–Texte–
Chaînes de Wright-Fisher

Mots-clefs : chaîne de Markov, martingale, temps d'absorption.

1 Modélisation

On souhaite étudier l'évolution de la fréquence d'un allèle dans une population de petite taille pour un gène se présentant sous deux allèles seulement. On considère une population haploïde : chaque individu possède un seul exemplaire du gène, sous la forme d'un des deux allèles A ou B. C'est par exemple le cas de l'ADN mitochondrial humain qui est uniquement transmis par la mère. L'étude de son évolution repose donc sur un modèle de population haploïde et asexué, car seul l'évolution de la population féminine conditionne l'évolution de cet ADN.

Présentons le cadre de notre modèle :

- le gène d'intérêt se présente sous deux allèles distincts A et B,
- la taille de la population reste constante au cours du temps, égale à N,
- les générations ne se chevauchent pas : à chaque instant k, la k-ième génération meurt et donne naissance aux N individus de la $(k+1)$-ième génération,
- chacun des enfants choisit son parent uniformément parmi tous les individus de la génération précédente et indépendamment des autres,
- la reproduction à l'instant k ne dépend pas des reproductions précédentes.

On note X_n le nombre d'allèles A dans la population au temps n. Au vu des hypothèses ci-dessus, il paraît naturel de modéliser la suite $(X_n)_{n\geqslant 0}$ par une chaîne de Markov à valeurs dans $\mathcal{S} = \{0, 1, \ldots, N\}$ déterminée par la propriété suivante : la loi de X_{n+1} sachant X_n est la loi binomiale $\mathcal{B}(N, X_n/N)$. En d'autres termes, la matrice de transition $P = (p_{ij})_{i,j}$ de la chaîne est donnée par

$$p_{ij} := \mathbb{P}(X_{n+1} = j | X_n = i) = \binom{N}{j} \psi_i^j (1-\psi_i)^{N-j} \tag{1}$$

pour $0 \leqslant i, j \leqslant N$ avec $\psi_i = i/N$.

Il semble intuitif qu'un allèle puisse finir par l'emporter sur l'autre car, si, à un instant n, X_n vaut 0 ou N alors il en sera de même dans tout le futur. C'est ce phénomène, appelé dérive génétique, qu'il s'agit à présent de décrire et quantifier.

2 Le modèle simple

Les états 0 et N sont absorbants. Tous les autres mènent à $\{0, N\}$, ils sont donc transients. L'espace d'états étant fini, le temps d'atteinte de $\{0, N\}$ est fini presque sûrement et même intégrable. La remarque suivante donne une justification de ce point dans le cas particulier de la chaîne de Wright-Fisher.

Remarque 2.1. Sachant que $X_n = i \notin \{0, N\}$, la probabilité que $X_{n+1} \in \{0, N\}$ est égale à $(1-\psi_i)^N + \psi_i^N$ qui est minimale pour $\psi_i = 1/2$ et vaut alors 2^{-N+1}. Ainsi, la chaîne atteindra les états absorbants avant qu'un lanceur de pièces ne fasse *pile* si la pièce est biaisée de telle sorte que *pile* sorte avec probabilité 2^{-N+1}. Cette borne est extrêmement pessimiste mais elle fournit facilement un contrôle sous géométrique explicite pour le temps d'atteinte étudié et confirme en particulier que le temps d'atteinte de $\{0, N\}$ est fini presque sûrement et même intégrable.

En plus d'être une chaîne de Markov, la suite $(X_n)_n$ a le bon goût d'être également une martingale.

Proposition 2.2. *La suite $(X_n)_n$ est une martingale pour sa filtration naturelle $(\mathcal{F}_n)_{n\geqslant 0}$.*

Cette propriété de $(X_n)_n$ permet de déterminer la probabilité de fixation (et la probabilité de disparation) de l'allèle A.

Corollaire 2.3. *La probabilité sachant que $X_0 = i$ que X atteigne N (avant 0) est égale à i/N.*

Démonstration. On applique le théorème d'arrêt à la martingale bornée X. Notons T le temps d'atteinte de l'ensemble $\{0, N\}$. Alors, pour $i = 0, 1, \ldots, N$,

$$\begin{cases} 1 = \mathbb{P}_i(T < +\infty) = \mathbb{P}_i(X_T = 0) + \mathbb{P}_i(X_T = N), \\ i = \mathbb{E}_i(X_0) = \mathbb{E}_i(X_T) = 0 \times \mathbb{P}_i(X_T = 0) + N \times \mathbb{P}_i(X_T = N), \end{cases}$$

d'où l'on tire $\mathbb{P}_i(X_T = N) = i/N$. □

La question suivante est de mesurer la vitesse de disparition de l'un des allèles. Voici quelques résultats dans cette direction.

Lemme 2.4. *Pour tout $n \geqslant 1$,*

$$\mathbb{E}(X_n(N - X_n)) = \left(1 - \frac{1}{N}\right)^n \mathbb{E}(X_0(N - X_0)).$$

Démonstration. On a

$$\mathbb{E}(X_n(N - X_n)) = N\mathbb{E}(X_n) - \mathbb{E}(X_n^2) = N\mathbb{E}(X_{n-1}) - \mathbb{E}\big(\mathbb{E}(X_n^2|X_{n-1})\big).$$

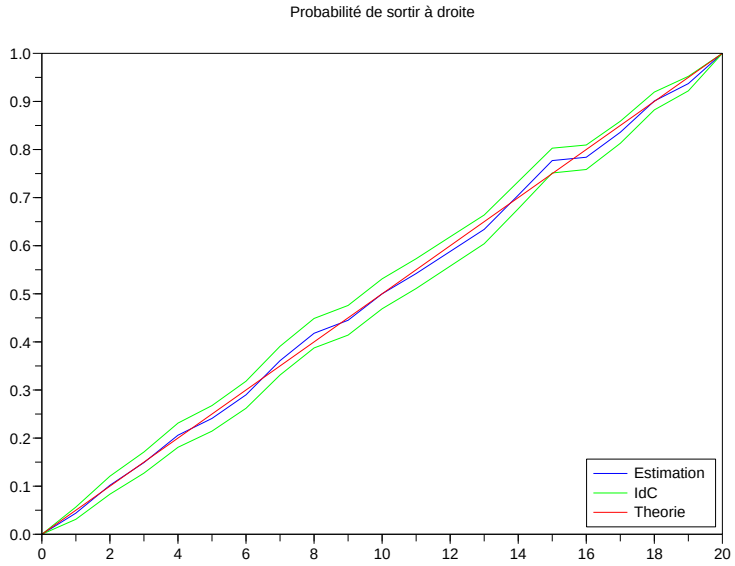

Fig. 1 – Probabilité de fixation de A pour $N = 20$.

On écrit alors

$$\mathbb{E}(X_n^2|X_{n-1}) = \mathrm{Var}(X_n|X_{n-1}) + (\mathbb{E}(X_n|X_{n-1}))^2 = X_{n-1}(1 - X_{n-1}/N) + X_{n-1}^2.$$

En regroupant les termes, on obtient bien

$$\mathbb{E}(X_n(N - X_n)) = \left(1 - \frac{1}{N}\right)\mathbb{E}(X_{n-1}(N - X_{n-1})).$$

ce qui fournit le résultat. □

Posons $\lambda = 1 - 1/N$. L'hétérozygotie est la probabilité que deux gènes choisis aléatoirement (sans remise) dans la population totale à la génération n soient représentés par des allèles différents. Elle est donnée par

$$H_n = \frac{2X_n(N - X_n)}{N(N - 1)}.$$

D'après le calcul précédent, l'hétérozygotie moyenne $h(n)$ vérifie

$$h_n := \mathbb{E}(H_n) = \lambda^n h_0.$$

Remarque 2.5. De même, la variance de X_n se calcule par un simple conditionnement :

$$\mathbb{V}(X_n) = \mathbb{E}(\mathbb{V}(X_n|X_{n-1})) + \mathbb{V}(\mathbb{E}(X_n|X_{n-1})).$$

On obtient alors, toujours avec $\lambda = 1 - 1/N$,

$$\mathbb{V}(X_n) = \mathbb{E}(X_0)(N - \mathbb{E}(X_0))(1 - \lambda^n) + \lambda^n \mathbb{V}(X_0).$$

Une question naturelle est aussi de mieux comprendre la loi de T le temps de disparition d'un des deux allèles, par exemple en estimant son espérance. Pour le modèle de Wright-Fisher cette question est délicate : il n'existe pas de formule simple pour tout N. En effet, si l'on note m_i l'espérance du temps d'absorption de X sachant que $X_0 = i$, on a $m_0 = m_N = 0$ et, grâce à la propriété de Markov,

$$m_i = 1 + \sum_{j=0}^{N} p_{ij} m_j.$$

Ce système à $N - 1$ inconnues n'est pas facile à résoudre. Il est possible d'utiliser l'ordinateur pour trouver une valeur approchée déterministe de la solution mais, dès que N est un peu grand, des problèmes dus au fait que les coefficients p_{ij} sont très petits risquent de fausser le résultat. On peut aussi utiliser une méthode probabiliste pour estimer la quantité m_i *via* une méthode de Monte-Carlo.

Il est toutefois possible de trouver un équivalent de m lorsque la taille de la population tend vers l'infini. Notons Z la chaîne sur $\mathcal{S}/N$ définie par $Z_n = X_n/N$. Bien entendu, le temps d'atteinte de $\{0, N\}$ pour X est égal au temps d'atteinte de $\{0, 1\}$ pour Z. Pour tout $x \in \mathcal{S}/N$, notons $t(x)$ l'espérance de T lorsque $Z_0 = x$. Enfin, sachant que $Z_0 = x$, alors Z_1 s'écrit $x + X$ où X est tel que $N(X - x)$ suit la loi binomiale $\mathcal{B}(N, x)$. On a alors, d'après la propriété de Markov et la définition de X,

$$t(x) = \mathbb{E}(t(x + X) + 1) = 1 + \mathbb{E}(t(x + X)).$$

Supposons que t soit proche d'une fonction de classe $\mathcal{C}^\infty$. Puisque la variable aléatoire X est bornée, on peut écrire

$$t(x + X) = t(x) + t'(x)X + \frac{t''(x)}{2} X^2 + O(|X|^3).$$

Or, il est clair que

$$\mathbb{E}(X) = 0, \quad \mathbb{E}(X^2) = \frac{x(1 - x)}{N}.$$

D'autre part, en vertu du théorème limite central, la loi de $\sqrt{N}X$ est proche de la loi $\mathcal{N}(0, x(1 - x))$ et ainsi, par exemple, $\mathbb{E}(X^4)$ est de l'ordre de N^{-2}. En résumé, on obtient l'expression suivante :

$$t(x) = 1 + t(x) + \frac{x(1 - x)}{2N} t''(x) + O(N^{-3/2}).$$

On pourrait donc dire que la fonction t doit ressembler à la fonction y solution de

$$\forall x \in]0,1[, \quad y''(x) = \frac{2N}{x(1-x)} \quad \text{et} \quad y(0) = y(1) = 0.$$

On obtient alors l'approximation suivante.

Proposition 2.6. *Si N est grand et $x = i/N$ alors*

$$m(i) \underset{N}{\sim} -2N(x \ln x + (1-x)\ln(1-x)).$$

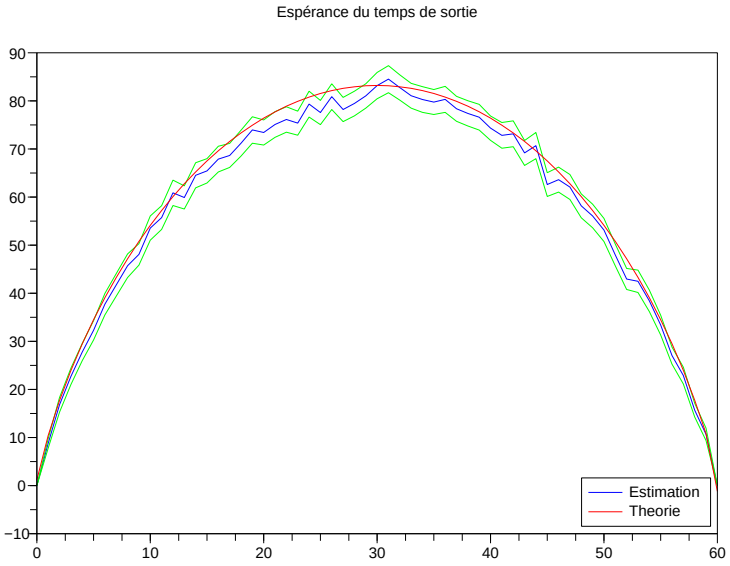

Fig. 2 – Estimation du temps de fixation pour $N = 40$.

2.1 Prise en compte de la sélection

Les différents allèles procurent à l'individu qui en est doté des capacités plus ou moins grandes dans tous les domaines de son développement et de sa reproduction : viabilité, potentiel attractif, fertilité etc. C'est la fameuse sélection naturelle. Tentons de quantifier ceci. Bien que l'adaptation d'un individu à son environnement soit déterminée par de nombreux facteurs, nous supposerons ici qu'il n'est déterminé que par le locus qui nous intéresse. Nous supposerons de plus que la sélection ne s'opère que sur le critère de viabilité. Supposons que les adaptabilités des deux allèles A et B soient données par $1+s$ et 1, c'est-à-dire que qu'un individu porteur de l'allèle A a $1+s$ fois plus de chances de survivre

qu'un individu porteur de l'allèle B. Pour construire la population à la date $n+1$, sachant que $X_n = k$, tout se passe comme si on tirait des allèles dans une urne qui contient une proportion $(1+s)k/((1+s)k+N-k)$. En d'autres termes, $(X_n)_{n\geqslant 0}$ est une chaîne de Markov sur $\mathcal{S}$ telle que la loi de X_{n+1} sachant $X_n = k$ est la loi binomiale de paramètres N et $(1+s)k/((1+s)k+N-k)$. Les états 0 et N sont toujours absorbants mais X n'est plus une martingale. Ceci complique le calcul des probabilités de fixation des allèles A et B. Notons $(\pi_i)_i$ les probabilités de fixation de la chaîne en N (avant 0) lorsque $X_0 = i$. La propriété de Markov assure que

$$\pi_i = \sum_{j=0}^{N} p_{ij}\pi_j, \quad \text{avec } \pi_0 = 0 \text{ et } \pi_N = 1.$$

La pression de sélection est très faible en pratique. Supposons que s soit de l'ordre N^{-1} et posons $\alpha = Ns$. Par le même raisonnement que dans la preuve de la proposition 2.6, on peut écrire, pour $x = i/N$

$$\pi(x) = \mathbb{E}(\pi(x+Z)) = \pi(x) + \pi'(x)\mathbb{E}(Z) + \frac{\pi''(x)}{2}\mathbb{E}(Z^2) + O(|Z|^3).$$

En tenant compte de la sélection, on a $Z = Y/N - x$ où Y suit la loi $\mathcal{B}(N, (1+s)x/(sx+1))$ donc

$$\begin{aligned}
\mathbb{E}(Z) &= \frac{(1+s)x}{sx+1} - x = \frac{sx(1-x)}{sx+1} = \frac{1}{N}\alpha x(1-x) + O(N^{-2}),\\
\mathbb{E}(Z^2) &= \frac{1}{N^2}\mathbb{V}(Y) + (\mathbb{E}(Y)/N - x)^2 = \frac{1}{N}x(1-x) + O(N^{-2}),\\
\mathbb{E}(|Z|^3) &= O\big(N^{-3/2}\big).
\end{aligned}$$

La fonction π semble donc proche de la solution de l'équation différentielle suivante :

$$z''(x) + 2\alpha z'(x) = 0, \quad \text{avec } z(0) = 0 \text{ et } z(1) = 1,$$

c'est-à-dire

$$\pi(x) \sim z(x) = \frac{1 - e^{-2\alpha x}}{1 - e^{-2\alpha}}.$$

La figure 3 propose une illustration de l'estimation du la probabilité de fixation de l'allèle A pour $N = 30$ et $\alpha = 2$.

Supposons que $N = 10^5$, $s = 10^{-4}$ et $x = 0,5$. Alors $\alpha = 20$ et $\pi = 0,999955$. En l'absence de sélection ($s = 0$), on a bien sûr $\pi(0,5) = 0,5$. Même le faible avantage $0,0001$ (inobservable en laboratoire ou par des mesures statistiques) est pourtant suffisamment grand pour avoir un effet déterminant sur la fixation des allèles. Bien que cet effet soit imperceptible sur une génération, il l'est jusqu'à l'instant de fixation car le temps de fixation est très grand.

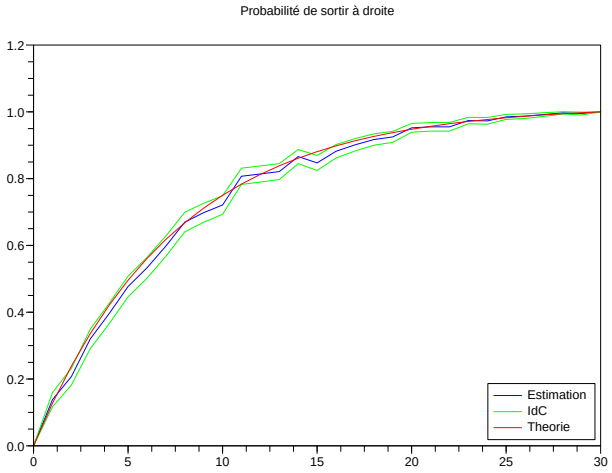

Fig. 3 – Probabilité de fixation de A pour $N = 30$ et $\alpha = 2$.

Remarque 2.7. Le même type de raisonnement conduirait à approcher l'espérance du temps de fixation par la solution de l'équation différentielle suivante :

$$z''(x) + 2\alpha z'(x) = -\frac{2N}{x(1-x)}, \quad \text{avec } z(0) = 0 \text{ et } z(1) = 0,$$

qui n'est pas facile à résoudre... La figure 4 propose une illustration de l'estimation du temps d'absorption pour $N = 30$ et $\alpha = 10$.

3 Suggestions

1. On pourra introduire le modèle de Wright-Fisher.
2. On pourra démontrer et/ou illustrer par la simulation le corollaire 2.3.
3. On pourra démontrer et/ou illustrer par la simulation la proposition 2.6.
4. On pourra expliquer comment prendre en compte la sélection et illustrer les résultats sur les probabilités et temps de fixation.

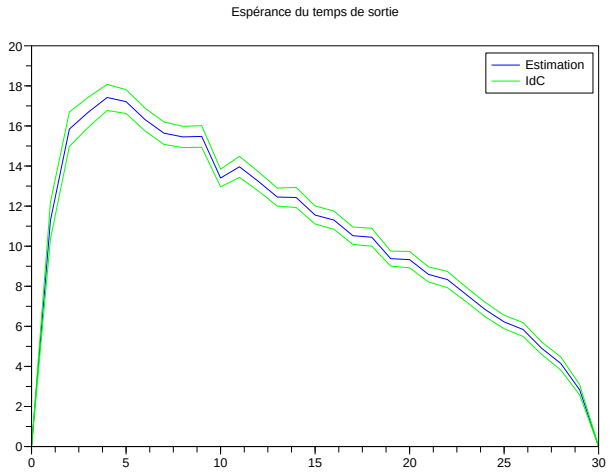

Fig. 4 – Espérance du temps de fixation pour $N = 30$ et $\alpha = 10$.