–Texte–
Ancêtres et descendants

1 Un modèle d'arbre généalogique

On souhaite ici présenter un modèle d'évolution génétique permettant d'étudier la structure d'arbres phylogéniques. Ces arbres décrivent l'évolution des espèces à partir d'un ancêtre commun, en précisant les instants de différenciation des espèces. Étant donné un groupe de k individus, on s'intéresse en particulier au temps d'apparition de l'Ancêtre Commun le Plus Récent (ACPR dans la suite) à ce groupe et à la taille de l'arbre généalogique de cet individu.

Présentons les hypothèses de notre modèle :

- on ne s'intéresse qu'à la population de sexe masculin (par exemple) ; chaque individu n'a donc qu'un parent à la génération précédente,
- la taille de la population reste constante au cours du temps, égale à N,
- les générations sont comptées à rebours du temps : la génération 0 est le présent, la génération 1 celle des parents des individus précédents, ...
- les générations ne se chevauchent pas : à chaque instant k, la k-ième génération est composée des N enfants de la génération $k+1$,
- on note a_i^k le parent à la génération $k-1$ de l'individu i à la génération k et on suppose, pour tout $k \in \mathbb{N}$, les variables aléatoires $(a_i^k)_{1\leqslant i\leqslant N}$ indépendantes et de même loi uniforme sur $\{1,\dots,N\}$,
- les reproductions à chaque génération sont indépendantes ; en d'autres termes, on suppose les N-uplets $((a_1^k,\dots,a_N^k))_{k\in\mathbb{N}}$ indépendants.

En résumé, compte tenu des hypothèses ci-dessus, un arbre généalogique est généré de la manière suivante : chaque individu de la génération k choisit un individu (son parent) dans la génération $k+1$ selon la loi uniforme et ce, indépendamment de tous les autres individus (de sa génération ou des autres).

Dans ce processus, chaque individu a un parent mais certains individus n'ont pas de descendants. On note ν_i^k le nombre de descendants à la génération $k+1$ de l'individu i vivant à la génération k. Les variables aléatoires $(\nu_i^k)_{1\leqslant i\leqslant N}$ ne sont pas indépendantes puisqu'elles doivent vérifier la relation $\nu_1^k+\cdots+\nu_N^k=N$.

Lemme 1.1. *Le N-uplet $\nu^k=(\nu_1^k,\dots,\nu_N^k)$ suit la loi multinomiale de paramètres N et $(1/N,\dots,1/N)$ que nous noterons μ_N, c'est-à-dire que*

$$\mathbb{P}\big(\nu_1^k=m_1,\dots,\nu_N^k=m_N\big)=\frac{N!}{m_1!\cdots m_N!}\frac{1}{N^N}\mathbf{1}_{\{m_1+\cdots+m_N=N\}}. \tag{1}$$

De plus, les variables aléatoires $(\nu^k)_{k\geqslant 1}$ sont indépendantes.

Remarque 1.2. Soit $Y_1, \dots, Y_N$ des variables aléatoires indépendantes et identiquement distribuées de loi de Poisson de paramètre θ. Alors, la loi de la variable aléatoire $(Y_1, \dots, Y_N)$, conditionnellement à l'événement $Y_1 + \cdots + Y_N = N$ suit la loi multinomiale μ_N. En particulier, la loi de ν_i^k est la loi binomiale $\mathcal{B}(N, 1/N)$.

2 Le processus ancestral à temps discret

Dans la suite, on ne s'intéresse pas à toute la structure de l'arbre mais uniquement au nombre d'ancêtres à chaque génération d'un groupe d'individus fixé au départ (inférieur ou égal à N). On appelle processus ancestral la suite $(A^N(r))_{r\in\mathbb{N}}$ telle que $A^N(r)$ soit le nombre d'ancêtres distincts à la génération r.

Proposition 2.1. *La suite $A^N(\cdot)$ est une chaîne de Markov à espace d'états $\{1, \dots, N\}$ dont la matrice de transition $G^{(N)}$ est diagonale inférieure. De plus, pour tous $i \geqslant j$, $G_{ij}^{(N)} > 0$. L'état 1 est absorbant les états $2, \dots, n$ sont transitoires.*

D'après les hypothèses du modèle, la probabilité que deux individus donnés aient deux parents distincts (à la génération précédente) est égale à $1 - 1/N$. On en déduit que $G_{22}^{(N)} = 1 - 1/N = 1 - G_{21}^{(N)}$. De même, on a

$$G_{33}^{(N)} = 1 - \frac{3N-2}{N^2}, \quad G_{32}^{(N)} = \frac{3}{N} - \frac{3}{N^2}, \quad G_{31}^{(N)} = \frac{1}{N^2}.$$

Les coefficients suivants de $G^{(N)}$ sont de plus en plus difficiles à expliciter (et on ne cherchera pas à le faire) mais le résultat suivant permet de simuler des trajectoires de A_n^N, $n \geqslant 0$, sans avoir à expliciter la matrice de transition.

Lemme 2.2. *La loi de $A^N(r+1)$ sachant que $A^N(r) = k$ est le nombre de coordonnées positives dans un vecteur de loi multinomiale de paramètres k et $(1/N, \dots, 1/N)$.*

Dater l'ancêtre commun (le plus récent) à n individus revient à s'intéresser au temps d'atteinte W^N de 1 par la chaîne A^N issue de n et notamment à son espérance et sa variance. Notons $e_n^N = \mathbb{E}_n[W^N]$. La propriété de Markov assure que

$$e_n^N = \frac{1}{1 - G_{nn}^{(N)}} \left(1 + \sum_{k=1}^{n-1} G_{nk}^{(N)} e_k^N \right).$$

Puisque les coefficients de $G^{(N)}$ ne sont pas explicites, ce système d'équations n'est d'aucune utilité. On peut bien entendu utiliser une méthode de Monte-Carlo pour estimer l'espérance et la variance de W^N sachant que $A^N(0) = n \dots$

L'autre question intéressante est de comprendre les propriétés de la longueur totale L_n^N de l'arbre généalogique reliant n individus c'est-à-dire le nombre total de descendants

de l'ancêtre commun. Cette variable aléatoire est définie par

$$L_n^N = \sum_{r=0}^{W^N} A^N(r),$$

avec $A^N(0) = n$. Encore une fois, il est illusoire d'espérer des résultats explicites sur la loi de cette variable aléatoire dès que n est un peu grand.

De plus, le temps d'apparition d'un ancêtre commun à au moins deux individus parmi n est une variable aléatoire de loi géométrique dont le paramètre est de l'ordre de $1/N$. Dans une grande population, ce temps sera donc très long.

3 Le modèle à temps continu

L'idée est ici de prendre du recul. Pour cela, la taille de la population sera choisie très grande et on ne va observer la situation qu'à des instants de plus en plus éloignés (fonctions de la taille de la population).

Remarque 3.1. Remarquons tout d'abord que si $[Nt]$ désigne la partie entière de Nt, alors la probabilité que deux individus donnés n'ait pas d'ancêtre commun avant la génération $r = [Nt]$ est donnée par

$$\left(1 - \frac{1}{N}\right)^{[Nt]} \xrightarrow[N\to+\infty]{} e^{-t}.$$

En d'autres termes, lorsque N est grand et que le temps est mesuré en unité de N générations, le temps d'apparition d'un ancêtre commun à deux individus donnés suit une loi exponentielle de paramètre 1, notée $\mathcal{E}(1)$.

3.1 Le passage à la limite

L'idée est donc d'étudier le comportement du processus ancestral sous cette renormalisation. Le résultat suivant (que l'on admettra) montre que cette approche a un sens.

Proposition 3.2. *Il existe un processus $(A(t))_{t\geqslant 0}$ à valeurs dans $\mathbb{N}^*$ tel que pour tous $1 \leqslant k, j \leqslant n$,*

$$\mathbb{P}_k(A(t) = j) = \lim_{N\to\infty} \mathbb{P}_k(A^N([Nt]) = j).$$

Ce processus peut être construit de la manière suivante : si $A(0) = n$ alors

$$A(t) = n - \sum_{k=-1}^{n-2} \{t\geqslant T_n+\cdots+T_{n-k}\},$$

où $T_2, T_3, \ldots, T_n$ sont des variables aléatoires indépendantes de lois exponentielles de paramètres respectifs $\binom{2}{2}$, $\binom{3}{2}$,$\ldots$,$\binom{n}{2}$ et $T_{n+1} = 0$ par convention.

Remarque 3.3. On peut reformuler le résultat ci-dessus de la façon suivante : si $A(0) = n$, alors pour tout $k \leqslant n$, $A(t) = k \iff T_n + \cdots + T_{k+1} \leqslant t < T_n + \cdots + T_k$, avec l'abus de notation $T_{n+1} = 0$. En d'autres termes, A reste en k durant un temps exponentiel de paramètre $\binom{k}{2}$ puis « saute » en $k-1$ *etc* (les temps de séjour étant indépendants). On peut aussi le voir comme un processus de vie et de mort avec $\lambda_k = 0$ (on parle de processus de mort pur), $\mu_k = \binom{k}{2}$ pour $k \geqslant 2$ et $\mu_1 = 0$.

3.2 Temps d'apparition de l'ACPR

Soit W_n le temps d'apparition de l'ACPR sachant que $A(0) = n$. On a

$$W_n = T_n + \cdots + T_2,$$

où les variables aléatoires $(T_k)_{2 \leqslant k \leqslant n}$ sont les temps de séjour dans les états $2, \ldots, n$. En particulier, on a donc

$$\mathbb{E}[W_n] = 2\left(1 - \frac{1}{n}\right) \quad \text{et} \quad 1 = \mathbb{V}(W_2) \leqslant \mathbb{V}(W_n) \leqslant \lim_{n\to\infty} \mathbb{V}(W_n) = 8\frac{\pi^2}{6} - 12 \approx 1.16.$$

Ces calculs suggèrent que la contribution essentielle semble provenir de T_2. De plus, on peut se demander quelles sont les propriétés de la limite de la suite $(W_n)_{n \geqslant 2}$.

Proposition 3.4. *La suite $(W_n)_{n\geqslant 2}$ converge presque sûrement vers une variable aléatoire W_∞ dont la transformée de Laplace est définie sur $]-\infty, 1/2[$. En particulier, elle admet des moments de tous ordres et*

$$\mathbb{E}[W_\infty] = 2 \quad \textit{et} \quad \mathbb{V}(W_\infty) = \frac{4\pi^2}{3} - 12.$$

Démonstration. Pour tout $n \geqslant 2$, la transformée de Laplace de W_n vaut

$$\mathbb{E}[\exp(\lambda W_n)] = \prod_{k=2}^{n} \mathbb{E}[\exp(\lambda T_k)] = \prod_{k=2}^{n} \left(1 + \frac{\lambda}{\binom{k}{2} - \lambda}\right).$$

Elle est donc finie sur $]-\infty, 1/2[$. Il reste à étudier la convergence de la suite de ces transformées de Laplace. □

3.3 Longueur de l'arbre

Par analogie avec le modèle à temps discret, on note L_n, et l'on appelle *longueur de l'arbre généalogique* la variable aléatoire égale à la somme des temps de vie de tous les individus de l'arbre. Elle s'exprime en fonction des temps d'apparition des ancêtres communs :

$$L_n = 2T_2 + \cdots + nT_n.$$

En particulier,

$$\mathbb{E}[L_n] \approx 2\log n \quad \text{et} \quad \mathbb{V}(L_n) \sim \frac{2\pi^2}{3}.$$

On peut en fait déterminer complètement la loi de L_n.

Proposition 3.5. *La variable aléatoire L_n suit la loi du maximum de $n-1$ variables aléatoires indépendantes et de même loi exponentielle de paramètre 1/2.*

Démonstration. Soit $(Z_n)_{n\geqslant 1}$ une suite de variables aléatoires indépendantes et de même loi exponentielle $\mathcal{E}(1/2)$. Pour tout $n \geqslant 1$, on note $Y_n = \max(Z_i, 1 \leqslant i \leqslant n)$. La transformée de Laplace de Y_n est donnée, pour $\lambda < 1/2$, par

$$G_n(\lambda) = \mathbb{E}(\exp(\lambda Y_n)) = \int_0^{+\infty} e^{\lambda y} \frac{n}{2} e^{-y/2} \big(1 - e^{-y/2}\big)^{n-1}\, dy.$$

On remarque alors que, pour tout $n \geqslant 2$, et tout $\lambda < 1/2$,

$$G_n(\lambda) = \frac{n}{1-2\lambda} G_{n-1}(\lambda - 1/2).$$

La transformée de Laplace de L_n notée H_n vaut quant à elle

$$H_n(\lambda) = \mathbb{E}\big[\exp(\lambda L_n)\big] = \prod_{k=2}^{n} \frac{k-1}{k-1-2\lambda}.$$

Une récurrence permet de déduire de ce qui précède que G_{n-1} et H_n sont égales. □

On veut à présent déterminer le comportement asymptotique de la suite $(L_n)_{n\geqslant 2}$.

Proposition 3.6. *La suite $(L_n/\log n)_{n\geqslant 2}$ converge presque sûrement vers 2 et $(L_n - 2\log n)_{n\geqslant 2}$ converge en loi vers la mesure de probabilité de fonction de répartition*

$$F(t) = \exp(-\exp(-t/2)).$$

Démonstration. On utilise les mêmes notations que dans la preuve de la Proposition 3.5. On a alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}(Y_n/\log n \leqslant 2(1-\varepsilon)) = \exp(-n^{\varepsilon}(1+o(1))),$$

ce qui assure que $\liminf(Y_n/\log n) \geqslant 2$ p.s. D'autre part,

$$\mathbb{P}(Y_n/\log n \geqslant 2(1+\varepsilon)) = n^{-\varepsilon(1+o(1))}.$$

Soit $\delta > 1/\varepsilon$. Pour tout $k \geqslant 1$, on pose $n_k = [(k+1)^{\delta}]$, où $[x]$ désigne la partie entière de x. La borne ci-dessus assure que $\limsup(Y_{n_k}/\log n_k) \leqslant 2$. En conséquence, $(Y_{n_k}/\log n_k)_{k\geqslant 1}$ converge vers 2 p.s. On conclut en utilisant l'encadrement suivant :

$$\frac{\log n_k}{\log n_{k+1}} \frac{Y_{n_k}}{\log n_k} \leqslant \frac{Y_n}{\log n} \leqslant \frac{Y_{n_{k+1}}}{\log n_{k+1}} \frac{\log n_{k+1}}{\log n_k},$$

où k est choisi tel que $n_k \leqslant n < n_{k+1}$. La convergence en loi se déduit de la Proposition 3.5. □

Remarque 3.7. On peut aussi obtenir la convergence presque sûre de $(L_n/\log n)_{n\geqslant 2}$ comme conséquence de résultat suivant (qui pourra être admis sans démonstration). Soit $(b_n)_{n\geqslant 1}$ une suite croissante qui tend vers $+\infty$. Soit $(V_n)_{n\geqslant 1}$ une suite de variables aléatoires indépendantes de carré intégrable telles que

$$\frac{1}{b_n}\sum_{k=1}^{n}\mathbb{E}(V_k) \xrightarrow[n\to\infty]{} m \quad \text{et} \quad \sum_{k=1}^{n}\frac{\mathbb{V}(V_k)}{b_k^2} < +\infty.$$

Alors

$$\frac{1}{b_n}\sum_{k=1}^{n}V_k \xrightarrow[n\to+\infty]{p.s.} m.$$

4 Suggestions

1. On pourra commenter le modèle, démontrer le Lemme 1.1 et/ou la Remarque 1.2.
2. On pourra démontrer la Proposition 2.1 et le Lemme 2.2 et en déduire un procédé pour simuler des trajectoires du processus à temps discret.
3. On pourra essayer de proposer une intuition pour le passage à la limite (arguments sur les temps de saut et les sauts pour deux ou trois individus ou simulations).
4. On pourra démontrer et illustrer par la simulation une partie des résultats concernant le temps d'apparition de l'ACPR.
5. On pourra démontrer la Proposition 3.5.
6. On pourra démontrer la Proposition 3.6.

