7

Estimation du taux de mutation de l'ADN

Depuis les années 1990, les modèles probabilistes pour l'évolution génétique connaissent un essor considérable. Ils permettent d'aborder pour ne citer que ces exemples :

- la reconstruction des arbres généalogiques ou phylogéniques avec des techniques récentes à partir des séquences d'ADN, voir les monographies [17 ou 9] et les nombreux sites consacrés aux algorithmes de reconstruction d'arbres,
- l'effet de l'histoire des populations, expansion ou récession, sur la diversité de l'ADN, voir par exemple [4],
- l'influence des mutations neutres ou favorables sur la diversité de l'ADN, voir les monographies [5 et 4].

Ce chapitre est une introduction au modèle d'évolution de Wright-Fisher, qui est un modèle élémentaire mais représentatif. Le lecteur intéressé pourra consulter les monographies de Ewens [5] (2004), Tavaré [17] (2001) et Durrett [4] (2002) ainsi que leurs références, où des modèles plus généraux sont étudiés et de nombreux thèmes liés à l'évolution génétique des populations sont traités.

Ce chapitre est organisé comme suit. Le paragraphe 7.1 présente le modèle d'évolution de Wright-Fisher développé à partir des années 1920. Le paragraphe 7.2 est consacré à l'étude des arbres généalogiques correspondants, qui repose sur les processus de coalescence. Le modèle de Wright-Fisher permet de vérifier que sans mutation la diversité biologique disparaît, voir le paragraphe 7.3. Enfin, le paragraphe 7.4 montre comment on peut, grâce aux modèles d'arbres généalogiques, estimer le taux de mutation de l'ADN. Ces méthodes d'estimation reposent sur les différences observées entre les séquences d'ADN au sein d'une même population. En revanche, la reconstruction d'un arbre généalogique, à partir des séquences d'ADN dépasse le cadre de ce chapitre.

7.1 Le modèle d'évolution de population

On présente le modèle d'évolution de population de Wright-Fisher, introduit par Fisher [6, 7] à partir de 1922 et par Wright [19] en 1931. Les modèles d'évolution de population ont été généralisés ultérieurement par Cannings [2, 3]. Par simplicité on considère une population haploïde (ce qui correspond à une population asexuée) : chaque individu possède un seul exemplaire de chaque double brin d'ADN. C'est par exemple le cas de la population humaine si l'on s'intéresse à l'ADN mitochondrial. Ce dernier est en fait uniquement transmis par la mère. L'étude de son évolution repose donc sur un modèle de population haploïde, car seule l'évolution de la population féminine conditionne l'évolution de cet ADN. On considère le modèle élémentaire suivant :

- La taille de la population reste constante au cours du temps, égale à N. Cette hypothèse est réaliste quand l'écosystème est stable. (En cas de colonisation ou de changement brusque de l'environnement tel que changement climatique ou épidémie par exemple, la taille de la population peut varier de manière importante. On peut modifier le modèle pour en tenir compte.)
- Les générations ne se chevauchent pas : à chaque instant $k \in \mathbb{N}$, la k-ième génération meurt et donne naissance aux N individus de la $(k+1)$-ième génération. Cette hypothèse est vérifiée par exemple pour les plantes annuelles. (Le modèle de Moran [13] est une version en temps continu du modèle présenté ; il permet de s'affranchir de cette hypothèse.)
- La reproduction est aléatoire. Plus précisément, si on note $a_i^{k+1} \in \{1, \dots, N\}$ le parent de l'individu i de la génération $k+1$, vivant à la génération k, alors les variables aléatoires $(a_i^{k+1}, i \in \{1, \dots, N\}, k \in \mathbb{N})$ sont indépendantes et de même loi uniforme sur $\{1, \dots, N\}$. Tout se passe comme si chaque individu choisissait de manière indépendante son parent dans la génération précédente. En particulier, ce modèle ne permet pas d'appréhender l'évolution de la population en présence d'avantage sélectif.

On note $\nu_i^k = \text{Card}\,\{r \in \{1, \dots, N\}; a_r^{k+1} = i\}$ le nombre d'enfants de l'individu $i \in \{1, \dots, N\}$ de la génération k. Bien sûr les variables aléatoires $\nu^k = (\nu_i^k, i \in \{1, \dots, N\})$ ne sont pas indépendantes car $\sum_{i=1}^N \nu_i^k = N$. Comme chaque enfant de la génération $k+1$ choisit uniformément et indépendamment son parent, on en déduit que la loi de ν^k est la loi multinomiale de paramètre $(N, (1/N, \dots, 1/N))$: pour $j_1, \dots, j_N \in \mathbb{N}$ tel que $\sum_{k=1}^N j_k = N$, on a

$$\mathbb{P}(\nu_1^k = j_1, \dots, \nu_N^k = j_N) = \frac{N!}{j_1! \dots j_N!} \frac{1}{N^N}\,.$$

Les variables aléatoires $(\nu^k, k \geq 0)$ sont indépendantes et de même loi. L'exercice suivant permet de donner une autre représentation du vecteur aléatoire ν^k.

Exercice 7.1.1. Soit $Y_1, \ldots, Y_N$ des variables aléatoires indépendantes de Poisson de paramètre $\theta > 0$.

1. Déterminer en utilisant les fonctions caractéristiques la loi de $\sum_{i=1}^{N} Y_i$.
2. Montrer que ν^k a même loi que $(Y_1, \ldots, Y_N)$ conditionnellement à l'événement $\{\sum_{i=1}^{N} Y_i = N\}$.
3. Vérifier que la loi de ν_i^k est la loi binomiale de paramètre $(N, 1/N)$.

♦

7.2 Étude de l'arbre phylogénique

Le but de ce paragraphe est de calculer, pour le modèle présenté au paragraphe 7.1, le temps d'apparition dans le passé de l'ancêtre commun le plus récent (ACPR) d'un groupe de r individus vivant aujourd'hui.

Le graphique 7.1 présente une réalisation de l'évolution d'une population de taille $N = 5$ sur 10 générations. L'arbre généalogique des individus vivant à la dernière génération est en trait plein. Le temps d'apparition de l'ACPR de toute la génération 10 est de 4 générations : l'ACPR apparaît à la sixième génération.

Le graphique 7.2 présente une réalisation de l'arbre généalogique des individus vivant à la dernière génération (pour une population de $N = 20$ individus) sur 60 générations. Le temps d'apparition de l'ACPR de toute la population est de 40 générations.

7.2.1 Temps d'apparition de l'ancêtre de deux individus

On considère deux individus à l'instant actuel. On désire savoir s'ils possèdent un ancêtre commun. On note $\tau_2 \in \mathbb{N}^* \cup \{+\infty\}$ le nombre de générations écoulées dans le passé pour observer leur premier ancêtre commun, avec la convention que $\tau_2 = +\infty$, si les deux individus n'ont pas d'ancêtre commun.

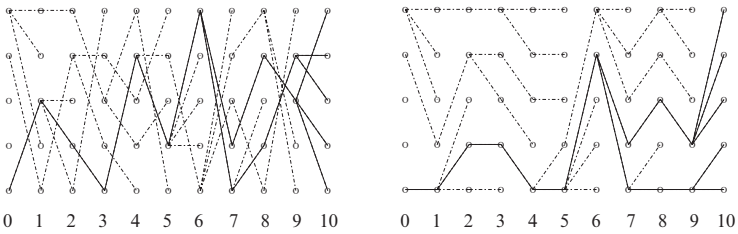

Fig. 7.1. Généalogie d'une population de $N = 5$ individus sur 10 générations. (Le graphique de droite reprend la même généalogie que le graphique de gauche, mais avec une numérotation des individus différente, de sorte que les branches ne se croisent pas)

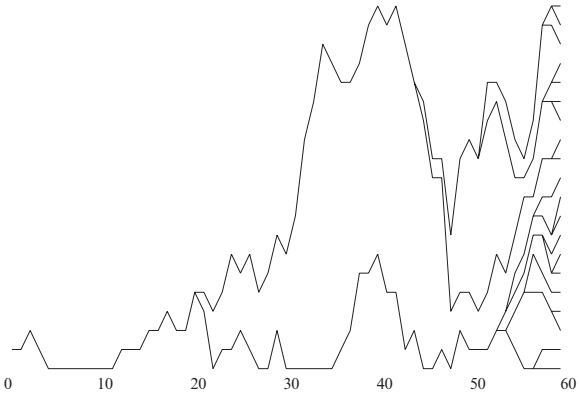

Fig. 7.2. Généalogie d'une population de $N = 20$ individus sur 60 générations

On dit que τ_2 est le temps de coalescence des deux individus. La probabilité que deux individus donnés différents aient des parents différents à la génération précédente (i.e. $\tau_2 > 1$) est

$$\mathbb{P}(\tau_2 > 1) = \frac{N(N-1)}{N^2} = 1 - \frac{1}{N}\,.$$

On a utilisé le fait que chaque individu choisit de manière uniforme son parent dans la génération précédente, et ce indépendamment des autres individus. Comme les choix des parents sont indépendants à chaque génération, on en déduit que

$$\mathbb{P}(\tau_2 > r) = \left(1 - \frac{1}{N}\right)^r.$$

La loi de τ_2 est donc la loi géométrique de paramètre $p_2' = 1/N$. En particulier, τ_2 est p.s. fini.

On étudie le modèle avec l'asymptotique N grand et la normalisation suivante où une unité de temps correspond à N générations. Le temps écoulé depuis l'apparition de l'ACPR est donc τ_2/N.

Le lemme suivant permet de déterminer la loi asymptotique de τ_2/N quand N tend vers l'infini.

Lemme 7.2.1. *Soit* $(V_n, n \geq 1)$ *une suite de variables aléatoires de lois géométriques de paramètres* $(\mu_n, n \geq 1)$ *telle que* $\lim_{n\to\infty} n\mu_n = \mu > 0$. *La suite* $(\dfrac{V_n}{n}, n \geq 1)$ *converge en loi vers* V *de loi exponentielle de paramètre* μ.

Démonstration. La transformée de Laplace de V_n/n, voir la remarque A.2.7, est donnée pour $\alpha \geq 0$ par :

$$\mathbb{E}[\mathrm{e}^{-\alpha V_n/n}] = \frac{\mu_n\,\mathrm{e}^{-\alpha/n}}{1-(1-\mu_n)\,\mathrm{e}^{-\alpha/n}} = \frac{\mu_n}{\mu_n - (1-\mathrm{e}^{\alpha/n})}.$$

On en déduit que

$$\lim_{n\to\infty} \mathbb{E}[\mathrm{e}^{-\alpha V_n/n}] = \frac{\mu}{\mu+\alpha}.$$

On reconnaît, pour le membre de droite, la transformée de Laplace de la loi exponentielle de paramètre μ. La conclusion découle alors du théorème A.3.9. □

Dans le cas d'une population comportant N individus, et dans une échelle de temps où une unité est égale à N générations, le temps d'apparition de l'ACPR pour deux individus donnés est asymptotiquement distribué suivant la loi exponentielle de paramètre 1. Comme l'espérance de la loi exponentielle de paramètre 1 est 1, on en déduit que $\mathbb{E}[\tau_2] \sim N$ quand $N \to \infty$.

7.2.2 Temps d'apparition de l'ancêtre de r individus

On note k le nombre d'ancêtres distincts que les r individus, vivant aujourd'hui, possèdent n générations dans le passé, et $A_1, \ldots, A_k$ la descendance actuelle de chacun des k ancêtres parmi les r individus. Ainsi $Y_n = \{A_1, \ldots, A_k\}$ forme une partition de $\{1, \ldots, r\}$. On note $\mathcal{P}_r$ l'ensemble fini des partitions de $\{1, \ldots, r\}$. Le processus $Y = (Y_n, n \geq 0)$, à valeurs dans $\mathcal{P}_r$, est le processus de coalescence en temps discret associé au processus d'évolution de la population. On remarque que Y_0 est la partition triviale formée de r singletons.

Comme à chaque génération les enfants choisissent leur parent de manière uniforme et indépendante, il s'en suit que le processus Y est une chaîne de Markov à valeurs dans $\mathcal{P}_r$. On note P sa matrice de transition. Pour expliciter P, on introduit un ordre partiel sur les partitions. On dit que $\eta = \{B_1, \ldots, B_\ell\}$ est une partition plus grossière que $\xi = \{A_1, \ldots, A_k\}$ si pour tout $1 \leq i \leq \ell$, B_i est la réunion d'un ou plusieurs éléments de ξ. On note alors $\eta \preccurlyeq \xi$. On note $|\xi| = k$ le nombre de sous-ensembles non vides qui forment la partition ξ. En particulier, si $\eta \preccurlyeq \xi$, on a $|\eta| \leq |\xi|$. On remarque que la partition triviale réduite à tout l'ensemble, $\xi_0 = \{\{1, \ldots, r\}\}$, est un élément absorbant de la chaîne de Markov.

On calcule la matrice de transition $P(\xi, \eta)$ pour $\xi \neq \xi_0$ (i.e. pour ξ tel que $|\xi| \geq 2$) :

- Si η n'est pas plus grossière que ξ, la transition est impossible et on a $P(\xi, \eta) = 0$.
- Si $\eta = \xi$, alors les $|\xi|$ individus ont des ancêtres tous distincts. Le premier individu a N choix possibles pour son ancêtre, le deuxième $N-1$, ... et le dernier $N - |\xi| + 1$. Il existe donc $N!/(N - |\xi|)!$ configurations qui conviennent parmi les $N^{|\xi|}$ configurations équiprobables pour le choix des ancêtres des $|\xi|$ individus. Il vient

$$P(\xi, \xi) = \frac{N!}{N^{|\xi|}(N - |\xi|)!} = \prod_{k=1}^{|\xi|-1} \left(1 - \frac{k}{N}\right) = 1 - \frac{|\xi|(|\xi| - 1)}{2N} + O(N^{-2}).$$

– Si $\eta \preccurlyeq \xi$ et $|\eta| = |\xi| - 1$, alors seulement deux individus fixés ont un ancêtre commun à la génération précédente. Ces deux individus ont N choix possibles pour leur ancêtre commun et les $|\xi| - 2$ autres individus ont respectivement $N-1, \ldots$ et $N - |\xi| + 2$ choix possibles. Il existe donc $N!/(N - |\xi| + 1)!$ configurations qui conviennent. On obtient

$$P(\xi, \eta) = \frac{N!}{N^{|\xi|}(N - |\xi| + 1)!} = \frac{1}{N} \prod_{k=1}^{|\xi| - 2} \left(1 - \frac{k}{N}\right) = \frac{1}{N} + O(N^{-2}),$$

avec la convention $\prod_{k=1}^{0} \left(1 - \frac{k}{N}\right) = 1$. On remarque qu'il existe $\binom{|\xi|}{2} = \frac{|\xi|(|\xi| - 1)}{2}$ choix possibles pour les deux individus qui ont un ancêtre commun à la génération précédente. Ainsi on a $\sum_{\eta \preccurlyeq \xi,\ |\eta| = |\xi| - 1} P(\xi, \eta) = \frac{|\xi|(|\xi| - 1)}{2N} + O(N^{-2})$.

– Comme $\sum_{\eta} P(\xi, \eta) = 1$ et $P(\xi, \xi) + \sum_{\eta \preccurlyeq \xi,\ |\eta| = |\xi| - 1} P(\xi, \eta) = 1 + O(N^{-2})$, on en déduit que pour $\eta \preccurlyeq \xi$ et $|\eta| < |\xi| - 1$, alors

$$P(\xi, \eta) = O(N^{-2}).$$

On introduit les temps successifs de sauts de la chaîne Y. On note $\tau_0 = 0$ et $S'_0 = 0$, et pour $k \geq 1$, on définit par récurrence

$$\tau_k = \inf\{n \geq 1; Y_{S'_{k-1} + n} \neq Y_{S'_{k-1}}\},$$

avec la convention $\inf \emptyset = 0$, et $S'_k = S'_{k-1} + \max(\tau_k, 1)$. Pour $k \in \mathbb{N}$, on pose $Z'_k = Y_{S'_k}$ et on note $R = \inf\{k \geq 0\,; Z'_k = \xi_0\}$, de sorte que $(Z'_k, k \in \{0, \ldots, R\})$ représente les états successifs différents de la chaîne Y.

On rappelle, voir théorème 1.2.2, que $Z' = (Z'_n, n \in \mathbb{N})$ est une chaîne de Markov, appelée chaîne trace, de matrice de transition Q' : pour $\xi \neq \xi_0$ et $\eta \neq \xi$,

$$Q'(\xi, \eta) = \frac{P(\xi, \eta)}{1 - P(\xi, \xi)} = \begin{cases} \dfrac{2}{|\xi|(|\xi| - 1)} + O(N^{-1}) & \text{si } \eta \preccurlyeq \xi \text{ et } |\eta| = |\xi| - 1, \\ O(N^{-1}) & \text{sinon.} \end{cases}$$

On remarque que $R \leq r - 1$ et $Z'_n = \xi_0$ pour $n \geq R$. D'après le théorème 1.2.2, les temps successifs de sauts sont conditionnellement à Z' des variables aléatoires indépendantes, et pour $1 \leq n \leq R$, la loi de τ_n est la loi géométrique de paramètre $1 - \mathbb{P}(Z'_{n-1}, Z'_{n-1}) = \frac{|Z'_{n-1}|(|Z'_{n-1}| - 1)}{2N} + O(N^{-2})$.

On démontre le résultat suivant sur la convergence de la chaîne Z' et des temps successifs de sauts renormalisés.

Théorème 7.2.2. *Soit $r \geq 2$. La suite $\left((Z'_0, \frac{\tau_1}{N}, \ldots, \frac{\tau_{r-1}}{N}, Z'_r), N \geq r\right)$ converge en loi vers $(Z_0, T_r, \ldots, T_2, Z_r)$, où :*

- *$Z = (Z_k, k \geq 0)$ est une chaîne de Markov sur $\mathcal{P}_r$, pour laquelle ξ_0 est un point absorbant et de matrice de transition Q définie : pour $\xi \neq \xi_0$ par*

$$Q(\xi, \eta) = \begin{cases} \dfrac{2}{|\xi|(|\xi|-1)} & \text{si } \eta \preccurlyeq \xi \text{ et } |\eta| = |\xi| - 1, \\ 0 & \text{sinon.} \end{cases}$$

  Et Z_0 est la partition triviale de $\{1, \ldots, r\}$ en r singletons.
- *Les variables aléatoires $(Z, T_r, \ldots, T_2)$ sont indépendantes.*
- *Pour $1 \leq k \leq r$, la loi de T_k est la loi exponentielle de paramètre $\dfrac{k(k-1)}{2}$.*

Remarque 7.2.3. On peut remarquer que $Z_k = \xi_0$ pour $k \geq r$ et que $|Z_k| = \max(r-k, 1)$. En particulier, lors de l'apparition d'un ancêtre commun, cela ne concerne que deux individus seulement. La variable T_k représente le temps à attendre pour observer l'apparition d'un ancêtre commun quand on considère une population de k individus. Cette dernière remarque justifie la numérotation de ces temps d'attente.

Dans l'approche asymptotique N grand, le temps d'attente de l'ACPR de r individus est donc de l'ordre de NW_r, où $W_r = T_2 + \ldots + T_r$. En moyenne, on a

$$\mathbb{E}[W_r] = \sum_{k=2}^{r} \mathbb{E}[T_k] = \sum_{k=2}^{r} \frac{2}{k(k-1)} = \sum_{k=2}^{r} \frac{2}{k-1} - \frac{2}{k} = 2 - \frac{2}{r}.$$

On remarque que la dernière coalescence nécessite une durée T_2 d'espérance 1, qui représente en moyenne plus de la moitié du temps d'atteinte de l'état absorbant ξ_0 (voir les graphiques 7.1 et 7.2). Ainsi, pour N grand, le temps d'apparition de l'ACPR est en moyenne de l'ordre de $2N\left(1 - \dfrac{1}{r}\right)$ générations. ◊

Démonstration du théorème 7.2.2. Rappelons que si τ suit la loi géométrique de paramètre $1-q$, alors on a, pour $\alpha \geq 0$, $\mathbb{E}[\mathrm{e}^{-\alpha\tau}] = \dfrac{1-q}{\mathrm{e}^{\alpha} - q}$. Soit $\alpha_1, \ldots, \alpha_{r-1} \in \mathbb{R}^+$ et $F_0, \ldots, F_r$ des fonctions réelles bornées définies sur $\mathcal{P}_r$. En conditionnant par rapport à $(Z'_0, \ldots, Z'_r)$, on obtient

$$\mathbb{E}[F_0(Z'_0)\,\mathrm{e}^{-\alpha_1\tau_1/N} \cdots \mathrm{e}^{-\alpha_{r-1}\tau_{r-1}/N}\, F_r(Z'_r)] = \mathbb{E}[G'_0(Z'_0) \cdots G'_{r-1}(Z'_{r-1}) F(Z'_r)],$$

où $G'_k(\xi) = F_k(\xi)\dfrac{1 - P(\xi,\xi)}{\mathrm{e}^{\alpha_k/N} - P(\xi,\xi)}$ si $\xi \neq \xi_0$ et $G'_k(\xi_0) = F_k(\xi_0)\,\mathrm{e}^{-\alpha_k/N}$. On en déduit que

$$\mathbb{E}[F_0(Z_0')\,\mathrm{e}^{-\alpha_1\tau_1/N}\cdots\mathrm{e}^{-\alpha_{r-1}\tau_{r-1}/N}\,F_r(Z_r')] = \left(\sum_{\eta_1,\ldots,\eta_r\in\mathcal{P}_r}\prod_{k=0}^{r-1}G_k'(\eta_k)Q'(\eta_k,\eta_{k+1})\right)F(\eta_r),$$

où η_0 est la partition triviale en r singletons. Par passage à la limite, on en déduit que

$$\lim_{N\to\infty}\mathbb{E}[F_0(Z_0')\,\mathrm{e}^{-\alpha_1\tau_1/N}\cdots\mathrm{e}^{-\alpha_{r-1}\tau_{r-1}/N}\,F_r(Z_r')] = \left(\sum_{\eta_1,\ldots,\eta_r\in\mathcal{P}_r}\prod_{k=0}^{r-1}G_k(\eta_k)Q(\eta_k,\eta_{k+1})\right)F(\eta_r), \quad (7.1)$$

où $G_k(\xi) = F_k(\xi)\dfrac{|\xi|(|\xi|-1)}{|\xi|(|\xi|-1)+2\alpha_k}$ si $\xi\neq\xi_0$ et $G_k(\xi_0) = F_k(\xi_0)$. En choisissant $\alpha_1 = \cdots = \alpha_{r-1} = 0$, on en déduit que

$$\lim_{N\to\infty}\mathbb{E}[F_0(Z_0')\cdots F_r(Z_r')] = \mathbb{E}[F_0(Z_0)\cdots F_r(Z_r)],$$

où $Z = (Z_k, k\geq 0)$ est une chaîne de Markov sur $\mathcal{P}_r$ issue de η_0 et de matrice de transition Q donnée dans le théorème. En particulier, on a p.s. que $|Z_k| = \max(r-k,1)$. Ainsi, dans le membre de droite de l'égalité (7.1), seuls les termes tels que $|\eta_k| = r-k$ pour $0\leq k\leq r-1$ et $\eta_r = \xi_0$ ont une contribution non nulle à la somme. On en déduit que

$$G_k(\eta_k) = F_k(\eta_k)\frac{(r-k)(r-k-1)}{(r-k)(r-k-1)+2\alpha_k}.$$

Il vient

$$\lim_{N\to\infty}\mathbb{E}[F_0(Z_0')\,\mathrm{e}^{-\alpha_1\tau_1/N}\cdots\mathrm{e}^{-\alpha_{r-1}\tau_{r-1}/N}\,F_r(Z_r')] = \mathbb{E}[F_0(Z_0)\cdots F_r(Z_r)]\prod_{k=0}^{r-1}\frac{(r-k)(r-k-1)}{(r-k)(r-k-1)+2\alpha_k}.$$

Comme $\dfrac{(r-k)(r-k-1)}{(r-k)(r-k-1)+2\alpha_k}$ est la transformée de Laplace de T_{r-k}, de loi exponentielle de paramètre $(r-k)(r-k-1)/2$, on a

$$\lim_{N\to\infty}\mathbb{E}[F_0(Z_0')\,\mathrm{e}^{-\alpha_1\tau_1/N}\cdots\mathrm{e}^{-\alpha_{r-1}\tau_{r-1}/N}\,F_r(Z_r')] = \mathbb{E}[F_0(Z_0)\cdots F_r(Z_r)]\prod_{j=2}^{r}\mathbb{E}[\mathrm{e}^{-\alpha_{r-j+1}T_j}].$$

Si l'on considère des variables $T_2,\ldots,T_r$ indépendantes entre elles, indépendantes de $Z = (Z_k, k\geq 0)$ et telles que T_k suit la loi exponentielle de paramètre $k(k-1)/2$, alors, d'après les résultats de convergence du para-

graphe A.3.1 en appendice, la suite $\left((Z'_0, \frac{\tau_1}{N}, \ldots, \frac{\tau_{r-1}}{N}, Z'_r), N \leq r\right)$ converge en loi vers $(Z_0, T_r, \ldots, T_2, Z_r)$. Ceci termine la démonstration du théorème. □

7.2.3 Processus de Kingman et commentaires

On conserve les notations du théorème 7.2.2. On pose pour $1 \leq k \leq r-1$, $S_k = \sum_{i=r-k+1}^{r} T_i$, $S_0 = 0$ et $S_r = +\infty$. On définit le processus à temps continu $U_t = Z_k$ pour $t \in [S_k, S_{k+1}[$. Ce processus est appelé processus de coalescence de Kingman [10]. (Il s'agit d'une chaîne de Markov à temps continu, voir le Chap. 8, et plus précisément le paragraphe 8.1). Plus généralement il peut être défini pour $r = +\infty$, de sorte que Z_0 est la partition triviale de $\mathbb{N}^*$ en singletons. On peut généraliser le processus de coalescence, voir Pitman [14], pour obtenir :

- des coalescences multiples qui modélisent le fait que pour le processus limite, plus que deux individus peuvent avoir le même ancêtre commun à la génération précédente,
- des coalescences simultanées qui modélisent le fait que plusieurs ancêtres communs apparaissent simultanément à la même génération.

Les processus de coalescence permettent de modéliser des phénomènes en chimie, physique, astronomie ou biologie, voir par exemple l'étude d'Aldous [1]. Le chapitre 12 est en partie consacré à la présentation et à la résolution d'équations de coagulation qui correspondent à des phénomènes de coalescence.

D'autres mécanismes de reproduction que « chaque enfant choisit indépendamment et uniformément son parent » permettent d'obtenir le processus de coalescence de Kingman ou des processus de coalescence plus généraux. L'étude de ces processus attire beaucoup l'attention depuis la fin des années 1990 et le début des années 2000. Voir par exemple les travaux de Möhle [11], Möhle et Sagitov [12], Sagitov [15], Schweinsberg [16] et leurs références.

7.3 Le modèle de Wright-Fisher

L'exemple suivant permet d'introduire un peu de vocabulaire.

Exemple 7.3.1. Un allèle est une version d'un gène. Ainsi chez l'homme, le gène qui code pour le groupe sanguin possède trois allèles différents : A, B et O. Étant donné que l'homme est diploïde, c'est-à-dire qu'il possède deux exemplaires de chaque chromosome, il existe six génotypes différents : AA, AB, AO, BB, BO, OO. En fait comme l'allèle O est récessif (en présence d'un allèle A ou B, l'allèle O n'est pas exprimé), on distingue seulement quatre phénotypes ou classes différentes de groupes sanguins : A, B, O et AB. ◊

Le modèle de Wright-Fisher concerne l'étude de l'évolution de la répartition de deux allèles, A et a, au sein d'une population. Le modèle d'évolution de la population haploïde utilisé est celui décrit au paragraphe 7.1.

On note X_k le nombre d'allèles A présents à la génération k dans la population. Comme l'évolution de la population à l'instant $k+1$ ne dépend des générations passées qu'au travers de la génération k, il est clair que $(X_k, k \geq 0)$ est une chaîne de Markov homogène. Si $X_k = i$, chaque enfant de la génération $k+1$ a une probabilité i/N d'avoir un parent possédant l'allèle A. Chaque enfant choisissant son parent de manière indépendante, on en déduit que conditionnellement à $X_k = i$, la loi de X_{k+1} est une loi binomiale de paramètre $(N, i/N)$. La matrice de transition, P, est donc : pour $i, j \in \{0, \ldots, N\}$

$$P(i,j) = \mathbb{P}(X_{k+1} = j | X_k = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}.$$

On remarque que si $X_k = 0$ (resp. $X_k = N$), alors pour tout $n \geq k$, $X_n = 0$ (resp. $X_n = N$). Les états 0 et N sont des états absorbants. Si à un instant donné la diversité disparaît, elle ne réapparaît plus. On note τ le premier instant de disparition de la diversité

$$\tau = \inf\{k \geq 0, X_k \in \{0, N\}\},$$

avec la convention $\inf \emptyset = +\infty$.

Lemme 7.3.2. *La variable aléatoire τ est p.s. finie. De plus, on a $\mathbb{P}(X_\tau = N | X_0 = i) = i/N$ pour tout $i \in \{0, \ldots, N\}$.*

Ainsi la probabilité que toute la population finisse par posséder l'allèle A (resp. a) est égale à la proportion initiale de l'allèle A (resp. a).

On remarque que dans ce modèle la diversité disparaît p.s. en temps fini. Pour tenir compte du fait que l'on observe de la diversité dans les populations actuelles, il est nécessaire de compléter le modèle de Wright-Fisher, en tenant compte par exemple des mutations. Une modélisation élémentaire des mutations est présentée au paragraphe suivant.

Démonstration. Si on pose $p = \min_{x \in [0,1]} x^N + (1-x)^N = 2^{-N+1}$, il vient pour tout $i \in \{1, \ldots, N-1\}$,

$$\mathbb{P}(\tau = 1 | X_0 = i) = \mathbb{P}(X_1 \in \{0, N\} | X_0 = i) = \left(\frac{i}{N}\right)^N + \left(1 - \frac{i}{N}\right)^N \geq p.$$

On pose $q = 1 - p$. Pour tout $i \in \{1, \ldots, N-1\}$, on a $\mathbb{P}(\tau > 1 | X_0 = i) \leq q$. En utilisant la propriété de Markov pour X, il vient pour $k \geq 2$,

$$\begin{aligned}
\mathbb{P}(\tau > k | X_0 = i) &= \sum_{j=1}^{N-1} \mathbb{P}(X_k \notin \{0, N\}, X_{k-1} = j | X_0 = i) \\
&= \sum_{j=1}^{N-1} \mathbb{P}(X_k \notin \{0, N\} | X_{k-1} = j, X_0 = i) \mathbb{P}(X_{k-1} = j | X_0 = i)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{N-1} \mathbb{P}(X_1 \notin \{0, N\} | X_0 = j) \mathbb{P}(X_{k-1} = j | X_0 = i) \\
&\leq q \sum_{j=1}^{N-1} \mathbb{P}(X_{k-1} = j | X_0 = i) \\
&= q\mathbb{P}(\tau > k-1 | X_0 = i).
\end{aligned}
$$

Par récurrence, on en déduit que

$$\mathbb{P}(\tau > k | X_0 = i) \leq q^{k-1} \mathbb{P}(\tau > 1 | X_0 = i) = q^k.$$

Comme $\{\tau = +\infty\}$ est la limite décroissante des événements $\{\tau > k\}$ quand k tend vers l'infini, on déduit de la propriété de convergence monotone des probabilités que pour tout $i \in \{1, \ldots, N-1\}$,

$$\mathbb{P}(\tau = +\infty | X_0 = i) = \lim_{k \to \infty} \mathbb{P}(\tau > k | X_0 = i) = 0.$$

Ainsi τ est p.s. fini. On remarque que

$$X_n = X_\tau \mathbf{1}_{\{\tau \leq n\}} + X_n \mathbf{1}_{\{\tau > n\}}.$$

Comme τ est fini p.s., on a donc p.s. $\lim_{n \to \infty} X_n = X_\tau$. Par convergence dominée ($0 \leq X_n \leq N$ pour tout $n \in \mathbb{N}$), il vient pour tout $i_0 \in \{1, \ldots, N-1\}$,

$$\lim_{n \to \infty} \mathbb{E}[X_n | X_0 = i_0] = \mathbb{E}[X_\tau | X_0 = i_0].$$

Comme, conditionnellement à $X_{n-1} = i$, X_n est distribué suivant la loi de Bernoulli de paramètre $(N, i/N)$, on a

$$\mathbb{E}[X_n | X_{n-1} = i, X_0 = i_0] = \mathbb{E}[X_n | X_{n-1} = i] = N \frac{i}{N} = i.$$

On en déduit que

$$
\begin{aligned}
\mathbb{E}[X_n | X_0 = i_0] &= \sum_{i=0}^{N} \mathbb{E}[X_n | X_{n-1} = i, X_0 = i_0] \mathbb{P}(X_{n-1} = i | X_0 = i_0) \\
&= \sum_{i=0}^{N} i \mathbb{P}(X_{n-1} = i | X_0 = i_0) \\
&= \mathbb{E}[X_{n-1} | X_0 = i_0],
\end{aligned}
$$

et par récurrence $\mathbb{E}[X_n | X_0 = i_0] = \mathbb{E}[X_0 | X_0 = i_0] = i_0$. On a donc montré que $\mathbb{E}[X_\tau | X_0 = i_0] = i_0$. De plus comme $X_\tau \in \{0, N\}$, on a $\mathbb{E}[X_\tau | X_0 = i_0] = N\mathbb{P}(X_\tau = N | X_0 = i_0)$. On en déduit donc que $\mathbb{P}(X_\tau = N | X_0 = i_0) = i_0/N$. □

Remarque 7.3.3. Il est intéressant d'étudier le temps moyen où disparaît la diversité : $t_i = \mathbb{E}[\tau|X_0 = i]$, pour $i \in \{0, \dots, N\}$. Bien sûr, on a $t_0 = t_N = 0$. Pour $1 \leq i \leq N-1$, on remarque que sur l'événement $\{X_0 = i\}$, $\tau = 1 + \inf\{k \geq 0, X_{k+1} \in \{0, N\}\}$, de sorte que, en utilisant la propriété de Markov, on a

$$\mathbb{E}[\tau|X_1 = j, X_0 = i] = 1 + \mathbb{E}[\tau|X_0 = j].$$

Il vient

$$\begin{aligned} t_i &= \sum_{j\in\{0,\dots,N\}} \mathbb{E}[\tau \mathbf{1}_{\{X_1=j\}}|X_0 = i] \\ &= \sum_{j\in\{0,\dots,N\}} \mathbb{E}[\tau|X_1 = j, X_0 = i]\mathbb{P}(X_1 = j|X_0 = i) \\ &= \sum_{j\in\{0,\dots,N\}} (1 + t_j)P(i, j) \\ &= 1 + \sum_{j\in\{0,\dots,N\}} P(i, j)t_j. \end{aligned}$$

Comme 0 et N sont des états absorbants, on a $t_i = \sum_{j\in\{0,\dots,N\}} P(i, j)t_j = 0$ pour $i \in \{0, N\}$. Si on note $T = (t_0, \dots, t_N)$, e_0 (resp. e_N) le vecteur de $\mathbb{R}^{N+1}$ ne comportant que des 0 sauf un 1 en première (resp. dernière) position, et $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^{N+1}$, on a

$$T = PT + \mathbf{1} - e_0 - e_N.$$

Le calcul des temps moyens où disparaît la diversité se fait donc en résolvant un système linéaire. Pour N grand, on a l'approximation suivante (voir [5]) : pour $x \in [0, 1]$,

$$\mathbb{E}[\tau|X_0 = [Nx]] \sim -2N\left(x\log(x) + (1 - x)\log(1 - x)\right),$$

où $[Nx]$ désigne la partie entière de Nx. Cette approximation est illustrée par le graphique 7.3. ◊

7.4 Modélisation des mutations

Pour rendre compte de la diversité des individus dans une population, on peut compléter le modèle de Wright-Fisher en tenant compte des possibilités de mutation de l'ADN, en particulier lors de sa réplication. Les mutations entraînent une diversification des enfants d'un même individu.

Rappelons qu'un chromosome est un double brin en hélice d'ADN. Chaque brin est constitué de plusieurs milliers de bases. Il existe quatre bases, ou nucléotides, différentes : Adénine (`A`), Guanine (`G`), Cytosine (`C`) et Thymine (`T`). Les deux brins d'ADN mettent en correspondance les bases `A` et `T` (par

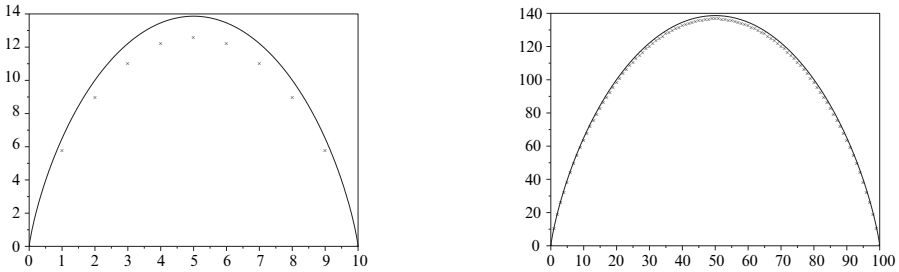

Fig. 7.3. Temps moyen de disparition de la diversité ($k \to \mathbb{E}[\tau | X_0 = k]$) et son approximation continue, $Nx \to 2N\,(x\log(x) + (1-x)\log(1-x))$, pour $N = 10$ (à gauche) et $N = 100$ (à droite)

deux liaisons d'hydrogène) et les bases `C` et `G` (par trois liaisons d'hydrogène). Les bases `A` et `G` (resp. `C` et `T`) ont des propriétés physiques proches ; elles sont appelées purines (resp. pyrimidines).

Le taux de mutation d'une base, i.e. le remplacement par erreur lors de la réplication d'une base par une autre, est très faible, de l'ordre de 10^{-5} à 10^{-8} mutation par base et par génération. On suppose pour simplifier que :

- Le taux ne dépend pas de la base concernée. (Une mutation qui conserve le type purine ou pyrimidine est appelée transition, sinon on parle de transversion. Les transversions sont plus rares que les transitions.)
- Le taux ne dépend pas de la position dans l'ADN. (Certaines mutations ne sont pas viables. D'autres mutations ne concernent pas les régions codantes de l'ADN. D'autres mutations encore sont silencieuses : la mutation qui affecte le codon, suite de trois bases qui code pour un des 20 acides aminés, ne change pas l'acide aminé concerné. Il est clair que le taux de mutation n'est pas constant sur l'ADN ! Toutefois, si l'on se restreint à des régions de l'ADN qui ont peu ou pas de rôle dans la production de protéines, alors on peut supposer que le taux de mutation est homogène.)
- Le taux est constant au cours du temps.

Ainsi, on suppose donc qu'à chaque génération un individu a une probabilité $\mu > 0$ d'avoir une mutation qui le différencie de son parent.

Si l'on considère une séquence relativement longue, le faible taux d'apparition des mutations fait que la probabilité que deux mutations concernent le même site de la séquence est négligeable devant les autres probabilités. On fera donc l'hypothèse simplificatrice que l'on a une infinité d'allèles possibles et que chaque mutation affecte un site différent. Ainsi une mutation donne toujours un nouvel allèle. Ce modèle est appelé « modèle avec une infinité de sites ». Enfin, le taux de mutation étant faible, on suppose que l'on a au plus une seule mutation entre un individu et son parent. Le temps d'apparition d'une mutation dans la lignée ancestrale d'un individu suit donc une loi géométrique de paramètre μ. On pose $\theta = 2\mu N$, et on suppose que $\theta = O(1)$. L'utilité de la

constante 2 apparaîtra ultérieurement dans les calculs. L'objectif est d'estimer le paramètre inconnu θ.

On s'intéresse à l'approche asymptotique N grand, et on considère la normalisation suivante : une unité de temps correspond à N générations. À la limite, quand N tend vers l'infini, on déduit du lemme 7.2.1, que le temps d'apparition d'une mutation, R, dans la lignée ancestrale d'un individu suit la loi exponentielle de paramètre $\theta/2$. On peut vérifier que la suite des temps entre les apparitions successives de mutations après R forme une suite de variables aléatoires indépendantes, indépendantes de R, et de même loi exponentielle de paramètre $\theta/2$.

Les deux processus de coalescence et de mutation sont d'origines aléatoires différentes. On les modélise donc par des processus indépendants.

7.4.1 Estimation du taux de mutation I

Pour estimer θ, on suppose que l'on dispose de n séquences d'ADN correspondant à n individus haploïdes. Plusieurs individus peuvent posséder le même allèle, i.e. la même séquence d'ADN. On note K_n le nombre d'allèles distincts. L'étude de la loi de K_n permettra de donner une estimation de θ et un intervalle de confiance, voir la proposition 7.4.3.

Proposition 7.4.1. *On a*

$$K_n = \sum_{k=1}^{n} \eta_k,$$

où les variables aléatoires $(\eta_k, k \geq 1)$ sont indépendantes de loi de Bernoulli de paramètre $\theta/(k-1+\theta)$.

La démonstration de cette proposition et l'interprétation des variables aléatoires $(\eta_k, k \geq 1)$ sont données à la fin de ce paragraphe. Le graphique 7.4 donne l'histogramme de la loi de K_n obtenu par simulation.

La proposition suivante présente une estimation de θ à l'aide de l'estimateur K_n. Voir la définition 5.2.5 pour les propriétés des estimateurs.

Proposition 7.4.2. *Pour $n \geq 1$, on a*

$$\mathbb{E}[K_n] = \theta \log(n) + O(1) \quad \textit{et} \quad \mathrm{Var}(K_n) = \theta \log(n) + O(1).$$

La suite $\left(\dfrac{K_n}{\log(n)}, n \geq 1\right)$ est un estimateur faiblement convergent de θ : la suite $\left(\dfrac{K_n}{\log(n)}, n \geq 1\right)$ converge en probabilité vers θ.

Démonstration. On a $\mathbb{E}[K_n] = \sum_{k=1}^{n} \mathbb{E}[\eta_k] = \sum_{k=1}^{n} \frac{\theta}{k-1+\theta}$. Comme

$$\int_{\theta}^{n+\theta} \frac{dx}{x} \leq \sum_{k=1}^{n} \frac{1}{k-1+\theta} \leq \frac{1}{\theta} + \int_{\theta}^{n-1+\theta} \frac{dx}{x},$$

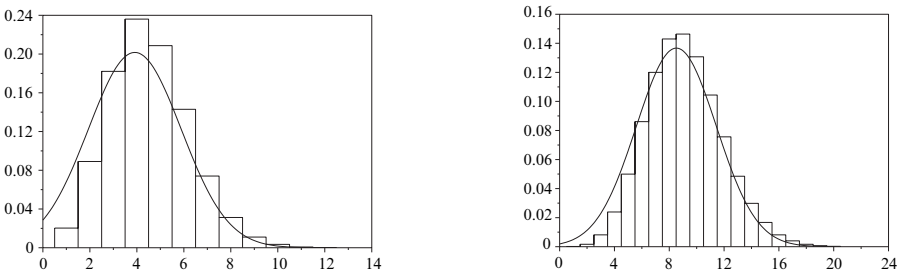

Fig. 7.4. Histogrammes de la loi de K_n pour $\theta = 1$, $n = 50$ (à gauche) et $n = 5\,000$ (à droite), obtenus à l'aide de 100 000 simulations de K_n, et densité de la loi gaussienne de moyenne et de variance $\theta \log(n)$

on a donc l'encadrement

$$\theta \log(n) + \theta \log\left(1 + \frac{\theta}{n}\right) - \theta \log(\theta) \leq \mathbb{E}[K_n] \leq \theta \log(n) + 1 + \theta \log\left(1 + \frac{\theta - 1}{n}\right) - \theta \log(\theta).$$

En particulier, il vient $\mathbb{E}[K_n] = \theta \log(n) + O(1)$.

On calcule la variance de K_n. En utilisant l'indépendance des variables aléatoires $\eta_1, \ldots, \eta_n$, il vient

$$\begin{aligned} \mathrm{Var}(K_n) &= \sum_{k=1}^{n} \mathrm{Var}(\eta_k) \\ &= \sum_{k=1}^{n} \frac{\theta}{k-1+\theta}\left(1 - \frac{\theta}{k-1+\theta}\right) \\ &= \mathbb{E}[K_n] - \theta^2 \sum_{k=1}^{n} \frac{1}{(k-1+\theta)^2}. \end{aligned}$$

Comme la série du dernier terme du membre de droite est convergente quand n tend vers l'infini, on en déduit que $\mathrm{Var}(K_n) = \theta \log(n) + O(1)$.

En utilisant l'inégalité de Tchebychev (A.2), on a pour tout $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}\left(\left|\frac{K_n}{\log(n)} - \theta\right| \geq \varepsilon\right) &\leq \frac{\mathbb{E}\left[\left(\frac{K_n}{\log(n)} - \theta\right)^2\right]}{\varepsilon^2} \\ &= \frac{\mathrm{Var}(K_n) + (\mathbb{E}[K_n] - \theta \log(n))^2}{\log(n)^2 \varepsilon^2} \\ &= \frac{1}{\varepsilon^2} O(\log(n)^{-1}). \end{aligned}$$

D'après la définition 5.2.5, l'estimateur est faiblement convergent. □

Proposition 7.4.3. *La suite* $\left(\dfrac{K_n - \theta\log(n)}{\sqrt{K_n}}, n \geq 1\right)$ *converge en loi vers* G *de loi gaussienne centrée réduite* $\mathcal{N}(0,1)$.

En particulier, si z est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$, alors on obtient un intervalle de confiance de niveau asymptotique $1-\alpha$ pour θ :

$$\Delta_n = \left[\frac{K_n}{\log(n)} \pm \frac{z\sqrt{K_n}}{\log(n)}\right],$$

c'est-à-dire $\lim_{n\to\infty} \mathbb{P}(\theta \in \Delta_n) = 1-\alpha$.

Démonstration. On utilise le théorème B.1 qui est une variante du théorème central limite. On pose $X_k = \eta_k - \mathbb{E}[\eta_k]$. On remarque que $\mathbb{E}[X_k^2] = \mathrm{Var}(\eta_k)$. Les conditions 1 et 2 du théorème B.1 sont vérifiées, d'après ce qui précède. Pour démontrer la condition 3, en remarquant que les variables aléatoires η_k sont des variables aléatoires de Bernoulli, on a

$$\left|X_k^3\right| = |\eta_k - \mathbb{E}[\eta_k]|^3 = \left(1 - \frac{\theta}{k-1+\theta}\right)^3 \eta_k + \left(\frac{\theta}{k-1+\theta}\right)^3 (1-\eta_k).$$

On en déduit donc que

$$\begin{aligned}\mathbb{E}[|X_k|^3] &= \left(1 - \frac{\theta}{k-1+\theta}\right)^3 \frac{\theta}{k-1+\theta} + \left(\frac{\theta}{k-1+\theta}\right)^3 \left(1 - \frac{\theta}{k-1+\theta}\right) \\ &\leq \frac{\theta}{k-1+\theta},\end{aligned}$$

où l'on utilise que $(1-x)((1-x)^2+x^2) \leq 1$ pour $x \in [0,1]$, avec $x = \theta/(k-1+\theta)$. Ainsi, on a $\sum_{k=1}^n \mathbb{E}[|X_k|^3] \leq \mathbb{E}[K_n]$, puis

$$\frac{\sum_{k=1}^n \mathbb{E}[X_k^3]}{(\sum_{k=1}^n \mathbb{E}[X_k^2])^{3/2}} \leq \frac{\mathbb{E}[K_n]}{\mathrm{Var}(K_n)^{3/2}} = O(\log(n)^{-1/2}).$$

La condition 3 du théorème B.1 est donc vérifiée. Ceci implique que la suite $\left(\dfrac{K_n - \mathbb{E}[K_n]}{\sqrt{\mathrm{Var}(K_n)}}, n \geq 1\right)$ converge en loi vers G de loi $\mathcal{N}(0,1)$. Comme $\mathbb{E}[K_n] = \theta\log(n) + O(1)$ et $\lim_{n\to\infty} K_n/\mathrm{Var}(K_n) = 1$ en probabilité, on déduit du théorème de Slutsky A.3.12 que la suite $\left(\dfrac{K_n - \theta\log(n)}{\sqrt{K_n}}, n \geq 1\right)$ converge en loi vers G de loi $\mathcal{N}(0,1)$. □

La suite de ce paragraphe est consacrée à la démonstration de la proposition 7.4.1. On établit le lemme préliminaire suivant.

Lemme 7.4.4. *Soit $V_1, \dots, V_n$ une suite de variables aléatoires indépendantes de loi exponentielle de paramètres respectifs $\mu_1, \dots, \mu_n$. On définit $V = \min_{1 \le i \le n} V_i$ et p.s. il existe un unique indice aléatoire, I, tel que $V = V_I$. Les variables aléatoires V et I sont indépendantes. De plus la loi de V est la loi exponentielle de paramètre $\sum_{i=1}^n \mu_i$, et pour tout $i_0 \in \{1, \dots, n\}$, on a $\mathbb{P}(I = i_0) = \dfrac{\mu_{i_0}}{\sum_{i=1}^n \mu_i}$.*

Démonstration. Comme les variables aléatoires $V_1, \dots, V_n$ sont indépendantes et continues, la probabilité que deux d'entre elles prennent la même valeur est nulle. En particulier, on en déduit que I est uniquement déterminé avec probabilité 1.

Pour $t \ge 0$, $i_0 \in \{1, \dots, n\}$, on a

$$\begin{aligned}
\mathbb{P}(I = i_0, V > t) &= \mathbb{P}(t < V_{i_0} < V_i,\ \forall i \neq i_0) \\
&= \mathbb{E}[\mathbf{1}_{\{t<V_{i_0}\}} \prod_{i \neq i_0} \mathbf{1}_{\{V_{i_0}<V_i\}}] \\
&= \int_{\mathbb{R}_+^n} dv_1 \dots dv_n \left(\prod_{i=1}^n \mu_i \, \mathrm{e}^{-\mu_i v_i} \right) \mathbf{1}_{\{t<v_{i_0}\}} \prod_{i \neq i_0} \mathbf{1}_{\{v_{i_0}<v_i\}} \\
&= \int \prod_{i \neq i_0} \mathrm{e}^{-\mu_i v_{i_0}} \ \mu_{i_0} \, \mathrm{e}^{-\mu_{i_0} v_{i_0}} \, \mathbf{1}_{\{t<v_{i_0}\}} \, dv_{i_0} \\
&= \frac{\mu_{i_0}}{\sum_{i=1}^n \mu_i} \, \mathrm{e}^{-(\sum_{i=1}^n \mu_i)t}.
\end{aligned}$$

En prenant $t = 0$, on obtient la loi de I. En sommant sur $i_0 \in \{1, \dots, n\}$, on obtient que pour $t \ge 0$,

$$\mathbb{P}(V > t) = \mathrm{e}^{-\sum_{i=1}^n \mu_i t}.$$

On reconnaît pour V la loi exponentielle de paramètre $\sum_{i=1}^n \mu_i$. Enfin, comme pour tous $t \ge 0$, $i_0 \in \{1, \dots, n\}$, on a

$$\mathbb{P}(I = i_0, V > t) = \mathbb{P}(I = i_0)\mathbb{P}(V > t),$$

on en déduit que les variables aléatoires I et V sont indépendantes. □

Démonstration de la proposition 7.4.1. Pour un groupe de n individus, on s'intéresse au premier temps dans le passé, U_n, d'apparition d'une coalescence (i.e. d'un ancêtre commun) ou d'une mutation. Tout se passe comme si U_n était le minimum entre T_n, premier temps de coalescence de loi exponentielle de paramètre $n(n-1)/2$, et $R_1, \dots, R_n$, premiers temps de mutation des individus $1, \dots, n$ de loi exponentielle de paramètre $\theta/2$.

Les variables aléatoires $T_n, R_1, \dots, R_n$ sont indépendantes. D'après le lemme précédent, la loi de U_n est la loi exponentielle de paramètre

$n(n-1+\theta)/2$. De plus la probabilité que ce premier phénomène soit une coalescence est $\dfrac{n(n-1)/2}{n(n-1+\theta)/2} = \dfrac{n-1}{n-1+\theta}$. La probabilité que ce phénomène soit une mutation est donc

$$1 - \frac{n-1}{n-1+\theta} = \frac{\theta}{n-1+\theta}.$$

On pose $\eta_n = 1$ si ce premier phénomène est une mutation, et $\eta_n = 0$ sinon. La variable aléatoire η_n est donc une variable aléatoire de Bernoulli de paramètre $\theta/(n-1+\theta)$.

- Si $\eta_n = 0$, i.e. le premier phénomène est une coalescence, alors le nombre d'ancêtres à l'instant U_n est $n-1$. Et le nombre d'allèles distincts dans ce groupe de $n-1$ personnes est $K_{n-1} = K_n$.
- Si $\eta_n = 1$, i.e. le premier phénomène est une mutation, alors l'individu qui a muté à l'instant U_n est à l'origine d'un allèle différent. Il correspond, dans la population initiale de n individus, à la présence d'un allèle distinct de tous les autres. De plus cet allèle est présent une seule fois dans la population des n individus. En retirant, à l'instant U_n, cet ancêtre de la population des n ancêtres, on obtient une population de $n-1$ individus possédant $K_{n-1} = K_n - 1$ allèles distincts.

Dans tous les cas, on a obtenu $K_n = K_{n-1} + \eta_n$, et on considère à l'instant U_n une population de $n-1$ individus possédant K_{n-1} allèles différents. Par récurrence descendante, on obtient que $K_n = K_1 + \sum_{k=2}^{n} \eta_k$. Les variables aléatoires η_k sont des variables de Bernoulli de paramètre $\theta/(k-1+\theta)$. On a $K_1 = 1$ (un seul allèle distinct dans une population d'un seul individu) et donc p.s. $K_1 = \eta_1$, où η_1 est une variable de Bernoulli de paramètre 1. On en déduit donc que $K_n = \sum_{k=1}^{n} \eta_k$.

Enfin, il reste à vérifier que les variables aléatoires $\eta_1, \ldots, \eta_n$ sont indépendantes. Pour cela on remarque que l'évolution dans le passé de la population des $n-1$ individus avant l'instant U_n est indépendante du phénomène observé, coalescence ou mutation, à l'instant U_n. On en déduit donc que η_n est indépendant des variables aléatoires $\eta_1, \ldots, \eta_{n-1}$. Par récurrence descendante, on en conclut que les variables aléatoires $\eta_1, \ldots, \eta_n$ sont indépendantes. □

7.4.2 Estimation du taux de mutation II

En fait les données dont on dispose en comparant les séquences d'ADN sont plus riches que la donnée du nombre d'allèles différents. En effet on dispose, pour un échantillon de n personnes, du nombre de sites, S_n, où ont eu lieu les mutations.

On remarque que dans le modèle initial, le nombre de mutations observées au cours de $[Nt]$ générations, pour les ancêtres d'un individu, suit une loi binomiale de paramètre $([Nt], \theta/2N)$. Quand $N \to \infty$, la loi binomiale converge en loi, dans ce cas, vers une variable de loi de Poisson de paramètre $\theta t/2$ (cf. l'exemple 6.3.1).

On note $S(t)$ le nombre de mutations durant une période t pour la lignée d'un individu. On pose $R_0 = 0$. On note R_j le temps qui s'est écoulé entre la $j-1$-ième mutation et la j-ème pour $j \geq 2$, R_1 étant le temps d'attente de la première mutation. Dans le modèle de mutation présenté, les variables aléatoires $(R_j, j \geq 1)$ sont indépendantes et de loi exponentielle de paramètre $\theta/2$. Par définition, on a pour $t \geq 0$,

$$S(t) = \inf\{k \geq 0\,; \sum_{j=1}^{k+1} R_j > t\}.$$

Remarque 7.4.5. D'après la définition 8.14, le processus $S = (S(t), t \geq 0)$ est un processus de Poisson de paramètre $\theta/2$. En particulier, on vérifie ainsi que la loi de $S(t)$ est la loi de Poisson de paramètre $\theta t/2$ (cf. la démonstration élémentaire du (ii) de la proposition 8.4.2). ◊

Soit $k \geq 2$. On note Y_k le nombre de mutations observées dans une population de k individus avant le premier temps de coalescence T_k. Le lemme suivant établit la loi de Y_k.

Lemme 7.4.6. *La variable Y_k a même loi que $\tilde{G} - 1$, où $\tilde{G}$ suit la loi géométrique de paramètre $p = \dfrac{k-1}{k-1+\theta}$.*

Démonstration. On a $Y_k = Y_k^{(1)} + \ldots + Y_k^{(k)}$, où $Y_k^{(i)}$ est le nombre de mutations de l'individu i avant T_k. En particulier, $Y_k^{(i)}$ a même loi que $S(T_k)$.

Avant toute coalescence, les processus de mutation des différents individus sont indépendants. On en déduit donc que $(Y_k^{(1)}, \ldots, Y_k^{(k)})$ a même loi que $(S^{(1)}(T_k), \ldots, S^{(k)}(T_k))$, où le processus $(S^{(i)}(t), t \geq 0)$ est défini par

$$S^{(i)}(t) = \inf\{k \geq 0\,; \sum_{j=1}^{k+1} R_j^{(i)} > t\},$$

et les variables $(R_j^{(i)}, j \geq 1, i \geq 1)$ sont indépendantes de loi exponentielle de paramètre $\theta/2$. En particulier les variables $S^{(1)}(t), \ldots, S^{(k)}(t)$ sont indépendantes. On a de plus que T_k est indépendant de la suite $(R_j^{(i)}, j \geq 1,$ $i \geq 1)$ et suit la loi exponentielle de paramètre $\frac{k(k-1)}{2}$. En écrivant l'événement

$$\{S^{(i)}(T_k) = r_i\} = \left\{ \sum_{l=1}^{r_i} R_l^{(i)} \leq T_k < \sum_{l=1}^{r_i+1} R_l^{(i)} \right\},$$

avec la convention $\sum_{l=1}^{0} R_l^{(i)} = 0$, et en utilisant l'indépendance des variables aléatoires, on obtient pour $r_1, \ldots, r_k \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{P}(S^{(1)}(T_k) = r_1, \ldots, S^{(k)}(T_k) = r_k) \\
&\quad = \int_0^\infty \frac{k(k-1)}{2}\, \mathrm{e}^{-k(k-1)t/2}\, \mathbf{1}_{\{t>0\}} \mathbb{P}(S^{(1)}(t) = r_1, \ldots, S^{(k)}(t) = r_k)\, dt.
\end{aligned}$$

On en déduit la fonction caractéristique de Y_k : pour $u \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{iuY_k}] &= \mathbb{E}[\mathrm{e}^{iu\sum_{i=1}^{k} S^{(i)}(T_k)}] \\
&= \int_0^\infty \frac{k(k-1)}{2}\,\mathrm{e}^{-k(k-1)t/2}\,\mathbf{1}_{\{t>0\}}\mathbb{E}[\mathrm{e}^{iu\sum_{i=1}^{k} S^{(i)}(t)}]\,dt \\
&= \int_0^\infty \frac{k(k-1)}{2}\,\mathrm{e}^{-k(k-1)t/2}\,\mathbf{1}_{\{t>0\}}\prod_{i=1}^{k}\mathbb{E}[\mathrm{e}^{iuS^{(i)}(t)}]\,dt \\
&= \int_0^\infty \frac{k(k-1)}{2}\,\mathrm{e}^{-k(k-1)t/2}\,\mathbf{1}_{\{t>0\}}\,\mathrm{e}^{-k\theta t(1-\exp(iu))/2}\;dt \\
&= \frac{k(k-1)}{2}\,\frac{1}{k(k-1)/2 + k\theta(1-\exp(iu))/2} \\
&= \frac{p}{1-(1-p)\,\mathrm{e}^{iu}},
\end{aligned}$$

avec $p = \dfrac{k-1}{k-1+\theta}$. On a utilisé l'indépendance des variables $S^{(1)}(t)$,..., $S^{(k)}(t)$ pour la troisième égalité, et pour la quatrième égalité le fait que $S^{(i)}(t)$ est distribué suivant la loi de Poisson de paramètre $\theta t/2$ d'après la remarque 7.4.5. D'autre part, si $\tilde{G}$ est de loi géométrique de paramètre p, on a

$$\mathbb{E}[\mathrm{e}^{iu(\tilde{G}-1)}] = \sum_{n\geq 1} p(1-p)^{n-1}\,\mathrm{e}^{iu(n-1)} = \frac{p}{1-(1-p)\,\mathrm{e}^{iu}}.$$

On en déduit donc que Y_k et $\tilde{G}-1$ ont même loi. □

On compte Y_k mutations pendant la période aléatoire de durée T_k, où le nombre d'ancêtres des n individus considérés est égal à k. Comme les durées, $T_2, \ldots, T_n$, des périodes sont aléatoires et indépendantes, on en déduit que les variables $Y_2, \ldots, Y_n$ sont indépendantes. Leur loi est décrite par le lemme 7.4.6. Enfin, le nombre total de mutations observées sur un échantillon de n individus, est $S_n = Y_2 + \ldots + Y_n$. On calcule l'espérance et la variance de S_n.

On rappelle que si $\tilde{G}$ est une variable géométrique de paramètre p, alors $\mathbb{E}[\tilde{G}] = 1/p$ et $\mathrm{Var}(\tilde{G}) = (1-p)/p^2$ (voir le tableau A.1 page 396). On en déduit que

$$\mathbb{E}[S_n] = \sum_{k=2}^{n}\mathbb{E}[Y_k] = \sum_{k=2}^{n}\left(\frac{k-1+\theta}{k-1} - 1\right) = \theta a_n,$$

où $a_n = \displaystyle\sum_{k=1}^{n-1}\frac{1}{k}$ et

$$\mathrm{Var}(S_n) = \sum_{k=2}^{n}\mathrm{Var}(Y_k) = \theta\sum_{k=2}^{n}\frac{k-1+\theta}{(k-1)^2} = \theta\sum_{k=1}^{n-1}\frac{1}{k} + \theta^2\sum_{k=1}^{n-1}\frac{1}{k^2} = \theta a_n + \theta^2 b_n,$$

où $b_n = \sum_{k=1}^{n-1} \frac{1}{k^2}$. Watterson [18] a proposé $\frac{S_n}{a_n}$ comme estimateur sans biais de θ. On remarque que $a_n \sim \log(n)$ quand n tend vers l'infini et que la suite $(b_n, n \geq 2)$ converge.

Proposition 7.4.7. *L'estimateur de θ, $(S_n/a_n, n \geq 2)$, est faiblement convergent (i.e. la suite $(S_n/a_n, n \geq 2)$ converge en probabilité vers θ). La suite $\left(\frac{S_n - \theta a_n}{\sqrt{S_n}}, n \geq 1\right)$ converge en loi vers G de loi gaussienne centrée réduite $\mathcal{N}(0,1)$.*

La convergence en loi de la proposition est illustrée par le graphique 7.5. La démonstration de cette proposition est indiquée dans l'exercice suivant.

Exercice 7.4.8. Le but de cet exercice, est de démontrer la proposition 7.4.7.

1. En utilisant l'inégalité de Tchebychev, et en s'inspirant des arguments utilisés dans la démonstration de la proposition 7.4.2, vérifier que l'estimateur de Watterson est faiblement convergent.
2. Soit $\tilde{G}$ une variable aléatoire de loi géométrique de paramètre $p \in]0,1[$. Vérifier que $\mathbb{E}[|\tilde{G}-1|^3] = \mathbb{E}[(\tilde{G}-1)^3] = \frac{6(1-p)^2}{p^3} + \frac{1-p}{p}$. Puis, en utilisant l'inégalité $(x+y)^3 \leq 4x^3 + 4y^3$ pour $x, y \geq 0$, montrer que
$$\mathbb{E}\Big[\big|\tilde{G} - \frac{1}{p}\big|^3\Big] \leq 32\frac{1-p}{p^3}.$$
3. On pose $X_k = Y_k - \mathbb{E}[Y_k]$. Montrer que $\mathbb{E}[|X_k|^3] \leq 32\theta(1+\theta)^2 \frac{1}{k-1}$. Puis vérifier les hypothèses du théorème B.1.
4. Démontrer la deuxième partie de la proposition 7.4.7 en s'inspirant des arguments utilisés dans la démonstration de la proposition 7.4.3.

♦

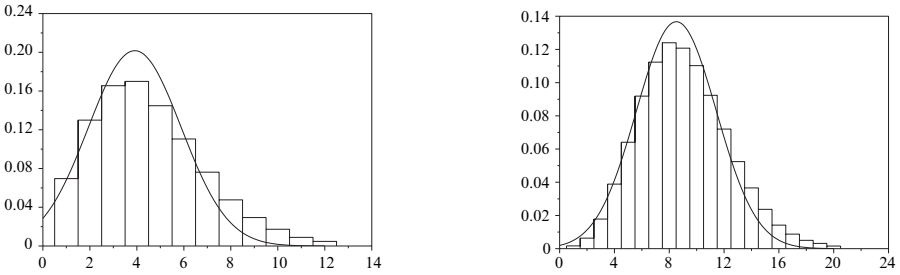

Fig. 7.5. Histogramme de la loi de S_n pour $\theta = 1$, $n = 50$ (à gauche) et $n = 5\,000$ (à droite), obtenu à l'aide de 100 000 simulations de S_n, et densité de la loi gaussienne de moyenne θa_n et de variance $\theta a_n + \theta^2 b_n$

En particulier, si z est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$, alors on obtient un intervalle de confiance de niveau asymptotique $1-\alpha$ pour θ :

$$\Delta_n = \left[\frac{S_n}{a_n} \pm \frac{z\sqrt{S_n}}{a_n}\right],$$

c'est-à-dire $\lim_{n\to\infty} \mathbb{P}(\theta \in \Delta_n) = 1-\alpha$.

7.4.3 Conclusion sur l'estimation du taux de mutation

Il n'y a pas une grande différence entre K_n et S_n pour l'estimation de θ. Comme K_n et S_n dont de l'ordre de $\log(n)$, la longueur de l'intervalle de confiance est de l'ordre de $1/\sqrt{\log(n)}$. Pour diviser la longueur de l'intervalle de confiance par 2, il faut observer n^4 individus au lieu de n. Ceci semble peu réalisable. Les estimateurs $K_n/\log(n)$ et S_n/a_n de θ sont très imprécis (voir les histogrammes des loi de K_n et S_n, graphiques 7.4 et 7.5). Malheureusement la vitesse de convergence en $(\log n)^{-1/2}$ est générique pour l'estimation de θ. La vitesse de convergence de l'estimateur du maximum de vraisemblance de θ, voir définition 5.2.2, quand on connaît l'arbre généalogique (forme et longueur des branches i.e. la suite $T_2, \ldots, T_n$) et le nombre de mutations pour chaque branche est aussi en $(\log n)^{-1/2}$, voir [8]. En revanche, on peut améliorer l'estimation de θ, à horizon fini, en tenant compte des arbres phylogéniques compatibles avec les données observées, voir par exemple [17], Chaps. 5 et 6.

Références

1. D.J. Aldous. Deterministic and stochastic models for coalescence (aggregation and coagulation) : a review of the mean-field theory for probabilists. *Bernoulli*, 5(1) : 3–48, 1999.
2. C. Cannings. The latent roots of certain Markov chains arising in genetics : a new approach. I. Haploid models. *Advances in Appl. Probability*, 6 : 260–290, 1974.
3. C. Cannings. The latent roots of certain Markov chains arising in genetics : a new approach. II. Further haploid models. *Advances in Appl. Probability*, 7 : 264–282, 1975.
4. R. Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer-Verlag, New York, 2002.
5. W.J. Ewens. *Mathematical population genetics. I*, volume 27 de *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, seconde édition, 2004. Theoretical introduction.
6. R.A. Fisher. On the dominance ratio. *Proc. Roy. Soc. Edinburgh*, 42 : 321–341, 1922.
7. R.A. Fisher. *The genetical theory of natural selection*. Clarendon Press, Oxford, 1930.

8. Y.X. Fu et W.H. Li. Maximum likelihood estimation of population parameters. *Genetics*, 1993.
9. O. Gascuel. *Mathematics of Evolution and Phylogeny*. Oxford University Press, 2005.
10. J.F.C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3) : 235–248, 1982.
11. M. Möhle. Robustness results for the coalescent. *J. Appl. Probab.*, 35(2) : 438–447, 1998.
12. M. Möhle et S. Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29(4) : 1547–1562, 2001.
13. P.A.P. Moran. Random processes in genetics. *Proc. Cambridge Philos. Soc.*, 54 : 60–71, 1958.
14. J. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27(4) : 1870–1902, 1999.
15. S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, 36(4) : 1116–1125, 1999.
16. J. Schweinsberg. Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Process. Appl.*, 106(1) : 107–139, 2003.
17. S. Tavaré et O. Zeitouni. *Lectures on probability theory and statistics*, volume 1837 de *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. 31$^{\text{ième}}$ école d'été de probabilité à Saint-Flour, 8–25 juillet 2001. Édité par Jean Picard.
18. G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoret. Population Biology*, 7 : 256–276, 1975.
19. S. Wright. Evolution in Mendelian populations. *Genetics*, 16 : 97–159, 1931.