,j\leqslant 2N.$$

Rappelons que $\mathbb{P}(B|A)=\dfrac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)}$ et si $\Omega=\cup_n A_n$, $\mathbb{P}(B)=\sum_n\mathbb{P}(B|A_n)\mathbb{P}(A_n)$. On peut donc prévoir la $(n+1)^{\text{e}}$ génération à partir de la n^{e} grâce à

$$\mathbb{P}(X_{n+1}=j)=\sum_{j=0}^{2N}\mathbb{P}(X_{n+1}=j\mid X_n=i)\mathbb{P}(X_n=i).$$

Ceci s'écrit en posant $\pi_i^{(n)}=\mathbb{P}(X_n=i)$,

$$\pi_j^{(n+1)}=\sum_{i=0}^{2N}\pi_i^{(n)}p_{ij}$$

soit, sous forme matricielle,

$$\pi^{(n+1)}=\pi^{(n)}P$$

où $\pi^{(n)}=(\pi_i^{(n)})_{0\leqslant i\leqslant 2N}$ est le vecteur-ligne représentant la loi de X_n et P est la matrice carrée $(p_{ij})_{0\leqslant i,j\leqslant 2N}$. La suite $(X_n)_{n\geqslant 1}$ ainsi construite est une chaîne de Markov sur l'espace d'états $\{0,1,\dots,2N\}$ puisque X_{n+1} ne dépend du passé que par l'intermédiaire de X_n (chaque génération ne dépend que de la précédente). P est la matrice des probabilités de transition de $(X_n)_{n\geqslant 1}$. Cette chaîne est homogène dans le temps car les p_{ij} ne dépendent pas de n. On peut alors, théoriquement, déterminer la loi de probabilité de X_n en fonction de la répartition initiale ($1^{\text{ère}}$ génération) :

$$\pi^{(n)}=\pi^{(0)}P^n$$

ou encore, en notant $P^n = (p_{ij}^{(n)})_{0\leqslant i,j\leqslant 2N}$,

$$\pi_j^{(n)} = \sum_{i=0}^{2N} \pi_i^{(0)} p_{ij}^{(n)}.$$

Une question intéressante concerne l'évolution de la population en temps grand : la variable aléatoire X_n admet-elle une loi limite lorsque $n \to +\infty$?

Remarquons que si $i = 0$, alors $p_{ij} = \begin{cases} 1 \text{ si } j = 0 \\ 0 \text{ si } j \geqslant 1 \end{cases}$, c'est-à-dire si $X_n = 0$ p.s. alors $X_{n+1} = 0$ p.s. et la population finira par être constituée uniquement d'allèles A (ligne pure).

De même, si $i = 2N$, alors $p_{ij} = \begin{cases} 0 \text{ si } j \leqslant 2N-1 \\ 1 \text{ si } j = 2N \end{cases}$, c'est-à-dire si $X_n = 2N$ p.s. alors $X_{n+1} = 2N$ p.s. et la population finira par être constituée uniquement d'allèles a (ligne pure).

On dit que les états 0 et $2N$ sont des états absorbants.

Si $1 \leqslant i \leqslant 2N-1$, $p_{ij} > 0$ pour tout état j, donc l'état j est accessible depuis l'état i. Il y aura nécessairement absorption en 0 ou $2N$, les états intermédiaires $1, \ldots, 2N-1$ étant transitoires. On va déterminer les probabilités d'absorption correspondantes (fixation de l'espèce) :

$$\begin{cases} q_{i,0} &= \mathbb{P}(\text{absorption en } 0 \mid X_0 = i) &= \mathbb{P}(\exists n \geqslant 1 : X_n = 0 \mid X_0 = i) \\ q_{i,2N} &= \mathbb{P}(\text{absorption en } 2N \mid X_0 = i) &= \mathbb{P}(\exists n \geqslant 1 : X_n = 2N \mid X_0 = i). \end{cases}$$

En observant que

$$\mathbb{E}(X_{n+1} \mid X_n = i) = \sum_{j=0}^{2N} j p_{ij} = 2N p_i = i,$$

(propriété de martingale), on obtient par itérations successives :

$$\sum_{j=0}^{2N} j p_{ij}^{(n)} = i.$$

Les états $1, \ldots, 2N-1$ étant transitoires, $p_{ij}^{(n)} \underset{n\to\infty}{\longrightarrow} 0$ pour tout $j \in \{1, \ldots, 2N-1\}$ et alors $2N p_{i\,2N}^{(n)} \underset{n\to\infty}{\longrightarrow} i$; d'où les probabilités d'absorption :

$$\boxed{\begin{aligned} q_{i,0} &= \mathbb{P}(\text{absorption en } 0 \mid X_0 = i) = 1 - \frac{i}{2N}, \\ q_{i,2N} &= \mathbb{P}(\text{absorption en } 2N \mid X_0 = i) = \frac{i}{2N}. \end{aligned}}$$

Remarques :
- le nombre pair $2N$ correspond à un organisme diploïde. L'étude précédente (historique) est en fait valable pour n'importe quel type d'organisme (haploïde ou polyploïde) ;
- le modèle de Fisher & Wright est rudimentaire. On peut le rendre plus réaliste en prenant en compte des pressions de mutation ainsi que des forces de sélection.

1.2 Mutation

On suppose que les pressions de mutation agissent sur la n^{e} génération avant le tirage au sort de la suivante. Notons α (resp. β) la probabilité de mutation de l'allèle A (resp. a) en a (resp. A). Si $X_n = i$, il convient alors de choisir pour p_i et $1 - p_i$:

$$\begin{aligned} p_i &= \mathbb{P}(\text{l'allèle } A \text{ est prélevé}) \\ &= \mathbb{P}(A \text{ est prélevé} \mid A \text{ n'a pas muté})\mathbb{P}(A \text{ n'a pas muté}) \\ &\quad + \mathbb{P}(A \text{ est prélevé} \mid a \text{ a muté})\mathbb{P}(a \text{ a muté}) \\ &= (1-\alpha)\frac{i}{2N} + \beta\Big(1 - \frac{i}{2N}\Big), \\ 1 - p_i &= (1-\beta)\Big(1 - \frac{i}{2N}\Big) + \alpha\frac{i}{2N}. \end{aligned}$$

Si $0 < \alpha, \beta < 1$, on a pour tous états $i, j \in \{0, \dots, 2N\}$, $p_{ij} > 0$ et alors tous les états communiquent entre eux : la chaîne de Markov $(X_n)_{n \geqslant 1}$ est irréductible. Dans ce cas, il n'y a p.s. jamais fixation de la population. En fait, la v.a. X_n converge en loi vers une v.a. X telle que pour tout $i \in \{0, \dots, 2N\}$, $\mathbb{P}(X = i) > 0$; tous les états sont p.s. visités. Le vecteur-ligne π représentant la loi de X est l'unique solution de l'équation $\pi = \pi P$ tel que $\sum_{i=0}^{2N} \pi_i = 1$. On dit que la chaîne de Markov $(X_n)_{n \geqslant 1}$ est ergodique. La loi de X n'est pas connue explicitement mais on peut cependant calculer son espérance. Remarquant que

$$\mathbb{E}(X_{n+1} \mid X_n = i) = \sum_{j=0}^{2N} j p_{ij} = 2N p_i = 2N\beta + (1 - \alpha - \beta) i,$$

on déduit la récurrence

$$\mathbb{E}(X_{n+1}) = 2N\beta + (1 - \alpha - \beta)\mathbb{E}(X_n),$$

d'où

$$\mathbb{E}(X_n) = (1 - \alpha - \beta)^n \Big(\mathbb{E}(X_0) - \frac{2N\beta}{\alpha + \beta} \Big) + \frac{2N\beta}{\alpha + \beta}.$$

Faisons tendre n vers $+\infty$; on obtient le nombre moyen d'allèles A présents dans chaque génération en régime stationnaire :

$$\boxed{\mathbb{E}(X) = \frac{2N\beta}{\alpha + \beta}.}$$

1.3 Sélection

On suppose cette fois qu'une force de sélection s'exerce en faveur de l'allèle A ; on pondère alors les i allèles A présents dans la n^{e} génération par un poids $1 + \varepsilon$. Tout se passe comme si on disposait de $(1 + \varepsilon)i$ allèles A et $(2N - i)$ allèles a. On prendra donc

$$p_i = \frac{(1 + \varepsilon)i}{2N + \varepsilon i} \quad \text{et} \quad 1 - p_i = \frac{2N - i}{2N + \varepsilon i}.$$

Il y a fixation de l'espèce dans ce cas.

2 Modèle de reproduction II (I.V. Schensted, 1958)

On étudie l'évolution d'un gène constitué de N plasmides. Ces plasmides sont soit mutants (type A) soit non mutants (type a). Lors du dédoublement et de la division de la cellule mère (mitose), le gène se dédouble en deux gènes contenant chacun le même nombre N de plasmides et donne naissance à deux cellules filles. On considère qu'une seule des deux cellules filles survit, c'est-à-dire qu'on ne considère qu'une lignée de descendance. On note X_n le nombre de plasmides mutants présents dans la n^{e} génération.

Supposons $X_n = i$: la cellule mère de la n^{e} génération contient i plasmides mutants et $N - i$ non mutants. Lors de sa division, le gène se dédouble et l'on est en présence de $2i$ plasmides mutants et $2N - 2i$ non mutants. La cellule mère donne naissance à deux gènes fils constitués chacun de N plasmides ; l'un des deux comporte disons j mutants et $N - j$ non mutants, l'autre gène comporte alors $2i - j$ mutants et $N + j - 2i$ non mutants. En considérant par exemple que seule la première cellule fille survit, on a $X_{n+1} = j$. Si seule la deuxième cellule fille survivait, on aurait $X_{n+1} = 2i - j$. En fait, les deux événements $X_{n+1} = j$ et $X_{n+1} = 2i - j$ ont même probabilité.

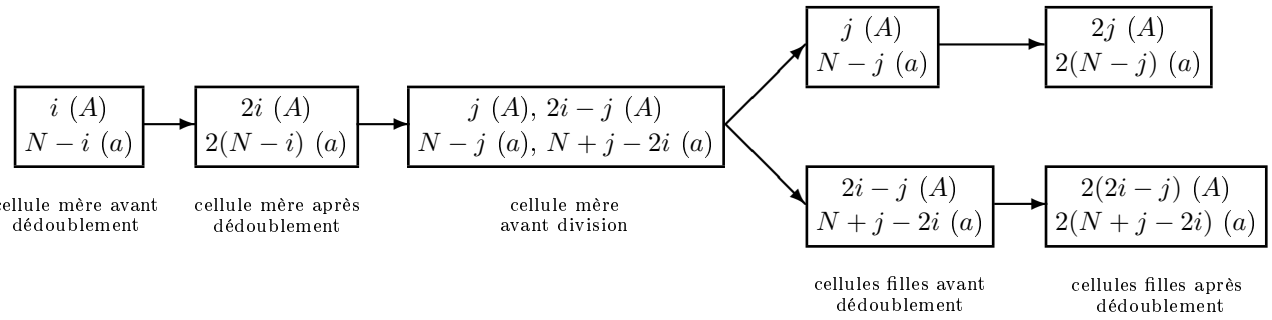

FIGURE 1 – Modèle à un paramètre

Le procédé ainsi décrit conduit à un modèle hypergéométrique :

$$p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i) = \begin{cases} \dfrac{C_{2i}^j C_{2N-2i}^{N-j}}{C_{2N}^N} & \text{si } \max(0, 2i-N) \leqslant j \leqslant \min(N, 2i), \\ 0 & \text{sinon} \end{cases}$$

c'est-à-dire X_{n+1} sachant $X_n = i$ suit la loi hypergéométrique $\mathcal{H}(2N, N, \frac{i}{N})$. La suite $(X_n)_{n\geqslant 1}$ ainsi construite est une chaîne de Markov sur l'espace d'états $\{0, 1, \ldots, N\}$. C'est même une martingale puisque

$$\mathbb{E}(X_{n+1} \mid X_n = i) = \sum_{j=\max(0,2i-N)}^{\min(N,2i)} j\, C_{2i}^j C_{2N-2i}^{N-j} / C_{2N}^N = 2i\, C_{2N-1}^{N-1} / C_{2N}^N = i.$$

En effet, partant de l'égalité $(x+1)^{2i}(y+1)^{2N-2i} = \sum_{k=0}^{2i} C_{2i}^k\, x^k \sum_{l=0}^{2N-2i} C_{2N-2i}^l\, y^l$, on obtient après dérivation par rapport à x puis substitution de y en x :

$$\begin{aligned} 2i(x+1)^{2N-1} &= \sum_{k=0}^{2i} k\, C_{2i}^k\, x^{k-1} \sum_{l=0}^{2N-2i} C_{2N-2i}^l\, x^l \\ &= \sum_{k=0}^{2i} \sum_{l=0}^{2N-2i} k\, C_{2i}^k\, C_{2N-2i}^l\, x^{k+l-1} \\ &= \sum_{n=0}^{2N-1} \Big(\sum_{j=\max(0,2i+n+1-2N)}^{\min(n+1,2i)} j\, C_{2i}^j C_{2N-2i}^{n+1-j} \Big) x^n. \end{aligned}$$

D'où, par identification des coefficients de x^{N-1},

$$\sum_{j=\max(0,2i-N)}^{\min(N,2i)} j\, C_{2i}^j C_{2N-2i}^{N-j} = 2i\, C_{2N-1}^{N-1},$$

puis le résultat annoncé. Les états $1, \ldots, N-1$ sont transitoires et les états 0 et N absorbants. L'état absorbant N signifie que p.s. tous les plasmides sont mutants, ce qui entraînera très vraisemblablement la mort de l'espèce, tandis que l'état absorbant 0 correspond à un gène qui ne produira p.s. plus de forme mutante. On obtient comme dans le modèle précédent les probabilités d'absorption

$$\boxed{\begin{aligned} q_{i,0} &= \mathbb{P}(\text{absorption en } 0 \mid X_0 = i) = 1 - \frac{i}{2N} \\ q_{i,2N} &= \mathbb{P}(\text{absorption en } 2N \mid X_0 = i) = \frac{i}{2N}. \end{aligned}}$$