Génétique des populations et inférence ancestrale sur des processus de branchement en temps continu

Benoît Henry

Faculté des Sciences et Technologies
Université de Lorraine

Génétique des populations et inférence ancestrale sur des processus de branchement en temps continu

Rapport de Stage

présenté pour l'obtention du diplôme de

Master

Spécialité : Mathématiques fondamentales et appliquées, MFA

par

Benoit Henry,
sous la direction de Nicolas Champagnat.

Soutenu le 18 septembre 2013.

Institut Elie Cartan de Lorraine, Laboratoire de Mathématiques, BP 239, 54 506 Vandœuvre-lès-Nancy Cedex

Table des matières

Remerciements

Je voudrais en premier lieu remercier de tout coeur Nicolas Champagnat, qui m'a dirigé lors de ce stage, pour ses conseils, sa patience, pour avoir su poser des questions stimulantes et pour ses nombreuses idées et critiques.

Je voudrais également remercier les autres personnes du laboratoire de mathématiques de Nancy pour leur accueil, leur soutient et leurs conseils. Je ne prendrai pas le risque de citer des noms puisqu'il me serais impossible de ne pas en oublier mais je ne doute pas qu'ils se reconnaitront.

Finalement, je souhaite remercier mes parents pour tout ce qu'ils ont pu faire pour moi.

1 Introduction

Avec la découverte de l'ADN, la biologie s'est retrouvée confrontée à de nouvelles problématiques dans l'étude du vivant qui lui ont permis d'étudier plus précisément les mécanismes en jeu dans l'évolution des espèces.

Notamment, les biologistes sont confrontés régulièrement au problème de reconstruire la généalogie d'une population pour en tirer des informations sur ses caractéristiques présentes en se basant sur les séquences d'ADN d'un échantillon. Cependant, de par la nature aléatoire de la reproduction ou de l'apparition de mutations, le génotype d'une population (l'ensemble des différents gènes et allèles la constituant) ne détermine pas de manière certaine sa généalogie. Des questions se posent alors :

- Etant donné le génotype d'un échantillon d'une population, quelle est la généalogie la plus probable dont il est issu ?
- A quelle fréquence apparaissent les mutations ?
- Quelle est la répartition des allèles ou des génotypes dans la population ?

Ces questions n'ont naturellement pas de réponse claire, et sont dépendantes du modèle utilisé pour décrire la dynamique de la population. Le présent mémoire sera divisé en deux parties : la première se consacrera à la présentation du modèle de Wright-Fisher, et à son pendant en temps rétrograde : le modèle de coalescence de Kingman. La seconde partie sera consacrée à l'étude d'un modèle de génétique des populations plus récent : le modèle des splitting trees, et son pendant en temps rétrograde : le processus ponctuel de coalescence. C'est cette seconde partie qui contient les travaux originaux de ce mémoire. Les deux parties suivront un cheminement similaire :

- Etude du modèle dans le sens normal du temps.
- Reconstruction de la généalogie en temps rétrograde.
- Introduction et étude de différents modèles de mutations sur les modèles en temps rétrograde.

2 Modèles de base en génétique des populations

Dans un premier temps, nous allons présenter une famille de modèles associés à des populations d'effectifs constants. Un cas particulier important sera le modèle de Wright-Fisher, lié au coalescent de Kingman, que nous présenterons comme un cas particulier du modèle de Cannings.

Il existe deux grandes familles de modèles : les modèles suivant le sens normal du temps décrivant la dynamique des naissances et des morts dans la population et les modèles en temps rétrograde qui servent en général à décrire les liens généalogiques passés reliant les individus d'une population. Le modèle de Cannings est un modèle en temps normal.

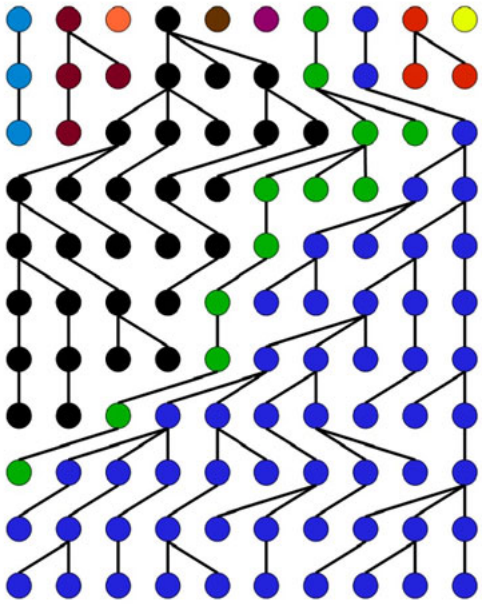


FIGURE 1 – Une dynamique de Cannings

Cette partie rassemble divers résultats classiques de la génétique des populations. Nous nous sommes basés pour notre présentation sur [11] et [13].

2.1 Modèle de Cannings

Le modèle de Cannings est très général et recouvre de nombreuses dynamiques (raisonnables) de populations monotypes et d'effectifs constants.

D'un point de vue modélisation, ses grandes faiblesses sont justement la constance de la taille de la population et le fait que les différentes générations soient disjointes. Voici les partis pris de modélisation supposés par le modèle de Cannings :

- On travaille avec une population de taille fixée $N \in \mathbb{N}$.
- Chaque génération est constituée des enfants de la génération précédente.
- A chaque génération, les nombres d'enfants d'un groupe d'individus de taille donné sont distribués suivant une loi de probabilité donnée, indépendante du groupe choisi.

Plus précisément :

Propriétés de la distribution des enfants

Soit $\mu = (\mu_1, ..., \mu_N)$ une variable aléatoire à valeur dans $\mathbb{N}^N$. Pour chaque individu i d'une génération donnée, μ_i représente le nombre de ses enfants dans la nouvelle génération. Pour que le modèle soit cohérent, il est nécessaire que la loi de μ satisfasse quelques propriétés :

- $\sum_{i=1}^{N} \mu_i = N$ p.s.
  (Pour conserver une population de taille fixe)

– La loi de μ est échangeable, i.e $\forall \sigma \in \mathcal{S}_N, \quad (\mu_1.., \mu_N) \overset{\mathcal{L}}{=} (\mu_{\sigma(1)}, ..., \mu_{\sigma(N)})$.
(Aucun individu n'est privilégié.)
– $\text{Var}(\mu_1) > 0$.
(Evite le cas trivial)

Remarque 2.1. *On a* $\mathbb{E} \sum \mu_i = N = N\ \mathbb{E}\mu_1$ *par échangeabilité. D'où* $\mathbb{E}\mu_i = 1$.

Dynamique de la population

La dynamique de la population est caractérisée seulement par une suite $(\mu_r)_{r\geq 0}$ de variables aléatoires i.i.d. de même loi que μ. Plus précisément, après avoir fixé une convention de numérotation des individus d'une génération à l'autre, la seule connaissance de la suite $(\mu_r)_{r\geq 0}$ permet de retrouver les relations de filiation entre individus, comme représentées dans la figure 1, avec la convention suivante : on numérote déjà les enfants de l'individu 1 de 1 à μ_1^r, puis ceux de l'individu 2, etc... pour tout $r \geq 0$, μ_i^r *représente le nombre d'individus descendants de l'individu* i *à la génération* r.

Dynamique d'un gène à deux allèles

Nous allons maintenant étudier le cas d'une population typée. Le cas le plus simple est l'étude d'un seul gène n'ayant que deux allèles différents. Supposons désormais que chaque individu de la population possède un type a ou A, de manière à ce que chaque individu transmet son type à ses descendants. Nous supposerons également que les types sont *neutres* : c'est-à-dire que le type n'influence pas la reproduction du porteur.

On note dès lors Y_n le nombre d'individus de type A dans la génération n. Alors le processus $(Y_n)_{n\geq 0}$ est une *chaîne de Markov*, telle que :

$$\mathbb{P}(Y_{n+1} = i | Y_n = j) = \mathbb{P}\left(\sum_{l=1}^{j} \mu_l = i\right) \ \forall i, j \in [\![0, N]\!] \ \text{(par écheangabilité)}.$$

D'autre part, par la remarque 2.1

$$\mathbb{E}[Y_{n+1} | X_n] = \sum_{l=1}^{Y_n} \mathbb{E}[\mu_l] = Y_n.$$

Le processus est donc également une *martingale*.

Probabilité de fixation de l'allèle observé

Il est clair, à la vue de ce qui précède, que les états 0 et N sont absorbants pour la chaîne Y. De plus, étant donné que Y est une martingale bornée (par 0 et N), celle-ci converge presque sûrement, et on voit facilement que la limite est l'un des deux points absorbants de la chaîne. Ainsi :

$$Y_\infty := \lim_{n\to\infty} Y_n \in \{0, N\}$$

On appelle fixation de l'allèle A (resp. a) l'atteinte de N (reps. 0) par la chaîne Y. Nous allons maintenant essayer de déterminer la probabilité de fixation de l'allèle A dans la population. Pour cela on pose :

$$\tau = \inf\{n \geq 0 \mid Y_n = N \ ou \ Y_n = 0\} < \infty \ p.s.$$

Par le théorème d'arrêt

$$\mathbb{E}_i[Y_\tau] = N\ \mathbb{P}_i(Y_\infty = N) = \mathbb{E}_i[Y_0] = i\ .$$

Finalement

$$\mathbb{P}_i(Y_\infty = N) = \frac{i}{N}\ .$$

Remarque 2.2. *Nous serons amenés dans la suite, pour un vecteur aléatoire $(Z_1, \ldots, Z_n)$, à utiliser la notation $\mathbb{P}(Z_{1:n} = z_{1:n})$ pour $\mathbb{P}(Z_1 = z_1, \ldots, z_n = z_n)$.*

2.2 Cas particulier : le modèle de Wright-Fisher neutre

Supposons désormais que μ suit une loi multinomiale sur $\mathbb{N}^N$ de paramètres $\left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$. On vérifie facilement que cette loi satisfait les conditions du modèle de Cannings.
D'autre part, on peut alors déterminer explicitement la probabilité de transition de la chaîne de Markov $(Y_n)_{n\geq 0}$:

$$\begin{aligned}
\mathbb{P}\left(\textstyle\sum_{l=1}^{j} \mu_l = k\right) &= \sum_{l_1+..+l_j=k} \mathbb{P}(Y_{1:j} = l_{1:j}) = \sum_{l_1+..+l_j=k} \ \sum_{l_{j+1}+..+l_N=N-k} \mathbb{P}(Y_{1:N} = l_{1:N}) \\
&= \sum_{l_1+..+l_j=k} \frac{N!}{(N-k)!l_1!...l_j!} \sum_{l_{j+1}+..+l_N=N-k} \frac{(N-k)!}{l_{j+1}!...l_N!} \\
&= \binom{N}{k} \left(\frac{j}{N}\right)^k \left(1 - \frac{j}{N}\right)^{N-k} \qquad \text{(par la formule du multinôme).}
\end{aligned}$$

On remarque alors que $\mathcal{L}(Y_n|Y_{n-1}) = \mathcal{B}\left(N, \frac{Y_{n-1}}{N}\right)$, où la notation $\mathcal{B}(n,p)$ désigne la loi binomiale de paramètres n et p.

Remarque 2.3. *Il y a une interprétation classique de la loi multinomiale : étant données N* boîtes *et N* balles*, la loi multinomiale correspond à la distribution du nombre de balles dans chacune des boîtes lorsque les balles sont déposées de manière indépendante et équiprobable dans chacune des N boîtes. Du point de vue de la dynamique des populations, on interprète ceci en disant :*

Dans le modèle de Wright-Fisher, chaque individu de la nouvelle génération choisit son parent uniformément dans la population de la génération précédente, et indépendamment des autres individus.

Il est possible de regarder quelques propriétés de la chaine Y.

Variance de Y_n

$$\mathrm{Var}(Y_n|Y_{n-1}) = N\ \frac{Y_{n-1}}{N}\left(1 - \frac{Y_{n-1}}{N}\right)$$

et

$$\mathrm{Var}\left(Y_n\right) = \mathbb{E}Y_0\left(N - \mathbb{E}Y_0\right)\left(1 - \left(1 - \frac{1}{N}\right)^n\right) + \left(1 - \frac{1}{N}\right)^n \mathbb{E}Y_0 \ .$$

2.2.1 Etude de l'évolution d'un allèle sous le modèle de Wright-Fisher

Le modèle de Wright-Fisher nous permet d'étudier plus en profondeur la dynamique d'une population porteuse d'un gène à deux allèles. Nous considérons ici encore que la population est scindée en deux types A et a. Nous allons étudier le comportement de la proportion d'un des allèles dans la population.

Hétérozygotie

Le notion mathématique d'hétérozygotie a pour but de *quantifier la variabilité génétique* dans la population. Elle est définie, ici, comme la *probabilité que deux individus tirés au hasard (avec remise) soient de type diffèrent.* Par exemple, si tous les individus de la population sont de type A, alors l'*hétérozygotie* est nulle. Dans le cas du modèle de Wright-Fisher, en notant h_n l'hétérozygotie de la population à la n-ième génération, on a :

$$\mathbb{E}\left[h_n \mid Y_n\right] = \frac{Y_n\left(N - Y_n\right)}{N^2}$$

On montre alors par récurrence que l'hétérozygotie moyenne h_n est donnée par

$$h_n = \left(1 - \frac{1}{N}\right)^n \ \mathbb{E}Y_0\left(N - Y_0\right) \ .$$

On remarque que $h_n \underset{n\to\infty}{\longrightarrow} 0$ *exponentiellement.* Cela signifie que la diversité va disparaître au sein de la population p.s., mais que peut-on dire du temps de fixation ? Il est compliqué de l'étudier, c'est pourquoi nous allons devoir l'approcher par un processus de diffusion.

Convergence en loi de la proportion d'un allèle

L'objectif de cette section est de montrer que la proportion de porteur de allèle A, modulo un changement de l'échelle de temps, converge en loi vers un processus de Markov. Pour cela introduisons déjà le théorème bien connu suivant, dont une preuve pourra trouver dans [12] :

Théorème 2.4 (Problème de martingale)**.** *Soit* $\left(\Omega, \mathcal{F}, \left(\mathcal{F}_t\right)_{t\geq 0}, \mathbb{P}\right)$ *un espace probabilisé filtré. Soit* b *et* σ *deux fonctions de* $\mathbb{R}$ *dans* $\mathbb{R}$ *vérifiant les conditions suivantes :*

- b *est Lipschitzienne.*
- *Il existe une fonction* h *telle que* $h\left(0\right) = 0$*,* h *strictement croissante et* $\int_0^\epsilon ds\ h^{-2}\left(s\right) = \infty \quad \forall \epsilon > 0$*, telle que* $\forall x, y \quad \mid \sigma\left(x\right) - \sigma\left(y\right) \mid \leq h\left(\left|x - y\right|\right)$*.*

Sous ces conditions, si $\left(X_t,\ t \in \mathbb{R}_+\right)$ *un processus, tel que* $\forall f \in \mathcal{C}_c^2$*, le processus* M^f *défini par :*

$$M_t^f = f\left(X_t\right) - f\left(X_0\right) - \int_0^t f''\left(X_s\right) b\left(X_s\right) ds$$

est une martingale. Alors X *a même loi que toute solution de l'EDS :*

$$dY_t = \sigma\left(Y_s\right) dB_t + b\left(Y_s\right) dt$$

Posons désormais pour tout $N \in \mathbb{N}$:

$$\forall t \in \mathbb{R}_+, \quad X_t^N = \frac{1}{N} Y_{[Nt]}$$

Théorème 2.5 (Convergence de la proportion d'allèle)**.**
Sous la condition que la suite de variables aléatoires $\left(X_0^N\right)_{N\geq 0}$ *converge en loi, alors :*

$$X^N \xrightarrow[N\to\infty]{\mathcal{L}} X \quad \textit{dans } \mathbb{D}(\mathbb{R}_+, [0,1])$$

pour la topologie de Skorohod, avec X *solution de l'EDS :*

$$dX_t = \sqrt{X_t(1-X_t)}dB_t, \quad \forall t \geq 0$$

Démonstration. La preuve est classique, elle se fait en deux étapes :
- Vérifier que la famille des lois des processus est uniformément tendue (c'est-à-dire relativement compacte pour la topologie de la convergence étroite dans $\mathcal{M}_1(\mathbb{D})$).
- Identifier une unique valeur d'adhérence : on utilisera pour cela le problème de Martingale.

Tension uniforme

Dans un premier temps, nous allons chercher à obtenir l'estimation suivante pour une constante positive C fixé :

$$\mathbb{E}\left[\left|X_t^N - X_s^N\right|^4\right] \leq C\left(t - s + \frac{t-s}{N}\right)^2$$

Soit $t, s \geq 0$ et $j = [Nt]$ $i = [Ns]$. On étudie

$$\begin{aligned}
\mathbb{E}\left[\left|X_t^N - X_s^N\right|^4\right] =& \frac{1}{N^4}\,\mathbb{E}\left[\left(\sum_{k=i}^{j-1}\left(Y_{k+1}^N - Y_k^N\right)\right)^4\right] \\
=& \mathbb{E}\left[\left(\sum_{k_1,..,k_4=i}^{j-1}\prod_{l=1}^{4}\left(Y_{k_l+1}^N - Y_{k_l}^N\right)\right)\right] \\
=& \frac{1}{N^4}\left(\mathbb{E}\left[\sum_{k=i}^{j-1}\left(Y_{k+1}^N - Y_k^N\right)^4\right] + 2\mathbb{E}\left[\sum_{i\leq k_1<k_2\leq j-1}\left(Y_{k_1+1}^N - Y_{k_1}^N\right)^2\left(Y_{k_2+1}^N - Y_{k_2}^N\right)^2\right]\right. \\
&+ \mathbb{E}\left[\sum_{i\leq k_1<k_2\leq j-1}\left(Y_{k_1+1}^N - Y_{k_1}^N\right)\left(Y_{k_2+1}^N - Y_{k_2}^N\right)^3\right] \\
&\left.+ 2\mathbb{E}\left[\sum_{i\leq k_1<k_2<k_3\leq j-1}\left(Y_{k_1+1}^N - Y_{k_1}^N\right)\left(Y_{k_2+1}^N - Y_{k_2}^N\right)\left(Y_{k_3+1}^N - Y_{k_3}^N\right)^2\right]\right).
\end{aligned}$$

L'idée est alors d'étudier chaque terme de cette dernière expression en exploitant le fait que $\mathcal{L}\left(Y_{n+1}|Y_n\right) = \mathcal{B}\left(N, \frac{Y_{n-1}}{N}\right)$.

Premier terme :

$$\mathbb{E}\left[\left(Y_{k+1}^N - Y_k^N\right)^4\right] = \mathbb{E}\left[\ \mathbb{E}\left[\left(Y_{k+1}^N - Y_k^N\right)^4 \mid Y_k^N\right]\right].$$

Soit $(Z_i)_{i\geq 0}$ une suite i.i.d. de Bernouilli de paramètre p, alors

$$\begin{aligned}\mathbb{E}\left[\left(Y_{k+1}^N - Y_k^N\right)^4 \mid Y_k^N = N\right] &= \mathbb{E}\left[\left(\sum_{i=1}^N (Z_i - p)\right)^4\right] \\ &= \sum_{i=1}^N (Z_i - p)^4 + 2\sum_{\substack{i,j\geq 1\\ i\neq j}} \mathbb{E}\,(Z_i - p)^2 (Z_j - p)^2 \\ &\leq N + 2N(N-1) \leq 3N^2\ .\end{aligned}$$

Ceci donne

$$\mathbb{E}\left[\left(Y_{k+1}^N - Y_k^N\right)^4\right] \leq 3N^2 \leq 3\left(\frac{j-i}{N}\right)^2 \leq 3\left(t - s + \frac{1}{N}\right)^2\ .$$

Alors

$$\frac{1}{N^4}\mathbb{E}\left[\sum_{k=i}^{j-1}\left(Y_{k+1}^N - Y_k^N\right)^4\right] \leq 3\frac{(j-i)}{N^2}.$$

Second terme : On remarque déjà que

$$\mathbb{E}\left[\left(Y_{k+1}^N - Y_k^N\right)^2 \mid Y_k^N\right] = \mathrm{Var}\left(Y_{k+1}^N \mid Y_k^N\right) = N\frac{Y_k^N}{N}\left(1 - \frac{Y_k^N}{N}\right) \leq N\ .$$

Alors

$$\begin{aligned}&\sum_{i\leq k_1<k_2\leq j-1} \mathbb{E}\left[\left(Y_{k_1+1}^N - Y_{k_1}^N\right)^2\left(Y_{k_2+1}^N - Y_{k_2}^N\right)^2 \mid \sigma\left(Y_l^N, l\leq k_2\right)\right] \\ = &\sum_{i\leq k_1<k_2\leq j-1}\left(Y_{k_1+1}^N - Y_{k_1}^N\right)^2 \mathbb{E}\left[\left(Y_{k_2+1}^N - Y_{k_2}^N\right)^2 \mid Y_{k_2}^N\right] \leq N\sum_{i\leq k_1<k_2\leq j-1}\left(Y_{k_1+1}^N - Y_{k_1}^N\right)^2\ .\end{aligned}$$

Finalement

$$\sum_{i\leq k_1<k_2\leq j-1} \mathbb{E}\left[\left(Y_{k_1+1}^N - Y_{k_1}^N\right)^2\left(Y_{k_2+1}^N - Y_{k_2}^N\right)^2\right] \leq \frac{N^2}{2}(i-j)^2\ .$$

Troisième terme :

$$\mathbb{E}\left[\left(Y_{k+1}^N - Y_k^N\right)^3 \mid Y_k^N\right] = Y_k^N\left(1 - \frac{Y_k^N}{N}\right)\left(1 - 2\frac{Y_k^N}{N}\right)\ .$$

Donc

$$\mathbb{E}\left[\left(Y_{k_1+1}^N - Y_{k_1}^N\right)^3\left(Y_{k_2+1}^N - Y_{k_2}^N\right)\right] \leq N^2\ ,$$

ce qui mène au résultat voulu.

Quatrième terme :

on a

$$\mathbb{E}\left[\sum_{i\le k_1<k_2<k_3\le j-1}\left(Y^N_{k_1+1}-Y^N_{k_1}\right)\left(Y^N_{k_2+1}-Y^N_{k_2}\right)\left(Y^N_{k_3+1}-Y^N_{k_3}\right)^2\right]\le \mathbb{E}\left[\sum_{i<k\le j-1}\left(Y^N_k-Y^N_i\right)^2\left(Y^N_{k+1}-Y^N_k\right)^2\right]$$
$$\le N\sum_{i<k\le j-1}\mathbb{E}\left[\left(Y^N_k-Y^N_i\right)^2\right]$$
$$= N\sum_{i<k\le j-1}\sum_{i\le l<k}\mathbb{E}\left[\left(Y^N_{l+1}-Y^N_l\right)^2\right]$$
$$\le N^2\sum_{i<k\le j-1}(k-i)$$
$$\le N^4(t-s)^2\ .$$

On en déduit finalement

$$\mathbb{E}\left[\left|X^N_t-X^N_s\right|^4\right]\le C\left((t-s)^2+\frac{(t-s)}{N}\right)\ .$$

Pour $T>0$, en appliquant l'inégalité maximale à la martingale $\left(X^N_s-X^N_T, s\ge T\right)$, on obtient

$$\forall T>0,\ \ \forall\epsilon>0,\ \ \forall\eta>0,\ \ \mathbb{P}\left(\sup_{T\le s\le T+\delta}\left|X^N_s-X^N_T\right|>\epsilon\right)\le\frac{\mathbb{E}\left[\left|X^N_{T+\delta}-X^N_T\right|^4\right]}{\epsilon^4}\le C\frac{\delta^2}{\epsilon^4}+\frac{\delta}{N\epsilon^4}\ .$$

Ainsi

$$\boxed{\forall T,\epsilon,\eta>0,\ \exists\delta\in(0,T),\ \forall t\in[0,T),\ \exists N_0\in\mathbb{N},\ \forall N>N_0,\quad \delta^{-1}\mathbb{P}\left(\sup_{t\le s\le t+\delta}\left|X^N_s-X^N_t\right|>\epsilon\right)<\eta\ .}\quad(\star)$$

Lemme 2.6. $(\star)$ *implique que*

$$\boxed{\forall T,\epsilon,\eta>0,\ \exists\delta\in(0,T),\ \forall t\in[0,T),\ \exists N_0\in\mathbb{N},\ \forall N>N_0,\quad \mathbb{P}\left(\sup_{0\le t\le s<T,\ |s-t|<\delta}\left|X^N_s-X^N_t\right|>\epsilon\right)<\eta\ .}$$

Démonstration. Soient : $T,\epsilon,\eta>0$. Il existe δ tel que

$$\mathbb{P}\left(\sup_{t\le s\le t+\delta}\left|X^N_s-X^N_t\right|>\epsilon\right)<\delta\frac{\eta}{2T}\quad\forall N\ .$$

Or, si $0\le s<t<T$ tel que $|t-s|<\delta$, alors

$$|x_t-x_s|\le|x_{\delta(i+1)}-x_t|+|x_{\delta(i+1)}-x_{\delta i}|+|x_{\delta i}-x_s|\le 3\sup_{i\in[\![0,[T/\delta]]\!]}\ \sup_{\delta i\le s\le\delta(i+1)}|X_s-X_{\delta(i+1)}|\ .$$

D'où

$$\left\{\sup_{0\le s<t<T,\ |t-s|<\delta}|X^N_t-X^N_s|>3\epsilon\right\}\subset\bigcup_{i=0}^{[T/\delta]}\left\{\sup_{\delta i\le s\le\delta(i+1)}|X_s-X_{\delta(i+1)}|>\epsilon\right\}$$

et

$$\mathbb{P}\left(\sup_{0\leq s<t<T,\ |t-s|<\delta}|X_t^N-X_s^N|>3\epsilon\right)\leq\sum_{i=0}^{[T/\delta]}\mathbb{P}\left(\sup_{\delta i\leq s\leq\delta(i+1)}|X_s-X_{\delta(i+1)}|>\epsilon\right)$$
$$\leq(1+[T/\delta])\,\mathbb{P}\left(\sup_{t\leq s\leq t+\delta}\left|X_s^N-X_t^N\right|>\epsilon\right)$$
$$\leq(1+[T/\delta])\,\delta\frac{\eta}{2T}<\eta\ .$$

□

On en déduit la tension uniforme de la famille des lois de X^N par le théorème 15.5 de [2].

Identification des valeurs d'adhérence :

Soit $N_k\to\infty$ une sous-suite telle que la suite des lois de X^{N_k} converge. Dans un souci de simplicité, on notera simplement N au lieu de N_k.
On pose

$$G^Nf(x):=\mathbb{E}\left[f\left(X^N_{t+1/N}\right)-f\left(X_t^N\right)\mid X_t^N=x\right]$$
$$=\frac{1}{N}\mathbb{E}\left[f\left(Y^N_{[Nt]+1}\right)-f\left(Y^N_{[Nt]}\right)\mid Y_{[Nt]}=Nx\right]\ .$$

En utilisant cette propriété conjointement avec la formule de Taylor-Largange à l'ordre 2 en x et le fait que $\mathcal{L}\left(Y^N_{[Nt]+1}\mid Y^N_{[Nt]}\right)=\mathcal{B}\left(N,\frac{Y^N_{[Nt]}}{N}\right)$, on obtient

$$G^Nf(x)=\frac{1}{2N}x\,(1-x)\,f''(x)+r_n(x)\,,$$

avec $r_N(x)=O\left(N^{-3/2}\right)$ uniformément en x. Alors

$$M_t^N:=f\left(X_t^N\right)-f\left(X_0^N\right)-\sum_{i=0}^{[Nt]-1}G^Nf\left(X^N_{i/N}\right)$$
$$=f\left(X_t^N\right)-f\left(X_0^N\right)-\frac{1}{2}\int_0^{[Nt]/N}X_s^N\left(1-X^Ns\right)f''\left(X_s^N\right)ds-\int_0^{[Nt]/N}Nr_N\left(X_s^N\right)ds,$$

est une martingale bornée (voir le théorème 5.12 en annexe). Ce qui est équivalent à la formulation suivante : pour tout ψ continue bornée sur $\mathbb{D}([0,s],[0,1])$ muni de la topologie de Skorohod, à valeur dans $\mathbb{R}$:

$$\forall s<t,\quad\mathbb{E}\left[M_t^N\psi\left(\left(X_r^N\right)_{r\in[0,s]}\right)\right]=\mathbb{E}\left[M_s^N\psi\left(\left(X_r^N\right)_{r\in[0,s]}\right)\right]$$

En passant à la limite, puisque l'on a convergence en loi, on obtient que sous la loi limite, M_t est une martingale. Par unicité de la solution du problème de Martingale du théorème 2.4, on en déduit le résultat. □

Temps et probabilité de fixation pour la diffusion X

Muni de notre diffusion X solution de l'EDS précédente, le but est ici d'étudier le temps de fixation et la probabilité de fixation d'un allèle donné.

Comme X représente la proportion de l'allèle A dans la population, on s'intéresse donc aux quantités : $\mathbb{P}_x(\tau_1 < \infty)$, et $\mathbb{E}_x(\tau)$, avec $\tau = \inf\{t \geq 0 \mid X_t = 1 \text{ ou } X_t = 0\}$ et $\tau_1 = \inf\{t \geq 0 \mid X_t = 1\}$. Puisque X est une martingale bornée, on obtient par le théorème d'arrêt

$$\mathbb{E}_x[X_\tau] = \mathbb{P}_x(\tau_1 < \infty) = \mathbb{E}_x[X_0] = x \ .$$

On obtient ainsi la probabilité de fixation.

On désire maintenant connaître la valeur de l'espérance de τ_1. Par la formule d'Itô, on a

$$\forall f \in \mathcal{C}^2([0,1],\mathbb{R}), \quad f(X_t) = f(X_0) + \int_0^t f'(X_s)\sqrt{X_s(1-X_s)}dB_s + \frac{1}{2}\int_0^t f''(X_s)(X_s(1-X_s))\,ds \ .$$

La propriété de martingale de l'intégrale stochastique contre un brownien permet d'obtenir :

$$\begin{aligned}\forall f \in \mathcal{C}^2([0,1],\mathbb{R}), \quad \mathbb{E}_x\left[\int_0^\tau f'(X_s)\sqrt{X_s(1-X_s)}dB_s\right] &= 0 \\ = \mathbb{E}f(X_t) - \mathbb{E}_x\left[\frac{1}{2}\int_0^\tau f''(X_s)(X_s(1-X_s))\,ds\right] &- x \ .\end{aligned}$$

Afin de tirer τ de cette expression, il faut trouver une fonction f telle que $f''(x) = \dfrac{1}{x(1-x)}$. Si de plus $f(0) = f(1) = 0$, on aurait alors

$$\mathbb{E}_x\left[\int_0^\tau ds\ 1\right] = \mathbb{E}_x(\tau) = 2f(x) - 2\mathbb{E}f(X_\tau) = 0 \ .$$

La solution de ce problème est

$$f(x) = \begin{cases} (1-x)\log\dfrac{1}{1-x} + x\log\dfrac{1}{x} & \text{si } x \in (0,1), \\ 0 & \text{si } x = 0 \text{ ou } 1 \ . \end{cases}$$

Cependant, la fonction f ci-dessus n'est de classe $\mathcal{C}^2$ que sur $(0,1)$. On pose alors

$$\forall n \in \mathbb{N}^* \quad \tau_n = \inf\{t \geq 0 \mid X_t = 1/n \text{ ou } X_t = 1 - 1/n\}$$

En appliquant la formule d'Itô comme ci-dessus mais sur $[0, \tau_n]$ on obtient

$$\mathbb{E}[\tau_n] = 2f(x) - 2\mathbb{E}f(X_{\tau_n}) \ .$$

Enfin par convergence monotone

$$\mathbb{E}_x(\tau) = 2(1-x)\log\frac{1}{1-x} + 2x\log\frac{1}{x} \ .$$

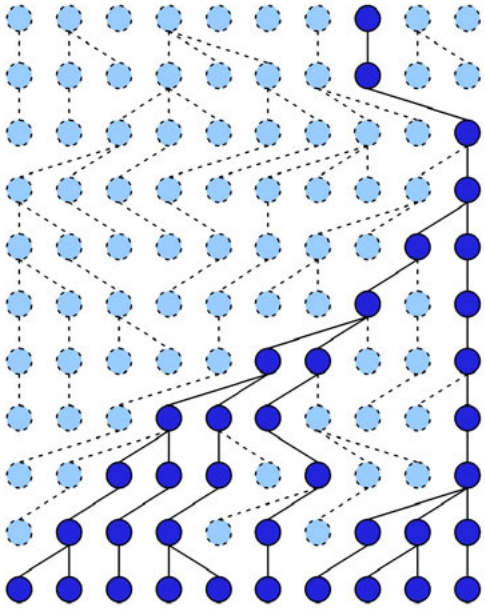


FIGURE 2 – Reconstruction de la généalogie de la popualtion

2.2.2 Reconstruire la généalogie d'un échantillon

Dans cette section, nous allons essayer de reconstruire la généalogie d'un échantillon d'une population suivant le modèle de Wright-Fisher. Cette question se justifie par le fait qu'en pratique, les biologistes ont rarement accès à d'autres informations que la population à l'instant présent, et qu'il peut leur être utile de reconstruire les relations généalogiques qui l'unissent.

Cas d'un échantillon de taille 2.

On note pour tout $k \in \mathbb{N}^*$, $A_N^{(n)}$ la variable aléatoire qui à un échantillon de taille n (tiré de manière uniforme sans remise dans une population de taille N) associe le nombre de générations avant le premier ancêtre commun de l'échantillon.

D'après la remarque 2.2,

$$\mathbb{P}(A_N^{(2)} = 1) = \mathbb{P}(\text{deux v.a.r indépendantes de loi uniforme sur } [\![1, N]\!] \text{ sont égales}) = \frac{1}{N}\,.$$

Puisque les générations sont indépendantes

$$\mathbb{P}(A_N^{(2)} = r) = \mathbb{P}(A_N^{(2)} = 1, \hat{A}_N^{(2)} = r-1) = \frac{1}{N}\left(1 - \frac{1}{N}\right)^{r-1}\,.$$

avec $A_N^{(n)}$ et $\hat{A}_N^{(n)}$ indépendantes et de même loi. Ainsi, A suit une loi *géométrique* de paramètre $\frac{1}{N}$. Alors

$$\mathbb{P}\left(A^{(2)} > Nt\right) = \left(1 - \frac{1}{N}\right)^{Nt} \underset{N\to\infty}{\longrightarrow} e^{-t} .$$

Ceci suggère que le temps avant le premier ancêtre commun converge en loi vers une exponentielle de paramètre 1 lorsque l'on compte le temps en Nt générations.

Cas d'un échantillon de taille n et processus ancestral

Soit $A_N^{(n)}(k)$ le nombre de parents distincts dans un échantillon de n individus de la population (de taille totale N), k générations dans le passé. Il est clair que ce processus est Markovien, puisque l'évolution de la population d'une génération à l'autre ne dépend pas des générations précédentes. De plus :

$$\begin{aligned}\mathbb{P}(A_N^{(n)}(k+1) = i \mid A_N^{(n)}(k) = j) &= \mathbb{P}\left(j \text{ individus ont } i \text{ parents distincts}\right) \\ &= \frac{N\left(N-1\right)\ldots\left(N-i+1\right)}{N^j}\,\mathcal{S}_j^{(i)} ,\end{aligned}$$

où $\mathcal{S}_j^{(i)}$ est le nombre de partitions de $[\![1, j]\!]$ en i sous-ensembles distincts, i.e. le nombre de façons de choisir pour les j individus les ancêtres correspondants.

2.2.3 Convergence du processus ancestral

Nous allons voir que l'on a pour les $A_N^{(n)}$ un résultat du même type que celui que l'on a trouvé pour $A_N^{(2)}$.

Proposition 2.7 (Convergence du processus ancestral)**.**
Soit $\left(A_N^{(n)}(k)\right)_{k\geq 0}$ tel que définit précédemment, alors on a

$$\left(A_N^{(n)}(Nt), t \in \mathbb{R}_+\right) \underset{N\to\infty}{\longrightarrow} A^{(n)} \quad \textit{dans } \mathbb{D}\left(\mathbb{R}_+, [\![0, n]\!]\right),$$

avec $A^{(n)}$ un processus de Markov de mort pure issu de n de $Q-$matrice donnée par

$$Q_{ij} = \begin{cases} \binom{i}{2} & \textit{si } j = i-1 \textit{ et } j \geq 0, \\ \\ 0 & \textit{sinon.} \end{cases}$$

Démonstration.
Soit G^N la matrice de transition de la chaine de Markov A_N. D'une part, on a

$$\begin{aligned} G^N_{k,k-1} =& \frac{N\left(N-1\right)\ldots\left(N-k+2\right)}{N^k}\,\mathcal{S}_{k-1}^{(k)} \\ =& \binom{k}{2}\frac{(N)_{(k-1)}}{N^k} \\ =& \binom{k}{2}\frac{1}{N} + O\left(N^{-2}\right), \end{aligned}$$

et

$$\begin{aligned} G^N_{k,k} =& \frac{(N)_{(k)}}{N^k} \\ =& 1 - \frac{1}{N}\frac{k\,(k+1)}{2} + O\left(N^{-2}\right) \ . \end{aligned}$$

Finalement, si $j < k-1$

$$G^N_{k,j} = \mathcal{S}^{(j)}_k \frac{(N)_{(j)}}{N^k} = O\left(N^{-2}\right) \ .$$

D'où

$$G^N = I + N^{-1}Q + o\left(N^{-1}\right) \ .$$

Comme $A^{(n)}_0$ est constant, égal à n, quelque soit n, on conclut par le théorème 5.12 des annexes. □

De Wright-Fisher au coalescent

Nous avons vu que, modulo un bon changement de temps, le nombre total d'ancêtres de notre échantillon converge lorsque $N \to \infty$ vers un processus qui diminue de un en un à taux $\binom{k}{2}$ lorsqu'il y a encore k ancêtres.

On fait l'approximation suivante : supposons N assez grand pour que la généalogie de l'échantillon se comporte à peu près comme ce processus. Moralement, ceci signifie qu'un couple de notre échantillon retrouve son ancêtre commun à chaque sonnerie d'une horloge exponentielle de paramètre 1.

Comme aucun couple n'est privilégié à priori ceci nous mène à considérer le modèle suivant, sous l'hypothèse d'une population d'effectif élevé :

- Partant d'un échantillon de taille n, un couple coalesce (retrouve son ancêtre commun) au bout d'un temps distribué suivant une exponentielle $\mathcal{E}\left(\binom{n}{2}\right)$
- Ce couple est choisi dans l'échantillon de manière uniforme parmi tous les couples possibles.
- On continue ensuite à reconstruire la généalogie de la même manière partant des $n-2$ individus restants, et de l'ancêtre nouvellement retrouvé.
- Le temps avant coalescence est alors distribué suivant une exponentielle indépendante des autres de paramètre $\binom{\tilde{n}}{2}$, avec $\tilde{n}$ l'effectif restant de l'échantillon.
- On s'arrête à 1 individu : c'est l'ancêtre commun de l'échantillon.

C'est sur cette idée que repose le modèle de Kingman.

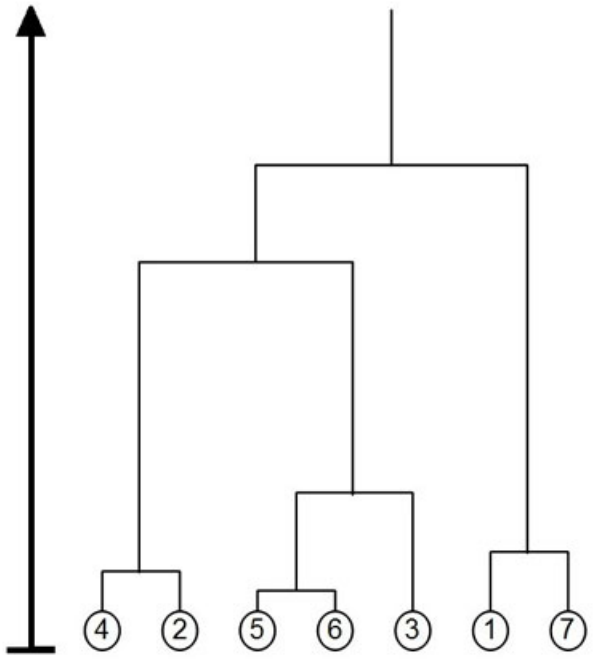


FIGURE 3 – Un processus de coalescence

2.3 Le coalescent de Kingman

2.3.1 Le n-coalescent

Dans ce paragraphe, nous allons définir formellement le processus de Kingman pour un échantillon de taille n. Avant cela, nous allons décrire la dynamique sur laquelle ce processus sera construit, qui est strictement équivalente à la dynamique décrite dans le paragraphe précédent. La dynamique est la suivante :

- on démarre en 0 avec n individus distincts.
- Chaque couple coalesce à taux 1, indépendamment des autres couples.
- Le processus s'arrête lorsqu'il ne reste plus qu'un individu : on est remonté jusqu'à l'ancêtre commun de l'échantillon.

Formellement, le processus s'écrit à l'aide d'une chaîne de Markov sur l'ensemble des partitions de $[\![1, n]\!]$ décrivant les groupes d'individus de l'échantillon ayant trouvé leur ancêtre commun, et tel que l'on démarre de la partition la plus fine pour arriver à la partition la moins fine.

On considère l'ensemble $[\![1, n]\!]$ des n premiers entiers, on note $\mathcal{P}_k$ l'ensemble des partitions de $[\![1, n]\!]$ de cardinal k, et $\mathcal{P} = \bigcup_{k=1}^{n} \mathcal{P}_k$.
Sur cet ensemble, on considère une chaîne de Markov $(P_t, t \in \mathbb{R}_+)$ en temps continu telle que $P_0 = (\{0\}, \{1\}, \ldots, \{n\})$ de Q-matrice :

$$\forall \zeta, \nu \in \mathcal{P} \quad \mathcal{Q}_{\zeta\nu} = \begin{cases} 1 & \text{si } \exists A, B \in \zeta \quad \nu = \{A \cup B\} \cup (\zeta \setminus \{A, B\}), \\ 0 & \text{sinon.} \end{cases}$$

Autrement dit, l'ensemble des valeurs accessibles depuis $\zeta \in \mathcal{P}$ est l'ensemble des partitions obtenue en fusionnant exactement deux classes de ζ. Le processus ainsi obtenu est appelé *n-coalescent de Kingman.*

Remarque 2.8. *On a la relation suivante avec le processus ancestral de la section 2.2.3 :*

$$|P_t| = A_t^n$$

De plus, si $|\zeta| = k$*, le taux de transition avant le prochain saut est donné par le nombre de partitions* ν *vérifiant* $\exists A, B \in \zeta, \quad \nu = \{A \cup B\} \cup (\zeta \setminus \{A, B\})$*. Il y en a en fait* $\dfrac{k(k+1)}{2}$*.*

2.3.2 Propriétés du modèle

Hauteur et longueur d'un n-coalescent

On note $(T_i)_{i \in [\![2,n]\!]}$ la suite des temps de saut d'un n-coalescent. Alors la hauteur de l'arbre H_n est donnée par

$$H_n = \sum_{l=2}^{n} T_l = \inf \{t \geq 0 \mid \ |P_t| = 1\} \ .$$

Par ailleurs, on vérifie que

- $\mathbb{E}(H_n) = 2 \left(1 - \dfrac{1}{n}\right)$,
- $Var(H_n) = \displaystyle\sum_{l=2}^{n} \frac{4}{k^2 (k-1)^2}$.

2.3.3 Modèle de mutation à une infinité d'allèles

D'un point de vue biologique il peut être intéressant, en plus d'étudier simplement les relations généalogiques des individus, de voir l'influence de la variabilité génétique sur l'échantillon. C'est pourquoi nous allons, dans cette section, considérer un modèle de Kingman avec mutation et en particulier un modèle de mutation dit à *une infinité d'allèles* (IMAM, infinitely many alleles model). La dynamique est la suivante :

- Conditionnellement à la généalogie, des mutations apparaissent de manière aléatoire sur chaque branche (et indépendamment des autres branches) aux instants de saut d'un processus de Poisson de taux $\theta \in \mathbb{R}_+$ donné.

- Chaque mutation donne lieu à un type unique encore jamais rencontré (infinité d'allèles).

- Chaque nouvelle mutation remplace le type précédent. (allèle par opposition à site : voir section 2.4)

- Les mutations sont neutres : elle n'ont pas d'influence sur la dynamique de la population et donc sur la généalogie.

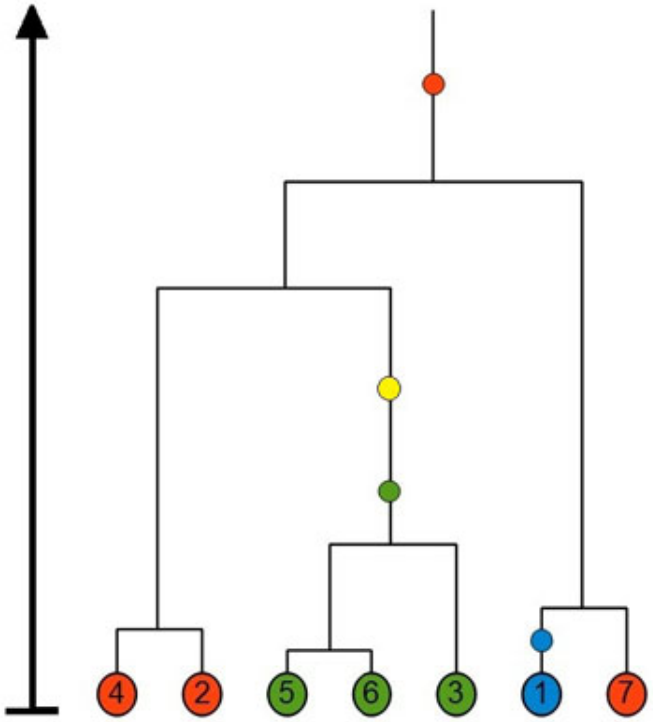


FIGURE 4 – Coalescent de Kingman avec mutation

Comme le modèle suppose que les nouveaux types remplacent les précédents, seule la dernière mutation subie compte réellement. Ainsi, nous allons nous borner à décrire ce qui se passe dans l'échantillon avant (en remontant le temps) les premières mutations.

On utilise alors la dynamique suivante :

- Comme auparavant, on part d'un échantillon de taille $n \in \mathbb{N}$.
- Chaque couple de classe (dans la partition du Kingman) coalesce à taux 1, indépendamment des autres classes.
- Chaque classe meurt à taux $\theta \in \mathbb{R}_+$ (le taux de mutation), indépendamment des autres classes.

Définition 2.9 (Coalescent avec mutation)**.**
On appelle Coalescent de Kingman avec mutation à une infinité d'allèles de le processus de Markov à valeur dans $\bigcup_{A\subset[\![1,n]\!]} \mathcal{P}(A)$ *de* Q-matrice *donnée par*

$$\forall \zeta, \nu, \quad \mathcal{Q}_{\zeta\nu} = \begin{cases} 1 & si\ \exists A, B \in \zeta \quad \nu = \{A \cup B\} \cup \zeta \setminus \{A, B\}, \\ \theta & si\ \exists A \in \zeta \quad \nu = \zeta \setminus A, \\ 0 & sinon. \end{cases}$$

Soit $(K_t, t \in \mathbb{R}_+)$ un coalescent de Kingman avec mutations IMAM, on définit alors le type d'un individu.

Définition 2.10 (Type des individus)**.**
Le type de l'individu $i \in [\![1, n]\!]$, noté par $\mathcal{T}(i)$ est défini comme

$$\mathcal{T}(i) = \inf \{t \geq 0 \mid \forall A \in K_t, i \notin A\} \ .$$

Remarque 2.11. *La définition précédente est bien cohérente au sens que, presque sûrement, aucune classe ne disparait aux mêmes instants et donc cette définition associe p.s. un type diffèrent à deux mutations diffèrentes.*

Cette manière très artificielle de définir ce processus permet néanmoins de formaliser clairement ce qu'est le coalescent de Kingman avec mutation. Ce n'est évidemment pas la seule manière de procéder pour rester cohérent avec le modèle.

Une question importante relative à ce modèle est celle du spectre de fréquence, c'est-à-dire de la répartition des types au sein de la population. En pratique, c'est presque la seule information dont disposent les biologistes. Par exemple sur la figure 4, nous sommes face à un 7-coalescent et le spectre de fréquence de l'échantillon est le suivant :

- Il y a un type représenté par 1 individu.
- Aucun type représenté par 2 individus.
- 2 types représentés par 3 individus.

Commençons par définir formellement le spectre de fréquence. La relation

$$\forall i, j \in [\![1, n]\!] \quad i \sim j \Leftrightarrow \mathcal{T}(i) = \mathcal{T}(j) \, ,$$

définit une relation d'équivalence sur $[\![1, n]\!]$. On note $\bar{k}$, pour $k \in [\![1, n]\!]$, la classe d'équivalence de k dans $([\![1, n]\!]/\sim)$ et $\text{Card}(\bar{k})$ le nombre d'éléments de $[\![1, n]\!]$ dans la classe de k.

Remarque 2.12. *Dans toute la suite le symbole # fait référence au cardinal d'un ensemble.*

Définition 2.13 (Spectre de fréquence)**.**
Pour tout $k \in [\![1, n]\!]$, on pose

$$C_k = \#\left\{ \ \bar{k} \in ([\![1, n]\!]/\sim) \mid Card\left(\bar{k}\right) = k \right\} ,$$

avec C_k est le nombre de types représentés dans l'échantillon par k individus. Le vecteur $C = (C_1, \ldots, C_n)$ est appelé le spectre de fréquence de l'échantillon.

L'étude du spectre de fréquence fait l'objet du la section 2.3.4.

Simulation d'un coalescent avec mutation IMAM

Une méthode simple de simulation d'une configuration issue du coalescent de Kingman consiste à utiliser un modèle équivalent.

Urne de Hoppe

Le modèle de l'Urne de Hoppe se décrit de la manière suivante :
On considère des balles noires de masse θ, et des balles colorées de masse 1.
On suppose qu'on a à disposition une infinité de couleurs différentes, et que tous les tirages sont faits avec remise.

- On part d'une urne contenant une unique balle noire de masse θ.

– A chaque tirage on tire aléatoirement une balle dans l'urne avec une probabilité de tirage proportionnelle à sa masse.
– Deux cas se présentent :
  → On tire une balle noire : on remet une balle d'une nouvelle couleur dans l'urne.
  → On tire une balle de couleur : on remet une balle de la même couleur dans l'urne.

Ainsi, au n-ième tirage, il y a n boules dans l'urne et on a un résultat permettant de faire le lien avec le modèle de Kingman.

Proposition 2.14 (Liens avec le coalescent)**.**
On considère un modèle de Hoppe arrêté au n-ième tirage et on note, pour tout $i \in [\![1, n]\!]$, H_i le nombre de couleurs représentées dans l'urne par i balles, alors le vecteur $(H_1, \dots, H_n)$ a exactement la loi du spectre de fréquence dans le coalescent de Kingman.

Démonstration. Nous ne faisons pas la démonstration qui est similaire à la démonstration de la proposition 2.17. □

Le modèle décrit ci-dessus permet de simuler facilement une réalisation du spectre de fréquence pour le Kingman avec mutation IMAM.

2.3.4 Formule d'échantillonnage d'Ewens

Comme nous l'avons évoqué précédemment, une des questions importantes relatives au coalescent de Kingman est celle de la loi du spectre de fréquence.

Cette question a été résolue par Warren Ewens (1972) avec le résultat suivant.

Théorème 2.15 (Formule d'échantillonnage d'Ewens)**.**
La loi du spectre de fréquence $C = (C_1, \dots, C_n)$ dans un n-coalescent IMAM est donné par

$$\forall c \in \mathbb{N}^n \qquad \mathbb{P}(C_{1:n} = c_{1:n}) = \mathbb{1}_{|c|=n} \frac{n!}{\theta_{[n]}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!},$$

avec $|c| = \sum_{i=1}^{n} i c_i$ et $\theta_{[n]} = \theta\,(\theta+1)\dots(\theta+n-1)$.

Avant d'entamer la preuve de ce résultat, nous allons introduire un autre modèle lié au coalescent avec mutation.

Modèle des arbres de Yule avec immigration

Nous considérons le modèle suivant où l'on regarde le comportement d'un système de particules :

– On démarre avec *zéro* particule.

– A taux θ, une nouvelle particule arrive, d'un nouveau type encore jamais vu.

– Chaque particule meurt en donnant naissance à deux nouvelles particules à taux 1, indépendamment des autres particules ou des phénomènes d'immigration.

Définition 2.16 (Processus de Yule)**.**
On appelle processus de Yule de paramètre θ le processus de naissance pure Markovien de taux $\theta_n = n\theta$ issu de 1.

On s'intéresse alors au nombre de particules de chaque type au cours du temps.

Soit $\left(Y^k\right)_{k\geq 1}$ une famille i.i.d. de processus de Yule de paramètre 1 et soit N un processus de Poisson de paramètre θ. On considère $Y = (Y_t)_{t\geq 0}$ le processus d'effectifs des particules au temps t, à valeur dans $\mathbb{N}^{\mathbb{N}}$ défini par

$$Y_t = \begin{cases} \left(Y_t^1, Y_{t-T_1}^2, .., Y_{t-T_{N_t}}^{N_t}, 0, 0, 0...\right) & si\ N_t > 0, \\ \mathbf{0} \in \mathbb{N}^{\mathbb{N}} & \text{sinon,} \end{cases}$$

avec $(T_i)_{i\geq 1}$ les temps de sauts du processus de Poisson N. On note $|Y_t| = \sum_{i=1}^{N_t} Y_t^i$.

Un premier résultat va nous permettre de faire le lien avec le coalescent de Kingman, nous permettant ensuite d'obtenir la formule d'Ewens.

Proposition 2.17 (Lien avec le coalescent de Kingman)**.**
Soit Y un processus suivant le modèle des arbres de Yules, alors pour tout $t \in \mathbb{R}_+$ la loi du vecteur

$$\left(\sum_{k=1}^{n} \mathbb{1}_{Y_t^k=1}, ..., \sum_{k=1}^{n} \mathbb{1}_{Y_t^k=n}\right),$$

conditionnellement à l'événement $\{|Y_t| = n\}$ est la même que celle du spectre de fréquence C de la définition 2.3.

Démonstration. Commençons par remarquer que dans le cadre du coalescent de Kingman, les temps auxquels ont lieu les différents événements (coalescence ou mutation) n'ont pas d'influence sur la configuration du spectre de fréquences : seul l'ordre des événements compte. Ainsi, en observant l'évolution du Kingman, seule compte la probabilité d'avoir une mutation ou une coalescence comme premier événement. Cette probabilité est donnée par :

$$\mathbb{P}\left(\text{“Le dernier événement est une coalescence”}\right) = \frac{n-1}{\theta+n-1},$$

$$\mathbb{P}\left(\text{“Le dernier événement est une mutation”}\right) = \frac{\theta}{\theta+n-1}\,.$$

Ceci s'observe via les propriétés du minimum de variables aléatoires exponentielles.

Dans le cadre des arbres de Yule, les coalescences (en remontant le temps) correspondent aux sauts des processus de Yule (i.e. à la scission d'une particule en deux), tandis que les mutations correspondent à l'arrivée de nouveaux immigrants (arrivée de nouvelles particules). On observe alors que (en temps rétrograde) ces probabilités sont les mêmes que pour le processus de Kingman, et cela suffit pour que les lois concordent.

□

Lemme 2.18.
Soit Y^1 un processus de Yule, alors

$$\forall t \geq 0, \quad \mathcal{L}\left(Y_t^1\right) = \mathcal{G}\left(e^{-t}\right).$$

Démonstration. Soit $(S_i)_{i\geq 1}$ une suite de variable aléatoire exponentielle telle que

$$\forall i \geq 1, \quad \mathcal{L}(S_i) = \mathcal{E}(i\lambda).$$

Alors

$$\begin{aligned}\mathbb{P}\left(Y_t^1 \geq k\right) &= \mathbb{P}\left(\sum_{i=1}^{k} S_i \leq t\right) \\ &= \int_0^t ds\ k\lambda\ e^{-k\lambda s}\ \mathbb{P}\left(\sum_{i=1}^{k-1} S_i \leq t-s\right)\ .\end{aligned}$$

On achève alors la preuve en faisant une récurrence, avec comme hypothèse de récurrence : $\mathcal{H}_n$: $\mathbb{P}\left(\sum_{i=1}^{n} S_i \leq t\right) = \left(1-e^{-\lambda t}\right)^n$. □

A partir d'ici on fixe un temps $t \in \mathbb{R}_+$; comment décrire la probabilité d'obtenir une configuration $(y_1, ..., y_k, 0...0....)$ via le processus Y, à l'instant t ?

Nous allons en fait décrire les événements subis par le processus Y à l'aide des atomes d'une mesure aléatoire de Poisson.

Soit N une mesure aléatoire de Poisson sur $[0,t] \times \mathbb{N}$ d'intensité $\theta ds \times \mathcal{G}\left(e^{-(t-s)}\right)$. On note $(S_k, J_k)_{k\geq 0}$ les atomes de la mesure N. Ces derniers ont la signification suivante :

- Les $S_i \in [0,t]$ sont les temps d'arrivée de nouvelles particules : en effet, ils correspondent aux sauts d'un processus de Poisson sur $[0,t]$.
- Les J_i correspondent aux valeurs des Yules, arrivés aux instants S_i, à l'instant t.

Notamment

$$\mathcal{L}(J_i \mid S_i) = \mathcal{G}\left(e^{-(t-S_i)}\right).$$

Ce qui implique par ailleurs que

$$\mathcal{L}\left(Y_t^1, \ldots, Y_t^{N_t} \mid S_i\right) = \mathcal{L}(J_1, \ldots, J_{N_t} \mid S_i).$$

Démonstration. Théorème 2.3 : Formule d'Ewens.
Par la proposition 2.3 on a

$$C \overset{\mathcal{L}}{=} \left(\sum_{k\geq 1} \mathbb{1}_{Y_t^k=1}, \sum_{k\geq 1} \mathbb{1}_{Y_t^k=2}, ..., \sum_{k\geq 1} \mathbb{1}_{Y_t^k=n} \,\middle|\, |Y_t| = n\right).$$

Du point de vue de la mesure N, cela correspond aux nombres d'atomes dont la deuxième composante a une valeur donnée, posons pour cela

$$\forall j \geq 1, \quad Z_j(t) = N([0,t] \times \{j\}).$$

On en déduit que

$$C \overset{\mathcal{L}}{=} \left(Z_1(t), ..., Z_n(t) \,\middle|\, \sum_{k\geq 1} kZ_k(t) = n\right).$$

D'autre part, comme les ensembles $[0,t]\times\{j\}$ et $[0,t]\times\{i\}$ sont disjoints pour $i\neq j$, les variables aléatoires $(Z_j(t))_{1\geq j\geq n}$ sont indépendantes et de lois données par

$$\forall j\geq 1,\quad \mathcal{L}\left(Z_j(t)\right)=\mathcal{P}\left(\theta\int_0^t e^{-(t-s)}\left(1-e^{-(t-s)}\right)^{j-1}ds\right)=\mathcal{P}\left(\frac{\theta}{j}\left(1-e^{-t}\right)^j\right).$$

Alors

$$\mathbb{P}\left(C_{1:n}=c_{1:n}\right)=\mathbb{P}\left(Z_{1:n}(t)=c_{1:n}\left|\sum_{k\geq 1}kZ_k(t)=n\right.\right).$$

Cette dernière expression étant vraie quel que soit $t>0$, lorsque $t\to\infty$ on obtient

$$\mathbb{P}\left(C_{1:n}=c_{1:n}\right)=\mathbb{P}\left(Z_{1:n}=c_{1:n}\left|\sum_{k\geq 1}kZ_k=n\right.\right),$$

où les $(Z_j, j\in[\![1,n]\!])$ sont indépendantes et de loi de Poisson $\mathcal{P}\left(\frac{\theta}{j}\right)$.

On conclut alors par la proposition suivante. □

Proposition 2.19.

Soit $(Z_j, j\in[\![1,n]\!])$ des variables aléatoires indépendantes de loi de Poisson $Z_j\overset{\mathcal{L}}{=}\mathcal{P}\left(\frac{\theta}{j}\right)$, alors

$$\mathbb{P}\left(Z_{1:n}=c_{1:n}\left|\sum_{k=1}^n kZ_k=n\right.\right)=\mathbb{1}_{|c|=n}\frac{n!}{\theta_{[n]}}\prod_{j=1}^n\left(\frac{\theta}{j}\right)^{c_j}\frac{1}{c_j!},$$

avec $|c|=\sum_{i=1}^n ic_i$.

Pour démontrer cette proposition commençons par introduire un lemme combinatoire qui ne sera démontré que plus tard afin de garder notre objectif en tête.

Lemme 2.20 (Permutation à k cycles)**.** *On note $s(n,k)$ le nombre de permutations de $\mathcal{S}_n$ composées de k cycles. Alors d'une part*

$$s(n,k)=\sum_{\substack{k_1+\cdots+k_n=k\\ \sum ik_i=n}}\frac{n!}{\prod_{j=1}^n j^{k_j}k_j!},$$

et d'autre part

$$\begin{cases} s(n,k)=s(n-1,k-1)+(n-1)\,s(n-1,k),\\ s(1,1)=1,\\ s(n,0)=0\quad\forall n,\\ s(n,0)=0\quad\forall n.\end{cases}$$

Démonstration.
Commençons par calculer la probabilité de l'événement $\{\sum_{k=1}^n kZ_k = n\}$:

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^n kZ_k = n\right) &= \sum_{\substack{k_1,\dots,k_n\geq 0\\ \sum ik_i = n}} \mathbb{P}\left(Z_1 = k_1, \dots, Z_n = k_n\right) \\
&= \sum_{\substack{k_1,\dots,k_n\geq 0\\ \sum ik_i = n}} \prod_{j=1}^n e^{-\theta/j} \frac{\theta^{k_j}}{j^{k_j} k_j!} \\
&= e^{-\theta\sum_{j=1}^n 1/j} \sum_{k=0}^n \sum_{\substack{k_1+\dots+k_n=k\\ \sum ik_i = n}} \prod_{j=1}^n \frac{1}{j^{k_j} k_j!}\ \theta^k \\
&= e^{-\theta\sum_{j=1}^n 1/j} \sum_{k=0}^n \frac{s\left(n,k\right)}{n!}\theta^k.
\end{aligned}$$

D'autre part on définit pour tout $n, k \in \mathbb{N}$, $a_n^{(k)}$ par

$$\theta\left(\theta+1\right)\dots\left(\theta+n-1\right) = \sum_{k=1}^n a_n^{(k)}\theta^k.$$

Alors

$$\sum_{k=1}^n a_n^{(k)}\theta^k = \theta\sum_{k=1}^{n-1} a_{n-1}^{(k-1)}\theta^k + \left(n-1\right)\sum_{k=1}^{n-1} a_{n-1}^{(k)}\theta^k.$$

Par identification des coefficients on obtient que

$$a_n^{(k)} = a_{n-1}^{(k-1)} + \left(n-1\right)a_{n-1}^{(k)}.$$

On voit alors que $\left(a_n^{(k)}\right)_{k,n}$ vérifie la relation de récurrence du lemme 2.20, d'où

$$\mathbb{P}\left(\sum_{k=1}^n kZ_k = n\right) = e^{-\theta\sum_{j=1}^n 1/j}\ \frac{\left(\theta\right)_{[n]}}{n!}.$$

On regarde ensuite la probabilité suivante :

$$\begin{aligned}
\mathbb{P}\left(Z_{1:n} = c_{1:n}, \sum_{k=1}^n kZ_k = n\right) &= \mathbb{1}_{|c|=n}\ \mathbb{P}\left(Z_{1:n} = c_{1:n}\right) \\
&= \prod_{j=1}^n e^{-\theta/j}\ \frac{\theta^{c_j}}{j^{c_j} c_j!}.
\end{aligned}$$

Ce qui achève la preuve. □

Passons à la preuve du lemme.

Démonstration. (Lemme 2.19)
Commençons par montrer la relation de récurrence :

On pose donc pour $k, n \in \mathbb{N}$

$$\begin{aligned} s(n,k) &= \#\{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k \text{ cycles}\}, \\ &= \#\{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k \text{ cycles dont le cycle } (n)\} + \#\{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k \text{ cycles mais pas le cycle } (n)\}. \end{aligned}$$

Soit $\sigma \in \mathcal{S}_n$ à k cycles telle que $\sigma(n) = n$. Dans ce cas, σ admet une restriction naturelle unique à l'ensemble $[\![1, n-1]\!]$ et cette restriction a $k-1$ cycles. Réciproquement, si on prend une permutation de $\mathcal{S}_{n-1}$ à $k-1$ cycles, on peut lui associer une unique permutation dans $\{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k \text{ cycles dont le cycle } (n)\}$.

On a donc une bijection de $\{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k \text{ cycles dont } (n)\}$ dans l'ensemble $\{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k-1 \text{ cycles}\}$, on en déduit que

$$\#\{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k \text{ cycles dont le cycle } (n)\} = s(n-1, k-1).$$

Soit maintenant $\sigma \in \{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k \text{ cycles mais pas le cycle } (n)\}$, on peut construire une restriction de cette permutation à $[\![1, n-1]\!]$ de la manière suivante :

$$\begin{array}{rccc} \sigma' & : & [\![1,n-1]\!] & \to & [\![1,n-1]\!] \\ & & i & \mapsto & \begin{cases} \sigma(i) & \text{si } \sigma(i) \neq n, \\ \sigma^2(i) & \text{si } \sigma(i) = n. \end{cases} \end{array}$$

On pose alors :

$$\begin{array}{rccc} T & : & \{\sigma \in \mathcal{S}_{n-1} \mid \sigma \text{ à } k \text{ cycles}\} \times [\![1,n-1]\!] & \to & \{\sigma \in \mathcal{S}_n \mid \sigma \text{ à } k \text{ cycles mais pas le cycle } (n)\} \\ & & (\sigma, i) & \mapsto & \tilde{\sigma}, \end{array}$$

avec

$$\tilde{\sigma}(k) = \begin{cases} \sigma(k) & \text{si } k \neq i \text{ et } k \neq n, \\ n & \text{si } k = i, \\ \sigma(i) & \text{si } k = n. \end{cases}$$

On vérifie que cette application est une bijection et on en déduit la relation de récurrence.

Nous allons à présent montrer la relation

$$s(n,k) = \sum_{\substack{k_1 + \cdots + k_n = k \\ \sum i k_i = n}} \#\{\sigma \text{ avec } k_j \text{ cycles de longueur } j \ \forall j\}.$$

Pour choisir une permutation dans un des ensembles du type $\{\sigma \text{ avec } k_j \text{ cycles de longueur } j \ \forall j\}$ il faut déjà choisir quels entiers vont dans chaque cycle. Ceci revient à mettre n balles (les entiers) dans k boîtes (les cycles) avec k_j balles dans la j-ième boîte : on a $\frac{n!}{\prod_{j=1}^{n} (j!)^{k_j}}$ possibilités.
Il faut ensuite choisir l'ordre des entiers dans chaque cycle : on a $(j-1)!$ possibilités pour un

j-cycle. Enfin, il faut tenir compte du fait que l'ordre des cycles n'a pas d'effet sur la permutation obtenue, on a donc

$$\begin{aligned} s(n,k) &= \sum_{\substack{k_1+\cdots+k_n=k \\ \sum i k_i = n}} \frac{n!}{\prod_{j=1}^{n} (j!)^{k_j}} \frac{1}{\prod_{j=1}^{n} k_j!} \prod_{j=1}^{n} (j-1)! \\ &= \sum_{\substack{k_1+\cdots+k_n=k \\ \sum i k_i = n}} \frac{n!}{\prod_{j=1}^{n} (j)^{k_j}} \frac{1}{\prod_{j=1}^{n} k_j!}. \end{aligned}$$

□

2.4 Quelques techniques d'inférence ancestrale pour le modèle de Kingman

L'objectif de cette section est : étant donné un échantillon d'obtenir des informations sur le paramètre θ du modèle. Pour exploiter au mieux les informations biologiques disponibles, nous allons devoir introduire un nouveau mécanisme de mutation. La méthode utilisée ici est issue de [13].

2.4.1 Modèle à une infinité de site

Une difficulté du modèle IMAM tient au fait que les configurations observées ne permettent pas de reconstruire la généalogie de l'échantillon précisément : on sait que deux individus de type identique doivent avoir coalescés avant d'avoir mutés (en temps rétrograde), mais on ne sait que dire de deux individus de types distincts. C'est pourquoi nous allons désormais considérer un nouveau modèle de mutation qui permet de tirer plus d'informations de l'échantillon observé.

Dans la suite, nous supposerons que chaque mutation, au lieu de donner naissance à un nouveau type, ne fait que modifier le type actuel de l'individu. Du point de vue biologique, on comprend ceci de la manière suivante : chaque mutation touche un locus différent (hypothèse justifiée dès lors qu'on suppose qu'il y a une infinité de loci) et on sait distinguer les loci de type ancestral des loci mutés : on parle du *infinitely many sites model*.

Plus précisément, tout individu i dans $[\![1,n]\!]$ peut par exemple être représenté par un s-uplet $x_i = (x_{i1}, ..., x_{is})$ où s est le nombre de mutations subies jusque là.
Et chaque nouvelle mutation fait apparaitre un nouvel élément dans le uplet représentant un individu. Ici, contrairement au modèle à une infinité d'allèles, les mutations *se cumulent au lieu de se remplacer*.

Mais ceci nous oblige à définir le coalescent de manière différente afin de prendre en compte ce nouveau mode de mutation.

Le nouveau mode de mutation est le suivant, il repose sur l'utilisation d'une mesure aléatoire de Poisson :

Soit $K = (K_s, s \geq 0)$ un coalescent de Kingman sans mutation comme défini dans la section 2.3.1. Alors, conditionnellement à K, soit $(\nu_t, t \geq 0)$ un processus ponctuel de Poisson tel que, pour tout t, v_t soit une mesure aléatoire de Poisson sur $K_t \times [0,t]$ d'intensité :

$$|K_t|\, \theta ds \times U_{K_t}, \quad \textit{avec } U_{K_t} \text{ la mesure uniforme sur les classes de } K_t.$$

On notera alors $(\zeta_j, S_j)_{j\geq 0}$ la suite des atomes de ν ordonnée par ordre croissant suivant les S_j. Munis de cet objet, nous sommes alors capables de définir à nouveau le type d'un individu, qui est en fait une mesure.

Définition 2.21 (Type des individus)**.**
Soit $i \in [\![1, n]\!]$, alors le type $\mathcal{T}(i)$ de l'individu i est un peigne de Dirac sur $\mathbb{R}_+$ défini par

$$\mathcal{T}(i) = \sum_{\substack{k\geq 0 \\ i\in\zeta_k}} \delta_k.$$

Ceci nous permet de ne nous intéresser qu'aux relations entre les types des individus. Par exemple, les individus i et j partagent-ils telle ou telle mutation ? Pour simplifier les notations on note

$$\Pi = \left\{ \sum_{k\geq 0} \delta_{u_k} \mid (u_k)_{k\geq 0} \in \mathbb{N}^{\mathbb{N}} \text{ strictement croissante} \right\}.$$

C'est l'ensemble de tous les types possibles pour un individu (et même beaucoup plus car seules les suites presque nulles nous intéressent en fait).

On appelle configuration un élément de Π^n. On note

$$\Delta : \sum_{k\geq 0} \delta_{u_k} \to \sum_{k\geq 0} \delta_{u_k} - \delta_{u^\star},$$

avec $u^\star = \inf\{u_k \mid k \in \mathbb{N}\}$. En somme, l'opérateur Δ supprime la mutation la plus récente (pour le cours normal du temps).

Dans le cadre du coalescent de Kingman, obtenir telle ou telle configuration ne dépend pas des temps d'apparition des événements (mutations ou coalescences), mais seulement de l'ordre dans lequel ces événements se produisent.

2.4.2 Estimation de la probabilité d'une configuration

Nous avons maintenant notre mécanisme de mutation. Nous choisissons alors une convention de représentation des données observées. On suppose les données obtenues sous la forme suivante :

$$(t, \mathbf{n}) \text{ avec } t = (t_1, ..., t_d),$$

où les t_i sont distincts et de la forme $t_i = \sum_{k\geq 0} \delta_{s_k^i}$, avec $(s_k^i) \subset \mathbb{R}$ croissante. On note également d le nombre de types distincts, et $\mathbf{n} = (n_1, .., n_d) \in \mathbb{N}^d$ représente la multiplicité des séquences dans l'échantillon, de sorte que $\sum_{i=1}^d n_i = n$.

Etant donné un échantillon tel que décrit précédemment, pour que celui-ci soit consistant avec un modèle généalogique, il faut que les différentes séquences présentent certaines propriétés, à savoir :

- Chaque séquence est distincte. (Comme supposé plus haut)

- Si deux séquences partagent une mutation, elles partagent toutes les mutations précédentes. Formellement :

$$(\exists x\ \mathbb{N}_+\ t_i\{y\} = 1 = t_j\{x\}) \Rightarrow (\forall y \leq x\ \ (t_i\{x\} = 1) \Leftrightarrow (t_j\{y\} = 1))$$

– On considère un état ancestral noté 0 : la mesure nulle.

On note T_n la variable aléatoire qui, à un n-coalescent, associe la configuration obtenue avec les notations ci-dessus. Nous allons donc étudier la probabilité d'obtenir une configuration donnée.

Théorème 2.22. *(Formule de récurrence sur la loi de X)*
Soit $(t,n) \in \bigcup_{l\geq 1} \left(\Pi^l \times \mathbb{N}^l\right)$, *alors*

$$\begin{aligned}\mathbb{P}(T_n=(t,n)) = &\sum_{k:n_k\geq 2} \frac{(n_k-1)\,n_k}{n\,(n-1+\theta)}\mathbb{P}\left(T_{n-1}=(t,n-e_k)\right)\\ &+\frac{\theta}{n\,(n-1+\theta)}\sum_{\substack{k:n_k=1\\ t_k\neq 0,\forall j\ t_j\neq t_k}} \mathbb{P}\left(T_n=(\Delta_k t,n)\right)\\ &+\frac{\theta}{n\,(n-1+\theta)}\sum_{\substack{k:n_k=1\\ t_k\neq 0}}\sum_{j:\Delta t_k=t_j} \mathbb{P}\left(T_n=(R_k t, R_k\,(n+e_j))\right),\end{aligned}$$

avec Δ_k *l'opérateur qui applique l'opérateur* Δ *à la* k*-ième composante de* t *et* R_k *l'opérateur qui supprime la* k*-ième composante.*

Démonstration.
On notera tout d'abord que T_n est fonction d'une famille i.i.d. de v.a exponentielle et de processus ponctuel de Poisson. On utilisera dans la suite le caractère Markovien de ces objets.

Pour démontrer ce résultat, nous allons conditionner par le premier (en temps backward, qui est le temps normal pour le coalescent) événement ayant eu lieu pour le processus : une coalescence ou une mutation.

On note
– E_{mut}="le dernier événement est une mutation",
$\rightarrow$ sur E_{mut}, on définit K_{mut} tel que l'individu ayant muté est de type $t_{K_{mut}}$,
– E_{coal}="le dernier événement est une coalescence",
$\rightarrow$ sur E_{mut}, on définit K_{coal} tel que l'individu né à l'instant de la coalescence est de type $t_{K_{coal}}$.

$$\begin{aligned}\mathbb{P}\left(T_n=(t,n)\right) &= \mathbb{P}\left(T_n=(t,n)\mid E_{coal}\right)\mathbb{P}\left(E_{coal}\right)+\mathbb{P}\left(T_n=(t,n)\mid E_{mut}\right)\mathbb{P}\left(E_{mut}\right)\\ &= \frac{n-1}{n-1+\theta}\,\mathbb{P}\left(T_{n-1}=(t,n-e_{K_{coal}})\middle|\, E_{coal}\right)\\ &\quad+\frac{\theta}{n-1+\theta}\,\mathbb{P}\left(T_n=(\Delta_{K_{mut}}t,n),\forall j\ t_j\neq \delta t_{K_{mut}}\middle|\, E_{mut}\right)\\ &\quad+\frac{\theta}{n-1+\theta}\,\mathbb{P}\left(T_n=(R_{K_{mut}}t,R_{K_{mut}}\,(n+e_j)),\exists j\ t_j=\delta t_{K_{mut}}\middle|\, E_{mut}\right)\end{aligned}$$

On notera que dans le dernier terme, si il existe, j est unique.
Puisque les indices K_{coal} et K_{mut} sont distribués uniformément dans les parties de la population susceptibles de subir de tels événements, K_{coal} et K_{mut} ont pour loi

$$\mathbb{P}\left(K_{coal}=k\mid E_{cola}\right)=\frac{n_k}{n}\mathbb{1}_{n_k\leq 2}$$

et

$$\mathbb{P}\left(K_{mut} \mid E_{mut}\right) = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{n_i=1}} \mathbb{1}_{n_k=1},$$

conditionnellement à, respectivement, E_{coal} et E_{mut}.

Le processus de Kingman étant Markovien, on peut enlever les conditionnements, car les types ne dépendent que de ce qui s'est passé après. Ce qui permet d'obtenir l'expression désirée.

On notera que deux cas de mutations ont été distingués : le cas où le type résultant de la suppression de la dernière mutation est encore représenté dans la population et le cas où le type résultant ne l'est pas. □

Cependant, dans ce modèle, les individus de l'échantillon sont ordonnés, ce qui n'est bien évidement pas le cas dans la nature, ni dans les données observées. Il est donc nécessaire pour être cohérent avec la pratique, de considérer toutes les trajectoires possibles menant à une même configuration (sans tenir compte de l'ordre). Pour cela, et comme on a symétrie entre les individus dans le processus de Kingman, il suffit de renormaliser par le nombre de permutations distinguables de la configuration, comme suit :

Soit $(t_0, \ldots, t_n), (l_0, \ldots, l_n) \in \Pi^n$. On note $\sim$ la relation d'équivalence

$$(t_0, \ldots, t_n) \sim (l_0, \ldots, l_n) \;\Leftrightarrow\; \exists \sigma \in \mathcal{S}_{n+1} \;\; (t_0, \ldots, t_n) = \left(l_{\sigma(0)}, \ldots, l_{\sigma(n)}\right).$$

On note $\tilde{T}$ la configuration indépendamment de l'ordre, c'est-à-dire sur l'espace quotient de $\cup\Pi$ par $\sim$, alors

$$\mathbb{P}\left(\tilde{T} \sim (t,n)\right) = \sum_{\substack{\sigma \in \mathcal{S}_n \\ \sigma \;\; distinguable}} \mathbb{P}\left(T = (t,n)\right) = \frac{n!}{n_1!...n_d!} \mathbb{P}\left(T = (t,n)\right).$$

Alors la formule récursive devient

$$\begin{aligned}
\mathbb{P}\left(\tilde{T} \sim (t,n)\right) = & \sum_{k:n_k \geq 2} \frac{(n_k - 1)}{(n - 1 + \theta)} \mathbb{P}\left(\tilde{T}_{n-1} \sim (t, n - e_k)\right) \\
& + \frac{\theta}{n\,(n-1+\theta)} \sum_{\substack{k:n_k=1 \\ t_k \neq 0, \forall j \;\; t_j \neq t_k}} \mathbb{P}\left(\tilde{T}_n \sim (\Delta_k t, n)\right) \\
& + \frac{\theta}{n\,(n-1+\theta)} \sum_{\substack{k:n_k=1 \\ t_k \neq 0}} \sum_{j:\Delta t_k = t_j} (n_j + 1)\, \mathbb{P}\left(\tilde{T}_n \sim (R_k t, R_k\,(n + e_j))\right).
\end{aligned}$$

C'est à partir de cette formule que nous allons calculer les probabilités des configurations et finalement inférer θ. La technique utilisée est basée sur une méthode de maximisation de vraisemblance rudimentaire : pour une configuration donnée $(t, \mathbf{n})$, on cherchera le maximum de l'application

$$\begin{array}{ccc} \mathbb{R}_+ & \to & [0,1] \\ \theta & \mapsto & \mathbb{P}_\theta\left(\tilde{T} \sim (t, \mathbf{n})\right). \end{array}$$

La vraie difficulté consiste à calculer la probabilité d'une configuration lorsque celle-ci est de grande taille : en effet, même si théoriquement cette formule récursive nous permettrait de le faire, en pratique le calcul est trop lourd par cette méthode. Nous allons donc utiliser un moyen détourné.

Calcul de vraisemblance pour les arbres par des méthodes MCMC

Commençons par introduire un lemme que nous ne démontrerons pas mais dont une version plus générale sera démontrée dans la seconde partie du mémoire.

Lemme 2.23.
Soit $(X_k, k \in \mathbb{N})$ une chaîne de Markov à valeur dans S au plus dénombrable, de probabilité de transition P. Soient également f une fonction sur S, $A \subset S$ et

$$\eta = \inf\{k \geq 0 \mid X_k \in A\}, \quad \textit{le temps d'atteinte de } A.$$

On suppose que $\mathbb{P}(\eta < \infty) = 1$ et on pose

$$\forall x \in S, \quad u_x(f) = \mathbb{E}_x\left[\prod_{k=0}^{\eta} f(X_k)\right]. \tag{1}$$

Alors

$$\forall x \in S \setminus A, \quad u_x(f) = f(x) \sum_{y \in S} p_{xy} u_y(f).$$

Remarque 2.24. $\forall x \in A \quad u_x(f) = f(x)$.

L'idée est la suivante : construire une chaîne de Markov sur notre espace et une fonction f, de telle manière que la relation 1 et la formule récursive concordent. Pour cela, soit f tel que

$$\begin{aligned} f(t,n) = &\sum_{k:n_k \geq 2} \frac{(n_k - 1) n_k}{n(n-1+\theta)} \\ &+ \frac{\theta}{n(n-1+\theta)} \sum_{\substack{k:n_k=1 \\ t_k \neq 0, \forall j \ t_j \neq t_k}} 1 \\ &+ \frac{\theta}{n(n-1+\theta)} \sum_{\substack{k:n_k=1 \\ t_k \neq 0}} \sum_{j:\delta t_k = t_j} 1, \end{aligned}$$

si $d > 2$ et

$$f(t,n) = \mathbb{P}\left(\tilde{T} \sim (t,n)\right), \text{ si } d = 2.$$

Sachant qu'il est facile de calculer le probabilité d'une configuration dans le cas $d = 2$ et de façon à ce que f renormalise notre formule récursive pour en faire une relation de type probabilité de transition Markovienne. On définit alors une fonction de transition de la manière suivante :

$$\forall (t,n) \in \bigcup_{l \geq 3} \left(\Pi^l \times \mathbb{N}^l\right), \ \forall (t',n') \in \bigcup_{l \geq 2} \left(\Pi^l \times \mathbb{N}^l\right),$$

$$P_{(t,n)(t',n')} = \begin{cases} \dfrac{n_k - 1}{f\left(t,n\right)\left(n-1+\theta\right)} & si\ \exists k\ \left(t',n'\right) = \left(t, n - e_k\right), \\ \dfrac{\theta}{f\left(t,n\right) n \left(n-1+\theta\right)} & si\ \exists k\ \left(t',n'\right) = \left(\delta_k t, n\right), \\ \dfrac{\theta\left(n_{j+1}\right)}{f\left(t,n\right) n \left(n-1+\theta\right)} & si\ \exists k,j\ \left(t',n'\right) = \left(R_k t, R_k\left(n+e_j\right)\right), \\ 0 & \text{dans tous les autres cas.} \end{cases}$$

On a en fait défini f de manière à ce que P soit effectivement une probabilité de transition.

Soit ensuite A l'ensemble des arbres n'ayant que *deux* séquences et η le temps d'atteinte de A. Alors par le lemme 2.3,

$$\begin{aligned} \mathbb{E}_{(t,n)}\left[\prod_{k=0}^{\eta-1} f\left(X_k\right)\mathbb{P}\left(\tilde{T} \sim X_\eta \mid X_\eta\right)\right] &= f\left(t,n\right)\ \Bigg[\sum_{k:n_k\geq 2} \frac{\left(n_k-1\right)}{f\left(t,n\right)\left(n-1+\theta\right)}\mathbb{P}\left(\tilde{T}_{n-1} \sim \left(t, n-e_k\right)\right) \\ &\quad + \frac{\theta}{f\left(t,n\right) n\left(n-1+\theta\right)} \sum_{\substack{k:n_k=1 \\ t_k\neq 0, \forall j\ t_j\neq t_k}} \mathbb{P}\left(\tilde{T}_n \sim \left(\delta_k t, n\right)\right) \\ &\quad + \frac{\theta}{f\left(t,n\right) n\left(n-1+\theta\right)} \sum_{\substack{k:n_k=1 \\ t_k\neq 0}} \sum_{j:\delta t_k = t_j} \left(n_j+1\right)\mathbb{P}\left(\tilde{T}_n \sim \left(R_k t, R_k\left(n+e_j\right)\right)\right)\Bigg] \\ &= \mathbb{P}\left(\tilde{T} \sim \left(t,n\right)\right). \end{aligned}$$

Méthode de calcul pratique

Soit maintenant $\left(T^i\right)_{i\geq 1}$ une suite i.i.d. de chaîne de Markov de loi P, alors par la loi forte des grands nombres

$$\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m}\left(\prod_{k=0}^{\eta_i} f\left(T_k^i\right)\right) = \mathbb{P}\left(\tilde{T} \sim \left(t,n\right)\right)\ p.s.$$

Ceci permet de calculer une approximation de la probabilité numériquement.

3 Splitting Tree, processus ponctuel de coalescence, et processus de contour

Cette seconde partie reprend la structure de la partie 1. C'est-à-dire que nous allons commencer par introduire une dynamique de population (temps foreward), puis nous allons reconstruire la généalogie d'une population évoluant sous cette dynamique (temps backward). La présentation de ce modèle est basée sur Lambert [9], qui s'inspire de travaux plus anciens de Geiger et Kersting [6] .

3.1 Introduction

Dans cette section nous allons étudier le modèle des splitting trees. Ce modèle prend la forme suivante : on se donne une mesure finie Λ à support sur $(0, \infty]$, appelée mesure de durée de vie, de masse totale $b \in \mathbb{R}$. On considère la dynamique suivante :

- Chaque individu a une durée de vie indépendante des autres individus, et se reproduit à taux b conditionnellement à sa durée vie et indépendamment des autres individus :
  c'est ce qui donne la propriété de branchement.

- La durée de vie de chaque individu est distribuée suivant la loi de probabilité Λ/b.
  Cette loi n'étant pas nécessairement sans mémoire, le modèle est non Markovien.

- L'arbre commence par un unique individu racine, appelé "ancêtre".

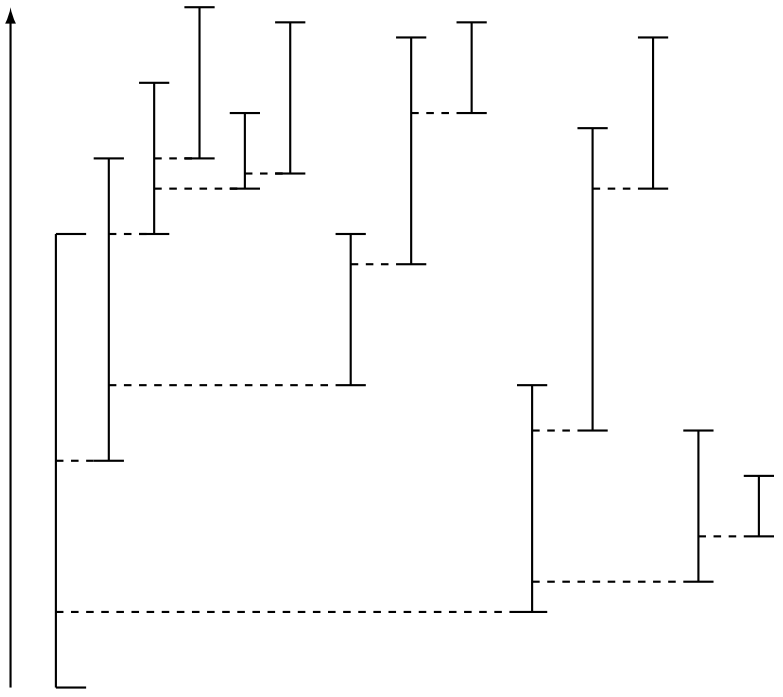


FIGURE 5 – Splitting Tree

Remarque 3.1. *On pourrait se demander pourquoi prendre b en particulier comme taux de reproduction. En fait, ce n'est absolument pas un choix contraignant puisque si on désire une autre*

valeur de taux c, il suffit de poser $\tilde{\Lambda} = \frac{c}{b}\,\Lambda$ *pour avoir le taux désiré, alors que la loi de la durée de vie reste inchangée.*
Cependant, cette convention permet d'obtenir des formules un peu plus compactes par la suite.

3.2 Construction formelle du modèle

3.2.1 $\mathbb{R}$-tree

Un $\mathbb{R}$-tree est la donnée d'une généalogie et d'une famille de réels représentant des temps (durées de vie et dates de naissance). La figure 5 est un exemple de représentation graphique d'un splitting tree, où :

- Chaque barre verticale représente un individu.
- La longueur de la barre correspond à la durée de vie de l'individu.
- Les connections pointillés représentent les filiations : pour deux individus connectés, celui de gauche est le père et celui de droite le fils.

En fait, les splitting trees ne sont que des $\mathbb{R}$-trees aléatoires. Afin de décrire formellement les $\mathbb{R}$-trees, commençons par introduire quelques définitions. On commence par définir ce qu'est une généalogie, codant les relations entre les individus sans aucune structure temporelle.

Définition 3.2 (Généalogie)**.**
Soit $\mathcal{U} = \bigcup_{n\geq 0} \mathbb{N}^n$*, on appelle généalogie tout sous-ensemble* $\mathcal{T} \subset \mathcal{U}$ *tel que*

- $\emptyset \in \mathcal{T}$,
  (L'ancêtre avec la notation $\mathbb{N}^0 = \emptyset$*)*
- $\forall j \in \mathbb{N},\ (uj \in \mathcal{T}) \Rightarrow (u \in \mathcal{T})$,
  (Consistance de la généalogie : si un individu est dans l'arbre, alors son ancêtre l'est aussi)
- $\forall j \in \mathbb{N},\ (uj \in \mathcal{T}) \Rightarrow (\forall 0 \leq i \leq j\ ui \in \mathcal{T})$.
  (Les enfants sont numérotés sans saut.)

Sur cette structure généalogique, on peut définir une relation d'ordre qui servira plus tard à construire une distance et une relation d'ordre sur les $\mathbb{R}$-trees.

Définition 3.3 (Relation d'ordre partielle sur $\mathcal{T}$)**.**
On définit une relation d'ordre $\prec$ *sur* $\mathcal{T}$ *par :*

$$\forall x, y \in \mathcal{T}, \quad x \prec y \text{ si et seulement si } \exists k \in \mathbb{N}^\star,\ \exists w \in \mathbb{N}^k,\ \ xw = y\ .$$

On dit que y est un descendant de x.

Puis on introduit la structure temporelle, on utilise pour cela l'ensemble $\mathbb{U} = \mathcal{U} \times [0, \infty)$ (on associe à chaque individu un intervalle réel correspondant à tout les instants où il est en vie) .

Soit p_1 la projection de $\mathbb{U}$ sur $\mathcal{U}$ et p_2 la projection de $\mathbb{U}$ sur $[0, \infty)$.

Définition 3.4 ($\mathbb{R}$-tree)**.**
Soit $\mathbb{U} = \mathcal{U} \times [0, \infty)$*, alors on appelle* $\mathbb{R}$*-tree tout sous-ensemble* $\mathbb{T} \subset \mathbb{U}$ *tel que*

- $(\emptyset, 0) \in \mathbb{T}$,
  (L'ancêtre né à l'instant zéro)

- $\mathcal{T} := p_1(\mathbb{T})$ *est une généalogie,*

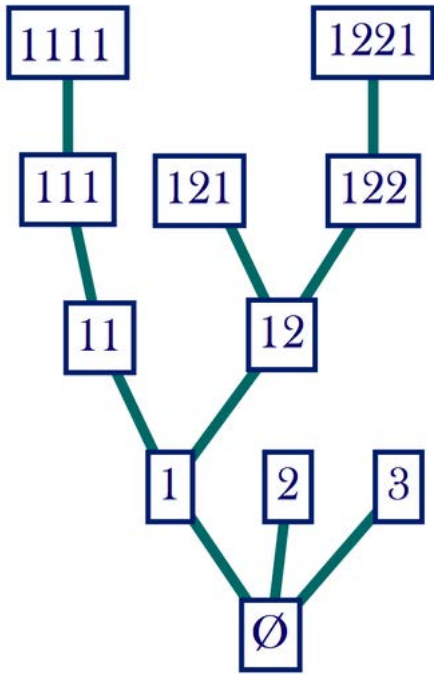


FIGURE 6 – Généalogie

– $\forall u \in \mathcal{T}$, *il existe* $\alpha(u) < \omega(u)$, *tels que* $(u, \sigma) \in \mathbb{T}$ *si et seulement si* $\sigma \in (\alpha(u), \omega(u)]$,

– $\forall u \in \mathcal{T},\ \forall j \in \mathbb{N},$ *tels que* $uj \in \mathcal{T},$ *alors* $\alpha(uj) \in (\alpha(u), \omega(u))$
(Les enfants naissent pendant la durée de vie de leur parents),

– $\forall u \in \mathcal{T},\ \forall i, j \in \mathbb{N},$ *tels que* $uj, ui \in \mathcal{T}$ *et* $i \neq j,$ *alors* $\alpha(ui) \neq \alpha(uj)$
(Et ne naissent jamais au mêmes instants).

La relation d'ordre suivante correspond à la relation de descendance :

Définition 3.5 (Relation d'ordre partielle sur $\mathbb{T}$)**.**
On peut également définir une relation d'ordre partielle sur $\mathbb{T}$ *par*

$$\forall x = (u, \sigma), y = (v, \eta) \in \mathbb{T}, \quad x \prec y \text{ si et seulement si } u \prec v, \text{et}$$

- *si* $u{=}v$, *alors* $\sigma < \eta$,
- *si* $u \neq v$, *alors* $\sigma \leq \alpha(uj)$ *avec* j *l'unique entier tel que* $uj \prec v$.

On dit que y est un descendant de x.

Cette relation n'a pour but que de permettre l'introduction d'une relation d'ordre total sur $\mathbb{T}$ qui est d'une grande importance pour "parcourir" l'arbre. On définit ensuite la notion de segment sur l'arbre. Cette notion nous permet d'étendre la mesure de Lebesgue sur $\mathbb{T}$.

Pour tout $x \in \mathbb{T}$, on note $[\emptyset, x] = \{y \in \mathbb{T} \mid y \prec x\}$.

Lemme 3.6 (MRCA, most recent common ancestor (Lambert 2010))**.**
Soit $x, y \in \mathbb{T}$, *alors il existe* $z \in \mathbb{T}$ *tel que* $[\emptyset, x] \cap [\emptyset, y] = [\emptyset, z]$. *On appelle* z *le plus récent ancêtre commun de* x *et* y, *et on le note* $y \wedge x$.

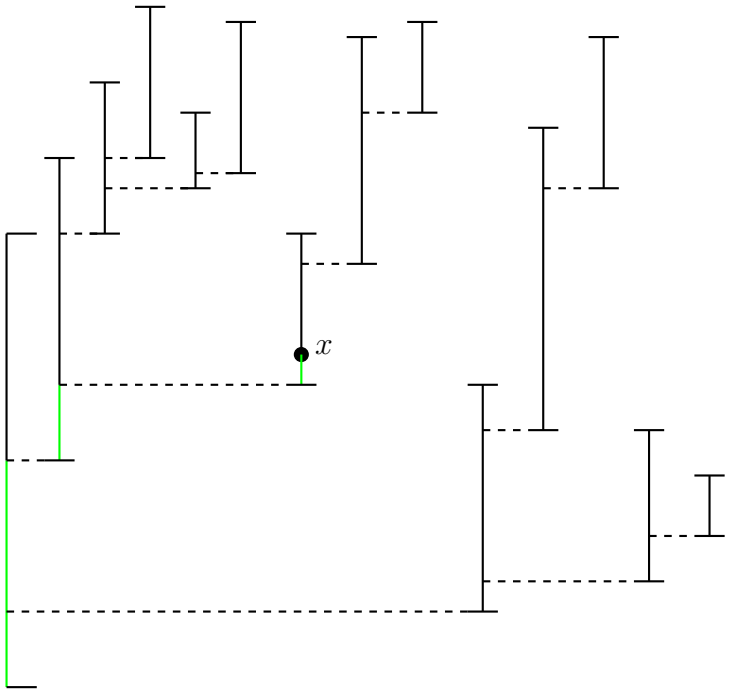


FIGURE 7 – Segment $[\rho, x]$

Définition 3.7 (Segment)**.**
Soit $x, y \in \mathbb{T}$, le segment reliant x à y, noté $[x, y]$, dans $\mathbb{T}$ est l'ensemble

$$[x, y] := ([\emptyset, x] \cap [\emptyset, y]) \setminus [\emptyset, x \wedge y].$$

Ces définitions sont illustrées par les figures 7 et 8.

Définition 3.8 (Distance sur les arbres)**.**
Soit $x, y \in \mathbb{T}$, alors l'application

$$d(x, y) := p_2(x) + p_2(y) - p_2(x \wedge y),$$

définit une distance sur l'arbre $\mathbb{T}$, faisant ainsi de l'arbre un espace métrique.

Nous venons de munir $\mathbb{T}$ d'une topologie et par la même d'une tribu borelienne. L'objectif suivant est de définir une mesure de Lebesgue sur l'arbre.

Proposition 3.9.
La tribu engendrée par l'ensemble des segments sur $\mathbb{T}$ est la tribu borélienne associée à la topologie induite par d.

Démonstration.
Soit $x \in \mathbb{T}$, et $\epsilon > 0$, alors

$$B(x, \epsilon) = \bigcup_{y \in \partial B(x, \epsilon)}]y, x].$$

L'ensemble $\partial B(x, \epsilon)$ étant au plus dénombrable, la preuve en découle facilement. □

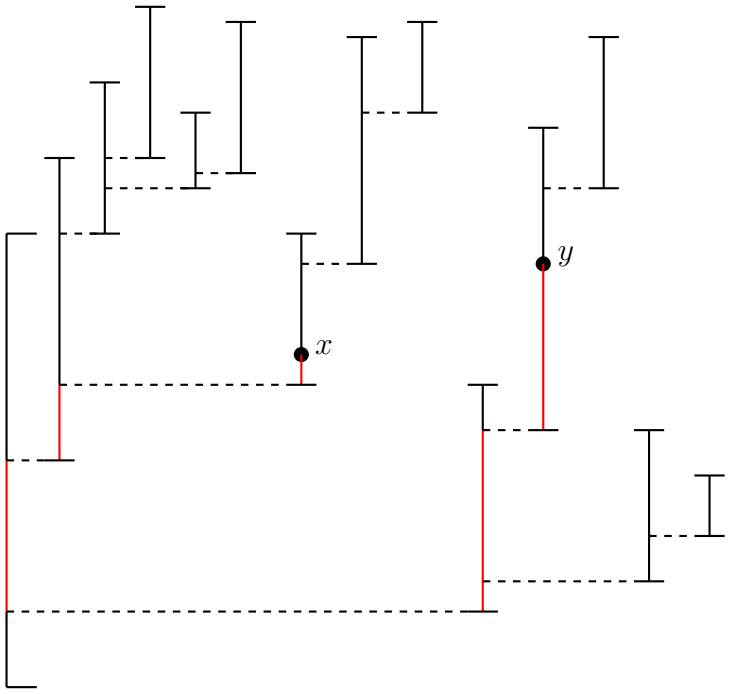


FIGURE 8 – Segment $[x, y]$

Construction de la mesure de Lebesgue

On vérifie que l'ensemble des segments satisfait les conditions du théorème de Caratheodory, puis on définit sur chaque segment

$$Leb\left([x,y]\right) = d(x,y).$$

Puis, par Caratheodory, on étend cette application en une mesure sur la tribu borélienne qui, on l'a vu, est engendrée par les segments.

Nous allons à présent construire une relation d'ordre totale sur $\mathbb{T}$, celle-ci sert à définir le processus de contour. On note $a\left(x\right) = \sup\left\{\sigma \in \left(0, \omega\left(\emptyset\right)\right) \mid \left(\emptyset, \sigma\right) \prec x\right\}$ et

$$A\left(x\right) = \left(\emptyset, a(x)\right) \in \mathbb{T}.$$

Le point $A(x)$ est un point de branchement, il coupe l'arbre en deux parties.

Définition 3.10. *Relation d'ordre totale sur $\mathbb{T}$ Soit $x, y \in \mathbb{T}$, on note*

$$y \leq x \;\Leftrightarrow\; y \in \left\{z \in \mathbb{T} \mid A\left(x\right) \prec z\right\} \setminus \left[A\left(x\right), x\right].$$

C'est-à-dire que $y \leq x$ si y descend de $A\left(x\right)$ mais n'est pas un ancêtre de x (Voir figure 10).

3.2.2 Construction de la loi des Splitting trees.

Définition 3.11 (Recollement d'arbre)**.**
Soit $\mathbb{T}$ et $\mathbb{T}'$ deux $\mathbb{R}$-trees. Soit $x = (u, \sigma) \in \mathbb{T}$ tel que
– $\forall k \in \mathbb{N}^\star,\ \alpha(ui) \neq \sigma$ *(x n'est pas un point de naissance),*

– $\omega(u) \neq \sigma$ *(x n'est pas le bout d'une feuille).*

On peut alors définir le recollement de $\mathbb{T}'$ *sur* $\mathbb{T}$ *en* x *comme descendant de* ui $(i \in \mathbb{N}^\star)$*. On pose*

$$\tilde{\mathbb{T}} = \left\{ \mathbb{T} \backslash \{(v,\eta) \in \mathbb{T} \mid v = ukw \text{ avec } k \geq i\} \right\} \cup \{(u(k+1),\eta) : (v,\eta) \in \mathbb{T} \text{ avec } v = ukw \text{ et } k \geq i\}.$$

Le recollement est alors

$$g(\mathbb{T}', \mathbb{T}, x, i) = \tilde{\mathbb{T}} \cup \{(uiv, \sigma + \eta) : (v,\eta) \in \mathbb{T}')\}.$$

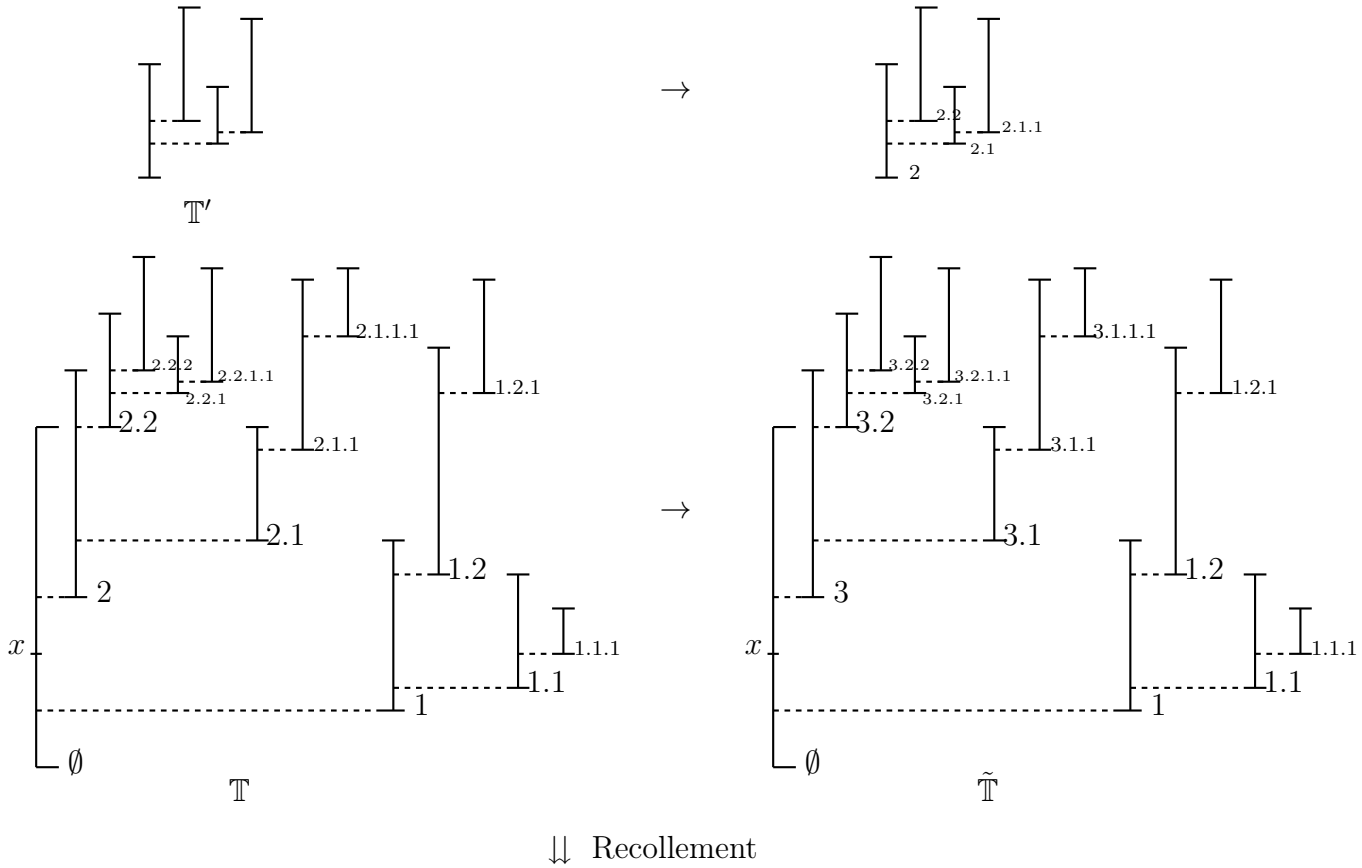
Muni de cette opération, on peut alors caractériser la mesure de probabilité, définie sur l'espace des arbres, associée a notre modèle.

On note $\mathbb{P}_\chi$, $\chi \in \mathbb{R}_+ \cup \{\infty\}$, la loi d'un splitting tree partant d'un ancêtre $\emptyset$ de durée de vie χ. Soit également $(\alpha_i, \zeta_i)_{i \geq 0}$ les atomes d'une mesure de Poisson sur $(0,\chi) \times (0,\infty]$ d'intensité Leb$\times\Lambda$ (on se réfèrera à [9] pour les détails : par exemple la méthode de numérotation des atomes).

Alors la loi $\mathbb{P} = (\mathbb{P}_\chi)_{\chi \in \mathbb{R}_+}$ des splitting tree est l'unique famille de mesures de probabilité telle que

$$\mathbb{P} = \mathcal{L}\left(\bigcup_{n \geq 1} g(T_n, \emptyset \times (0,\chi), (\emptyset, \alpha_n), n) \right),$$

avec, pour tout n, $\mathbb{T}_n$ suit la loi $\mathbb{P}_{\zeta_n}$ conditionnellement à la mesure de Poisson ci-dessus.



$\Downarrow\Downarrow$ Recollement

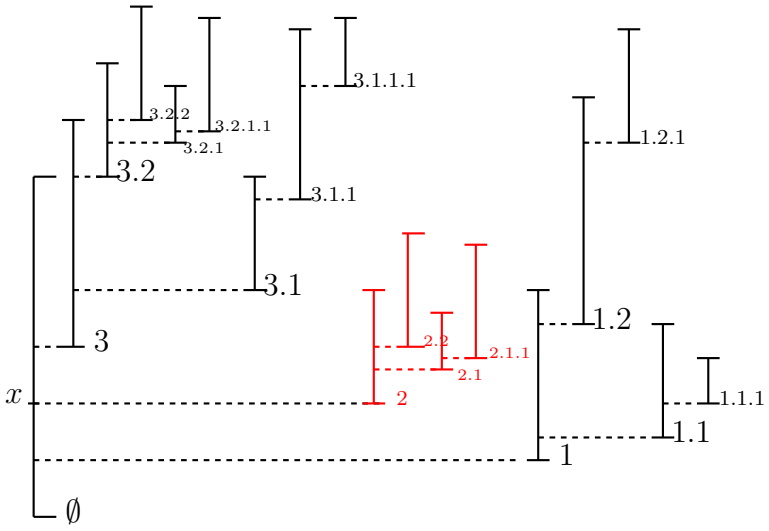


FIGURE 9 – Recollement de deux arbres :$g\,(\mathbb{T}', \mathbb{T}, x, 2)$

3.3 Processus de contour

Les processus de contour sont des outils importants de l'étude des arbres. Pour les Galton-Watson, le processus de contour standard s'appelle le processus de contour de Lukasiewicz (voir [10]). Ici, nous allons introduire une autre notion de processus de contour.

3.3.1 Présentation

Notre but ici est, comme cela a été fait pour le modèle de Wright-Fisher et le coalescent de Kingman, de reconstruire la généalogie de la population vivante à un temps t pour le modèle des splitting trees. Nous allons pour cela utiliser un outil intermédiaire, qui permet de caractériser la loi des temps de coalescence entre les individus vivants au temps t.

Nous définissons pour cela ce qu'on appelle le processus de contour de l'arbre (voir figure 12). C'est un processus cadlàg sans saut négatif où la hauteur des sauts est égale à la durée de vie des individus de l'arbre. Ce processus peut être décrit de la façon suivante :

- En 0, le processus est égal à la durée de vie de l'ancêtre.
- Le processus décroit linéairement avec une pente -1 pendant la durée de vie restant à l'ancêtre après la naissance de son dernier enfant.
- Il saute alors d'une hauteur égale à la durée de vie de l'enfant.
- Il parcourt alors en sens inverse la vie de l'enfant à vitesse -1 depuis la date de sa mort jusqu'à rencontrer une naissance (auquel cas il recommence comme ci-dessus) ou bien jusqu'à remonter à la naissance de l'individu (auquel cas il continue son parcours à l'envers sur la durée de vie de l'individu qui lui a donné naissance, en partant du même instant).
- Ainsi de suite, jusqu'à ce que le processus de contour ait atteint l'instant de naissance de l'ancêtre $\emptyset$.

3.3.2 Construction formelle

Nous allons à présent construire de manière rigoureuse le processus de contour d'un splitting tree de mesure total finie. Pour cela, nous le définissons dans le cas d'un arbre $\mathbb{T}$ déterministe de longueur totale finie $Leb(\mathbb{T}) = l$. Soit maintenant

$$\begin{array}{rccc} \varphi & : & \mathbb{T} & \to & [0,l] \\ & & x & \mapsto & Leb(\{y \in \mathbb{T} \mid y \leq x\}). \end{array}$$

(Voir figure 11.)

Proposition 3.12 (Lambert 2010)**.**
La fonction φ définie ci-dessus est une bijection mesurable de $\mathbb{T} \to [0, Leb(\mathbb{T})]$ croissante et transformant la mesure de Lebesgue sur $\mathbb{T}$ en la mesure de Lebesgue sur $[0, Leb(\mathbb{T})]$.

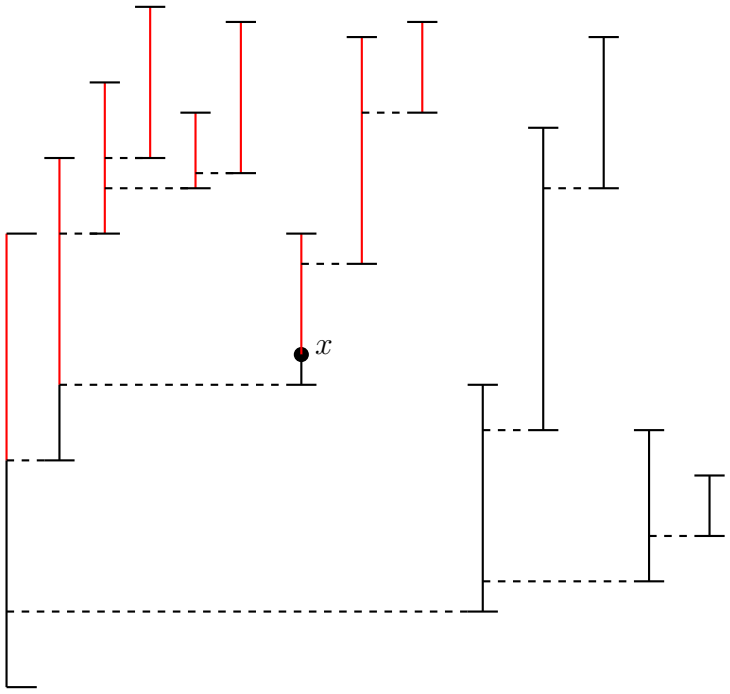


FIGURE 10 – En rouge l'ensemble $\{y \in \mathbb{T} \mid y \leq x\}$

Définition 3.13 (Processus de contour)**.**
On appelle processus de contour associé à $\mathbb{T}$, l'application

$$\begin{array}{rccc} X_t & : & [0,l] & \to & \mathbb{R}_+ \\ & & t & \mapsto & p_2\left(\varphi^{-1}(t)\right). \end{array}$$

(Voir figure 12.)

3.3.3 Reconstruire la généalogie

Soit $t \in \mathbb{R}_+$. On se propose de regarder la population vivante à l'instant t dans le splitting tree. Mais avant tout, explicitons ce qu'on entend par individu vivants.

Définition 3.14 (Individu vivant)**.**
L'ensemble des individus vivant à un instant $t \in \mathbb{R}_+$ donné est l'ensemble

$$\{(x,\sigma) \in \mathbb{T} \mid \sigma = t\}.$$

Dans l'exemple de la figure 13 on a 4 individus vivants au temps t qui sont, avec la notation de Neuveu : $\emptyset,\ 2,\ 21,\ 12$. Afin de reconstruire la généalogie de la population, intéressons-nous aux temps de coalescence entre les lignées de deux individus successifs vivants dans l'arbre à l'instant t (les temps $H_1,\ H_2,\ H_3$) et à l'ordre de coalescence. Dans l'exemple précédent :

– Dans un premier temps 21 et 2 coalescent au bout du temps H_2.

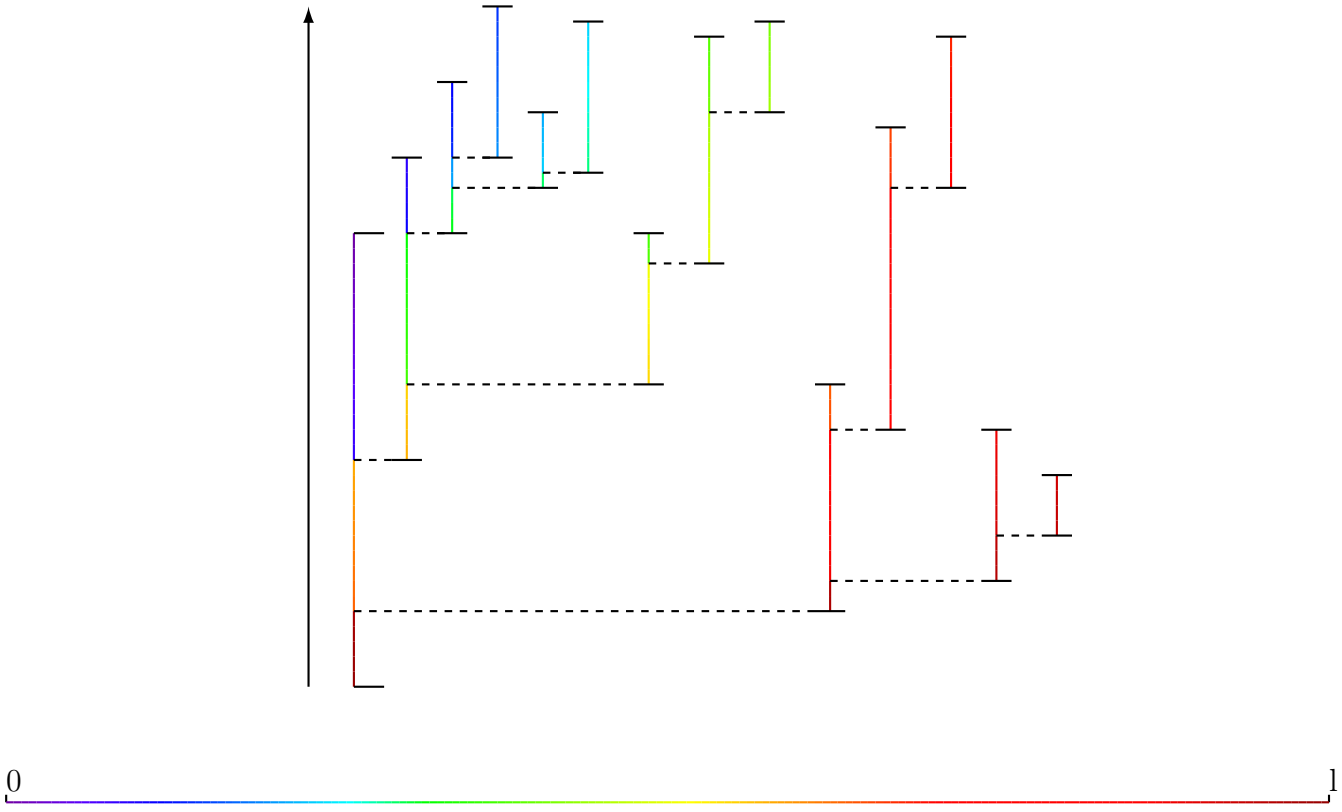


FIGURE 11 – Bijection φ entre $[0, Leb(\mathbb{T})]$ et l'arbre

– Puis la lignée $\{21, 2\}$ coalesce avec l'individu ancestral $\emptyset$ au bout du temps H_1.

– Enfin toutes les lignées se rejoignent au bout du temps H_3.

C'est dans le but de retrouver les temps de coalescence que le processus de contour va avoir son intérêt. En effet, il existe des connexions fortes entre ces temps et le processus de contour.

Proposition 3.15 (Temps de coalescence, Lambert 2010)**.**
Soit $\mathbb{T}_t := \{(x, \sigma) \in \mathbb{T} \mid \sigma = t\}$ *l'ensemble des individus vivants au temps* t*. On suppose que* $\mathbb{T}_t \neq \emptyset$*, alors*

$$\mathbb{T}_t = p_2^{-1}(t)$$

et

$$\mid \mathbb{T}_t \mid < \infty \ p.s.$$

Si, on note $(x_i)_{\geq 0}$ *la suite réordonnée de manière croissante pour l'ordre total* $\leq$ *de la population vivante au temps* t*, alors*

$$Coal(x_i, x_{i+1}) = \inf\{X_t \mid t \in [\varphi(x_i), \varphi(x_{i+1})]\}.$$

Chaque individu du splitting tree vivant au temps t est parcouru une fois par le processus de contour à la hauteur t. Il n'y a donc qu'un nombre fini de x_i dans la proposition précédente. La population vivante au temps t est illustrée par la figure 13. Ainsi, les temps de coalescences correspondent aux profondeurs des excursions du processus de contour sous t.

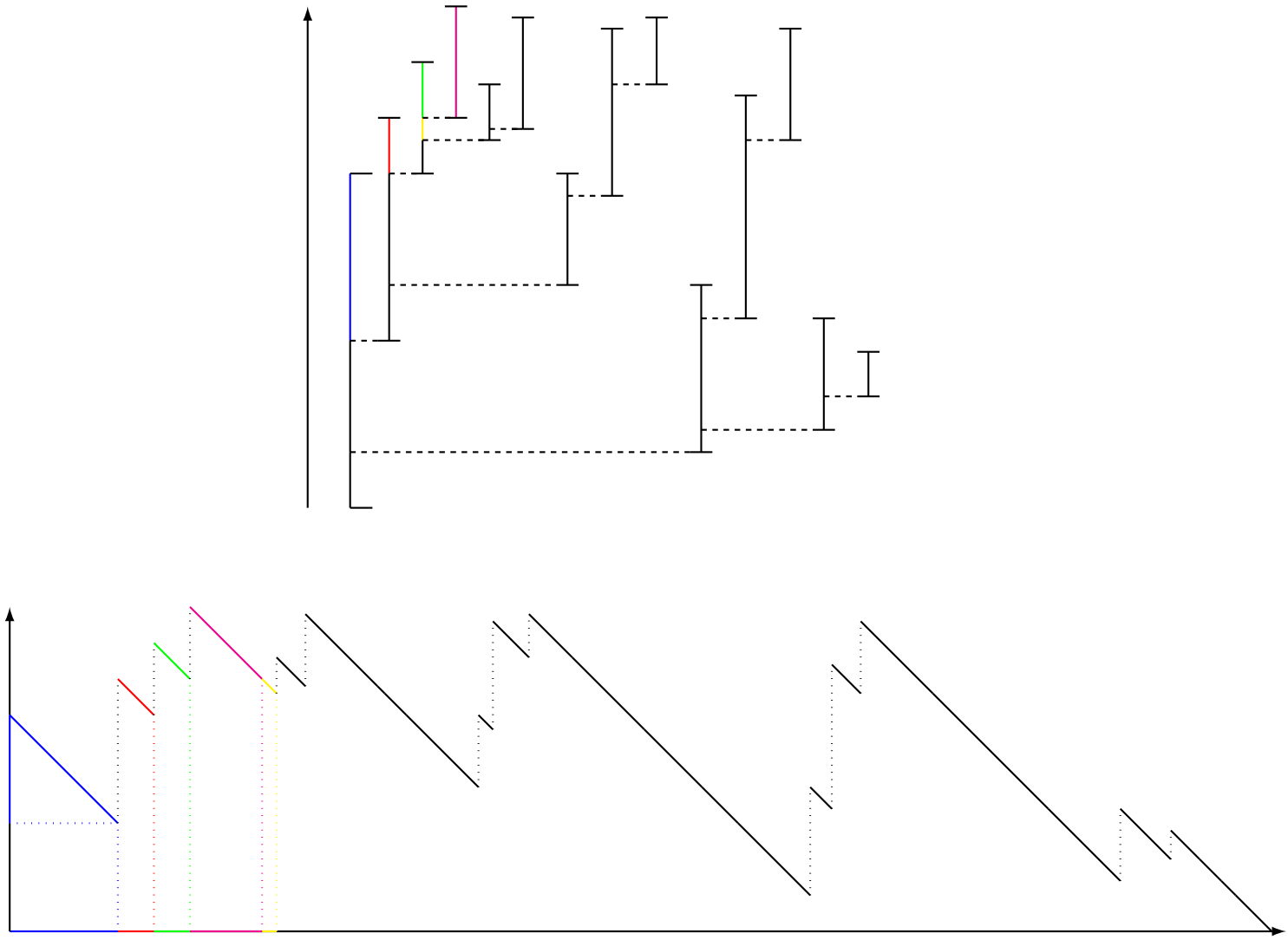


FIGURE 12 – Lien entre l'arbre et le processus de contour

3.4 Processus ponctuel de coalescence

3.4.1 Excursion du processus de contour et construction du PPC

Comme nous l'avons vu dans le paragraphe précédent, reconstruire la généalogie de la population revient à étudier la profondeur des excursions du processus de contour. Cette approche permet de résoudre le problème, cependant un splitting n'a pas toujours une longueur totale finie comme nous l'avons supposé plus haut.

Définition 3.16 (Splitting tree tronqué en t)**.**
Soit $\mathbb{T}$ un splitting tree et $t > 0$. Alors le splitting tree tronqué au temps t est défini par

$$C_t(\mathbb{T}) = \{(x, \sigma) \in \mathbb{T} \mid \sigma \leq t\}.$$

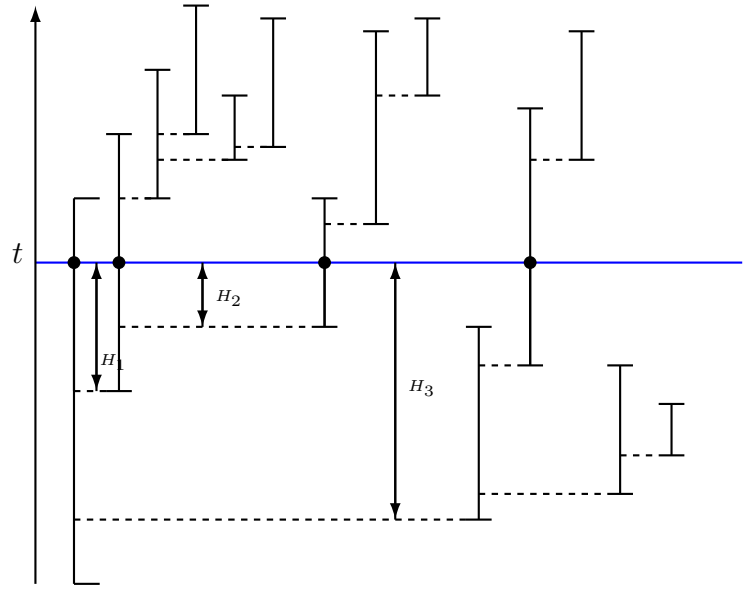


FIGURE 13 – Population vivante en t

Proposition 3.17 (Lambert (2010))**.**

$$Leb\left(C_t\left(\mathbb{T}\right)\right) < \infty \text{ } Leb\text{-}p.p. \text{ } p.s.$$

Nous pouvons alors passer au théorème principal qui va nous permettre de résoudre nos problèmes.

Théorème 3.18 (Lambert (2010))**.**
Le processus de contour associé à un Splitting Tree tronqué en t a la loi d'un processus de Lévy d'exposant de Laplace

$$\psi(x) = x - \int_{(0,\infty]} (1 - e^{-rx})\Lambda(dr),$$

tué en zéro et réfléchi en t.

De plus, le caractère fortement Markovien d'un tel processus de Lévy nous permet d'affirmer que la suite des profondeurs des excursions (correspondant aux temps de coalescence) forment une famille i.i.d. de variables aléatoires. La théorie des processus de Lévy sans saut négatif permet d'en caractériser la loi (Voir Bertoin [1]) :

$$\mathbb{P}(H > t) = \frac{1}{W(t)} \quad \forall t \geq 0,$$

où $W(t)$, appelée fonction d'échelle associée à l'exposant de Laplace ψ, est l'unique fonction continue strictement croissante sur $[0, \infty)$ telle que

$$\begin{cases} W(x) = 0 \ \forall x < 0, \\ \int_0^\infty dr \ e^{-xr} W(r) = \dfrac{1}{\psi(x)} \quad \forall x > \alpha, \end{cases}$$

où α est la plus grande racine de ψ (Voir figure 14).

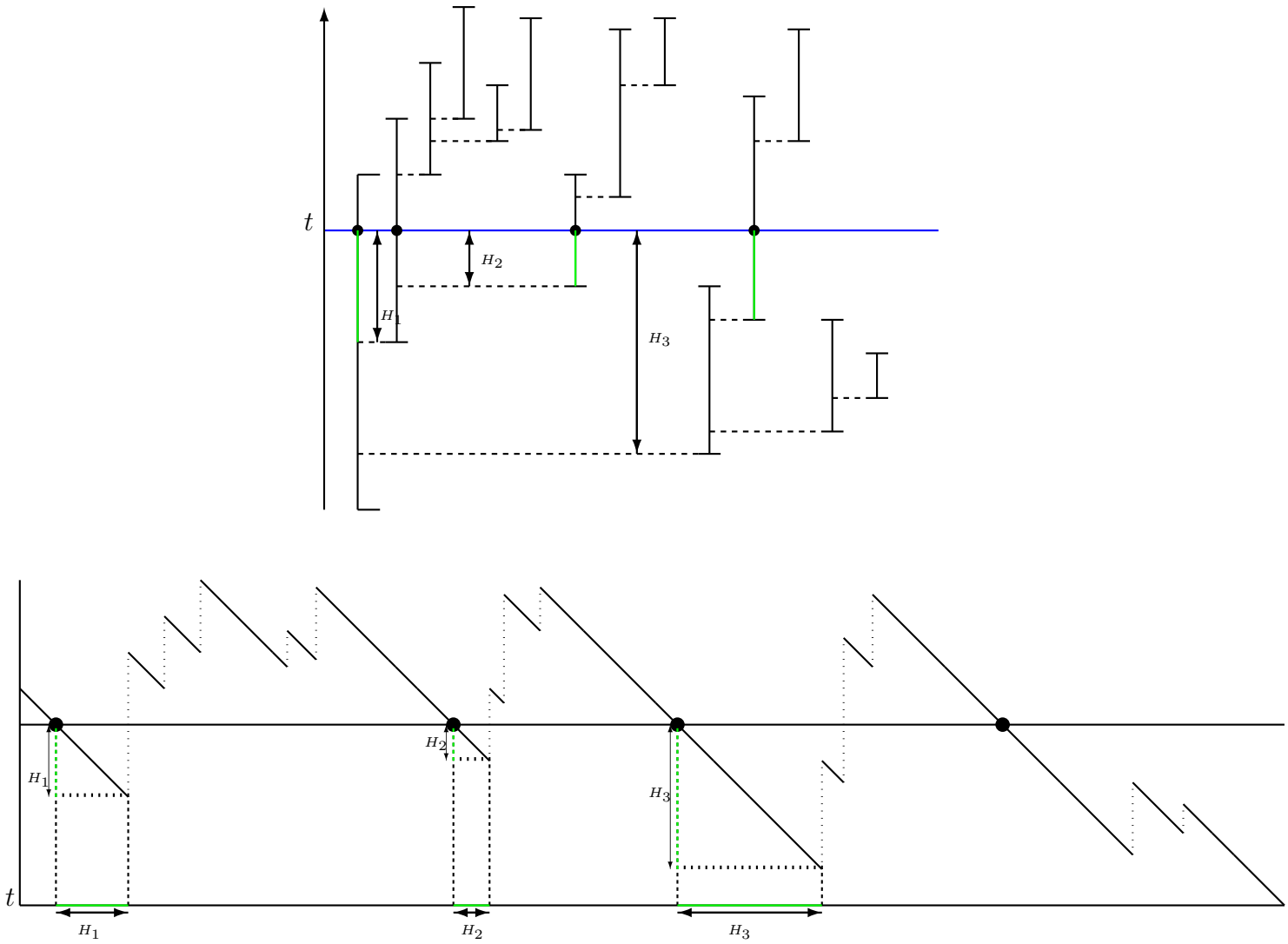


FIGURE 14 – Connexions entre le processus de contour l'arbre et les temps de coalescences.

Remarque 3.19. *Si on considère le processus de contour d'un splitting tree sur l'événement "le splitting tree est de mesure totale finie", il est possible de comprendre assez bien l'origine de la "nature Lévy" du processus. En effet, ce dernier décroit linéairement jusqu'à la rencontre d'une naissance : il saute alors d'une hauteur donnée par une loi constante, puis recommence à décroitre linéairement jusqu'à la prochaine naissance ou l'atteinte de* 0*. Ainsi, comme les naissances arrivent aux rythmes de processus de Poissons, les temps de saut sont distribués suivant des lois exponentielles et les hauteurs des sauts suivant la loi* $\Lambda(.)/b$ *.*

Dans ce cas, le processus de contour est un processus de Poisson composé et compensé de taux de sauts b *et de loi de saut* $\Lambda(.)/b$*.*

Remarque 3.20. *Dans le cas où* $\Lambda(\mathbb{T}) < \infty$*, le processus de contour tronqué en* t *s'obtient en éliminant toutes les parties du contour non tronqué au-dessus du niveau* t *(Voir figure 14).*

Ceci nous permet alors d'étudier directement la généalogie à partir de la suite des temps de coalescence, sans passer par le Splitting Tree : on appelle une telle suite un processus ponctuel de coalescence.

Définition 3.21 (Processus ponctuel de coalescence)**.**
Soit $W(t)$ la fonction d'échelle associée à un splitting tree. On appelle processus ponctuel de coalescence le processus P au temps t et associé à la loi $\mathbb{P}(H \in dx)$ défini par

$$P^{(t)} = (H_i)_{i \in [\![0, N_t - 1]\!]},$$

avec $H_0 = t$, $(H_i)_{i \geq 1}$ une suite i.i.d. de loi donnée par $\mathbb{P}(H > t) = \frac{1}{W(t)}$ et

$$N_t = \inf \{i \geq 1 \mid H_i \geq t\}.$$

Autrement dit, le PPC est une suite i.i.d. de variables aléatoires, de loi donnée par $\mathbb{P}(H > t) = \frac{1}{W(t)}$, tuée au premier élément dépassant t.

Graphiquement, on obtient l'arbre généalogique des individus vivants à t en traçant (dans l'ordre) les hauteurs (*branches*) H_i, puis en reliant les extrémités de chaque branche par un trait horizontal vers la gauche à la première branche plus grande (Voir figure 15). La branche H_i représente la lignée ancestrale du $(i+1)$-ième individu jusqu'au plus récent ancêtre commun entre cet individu et d'autres.

Revenons à l'exemple de la figure 13 pour voir ce que donne le PPC associé :

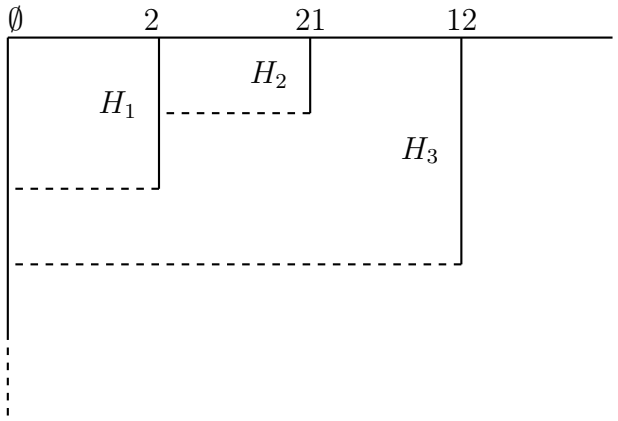


FIGURE 15 – PPC associé à l'exemple

Comme on le voit sur la figure 12, on retrouve la généalogie décrite précédemment :
- Dans un premier temps 21 et 2 coalescent au bout du temps H_2.
- Puis la lignée $\{21, 2\}$ coalesce avec l'individu ancestral $\emptyset$ au bout du temps H_1.
- Enfin toutes les lignées se rejoignent au bout du temps H_3.

Remarque 3.22.

- *Dans le cas où $\Lambda(du) = b\delta_\infty(du)$, le nombre d'individus vivant (indexé par t) est un processus de Yule de paramètre b et le splitting tree associé est un arbre de Yule.*

– *Si* $\Lambda(du) = bde^{-du}du$*, alors le nombre d'individus vivant est un processus de naissance et de mort linéaire.*

(Voir [5].)

3.4.2 Splitting tree et PPC avec mutation

Nous allons maintenant définir, comme cela a été fait pour le coalescent de Kingman, le PPC avec mutation. On choisit pour l'instant le modèle à une infinité d'allèles.

PPC avec mutation

Soit $P^{(t)}$ un processus ponctuel de coalescence arrêté en t. Alors conditionnellement à $P^{(t)}$, soit ν_t une mesure aléatoire de Poisson sur $\mathbb{R}_+ \times [\![0, N_t - 1]\!]$ d'intensité

$$N_t \theta ds \times \mathcal{U}_{[\![0,N_t-1]\!]},$$

avec $U_{[\![0,N_t-1]\!]}$ la mesure uniforme sur $[\![0, N_t - 1]\!]$.

On note aussi $(T_k, J_k)_{k\geq 0}$ l'ensemble des atomes de cette mesure ordonné par ordre croissant suivant les T_k.

Cette mesure régira les mutations apparaissant sur l'arbre. Pour le moment, on ne tient pas compte de la hauteur des H_i, donc certains atomes de la mesure de Poisson ne correspondront pas à une mutation. Les définitions suivantes ont pour but de préciser cela. Il reste donc encore à savoir quelles mutations concernent quels individus.

Définition 3.23 (Suite des ancêtres d'un individu)**.**
Soit $i < N_t$*, on note alors*

$$A(i) = \sup\{0 \leq j < i \mid H_j > H_i\}$$

et

$$A(0) = 0.$$

On note également pour tout $n \in \mathbb{N}$*,* $A^{(n)}(i)$ *l'itération n-ième de A pour la composition. Cette suite représente la suite des lignées ancêtres de i.*

Définition 3.24 (Mutation d'un individu)**.**
On note $M(i)$ *l'ensemble*

$$M(i) = \left\{T_k,\ k \in \mathbb{N} \mid \exists n \in \mathbb{N}\ J_k = A^{(n)}(i)\ \text{ et } H_{A^{(n-1)}(i)} < T_k < H_{J_k}\right\}.$$

Comme nous nous plaçons pour l'instant dans le modèle de mutation IMAM, seule la dernière mutation compte dans le type des individus.

Définition 3.25 (Type des individus sous IMAM)**.**
Soit $i > N_t$*, alors le type* $\mathcal{T}(i)$ *de l'individu i est défini par*

$$\mathcal{T}(i) = \begin{cases} \inf\{T_k \mid T_k \in M(i)\} & si\ M(i) \neq \emptyset, \\ 0 & sinon. \end{cases}$$

La figure 16 illustre le modèle à une infinité d'allèles.

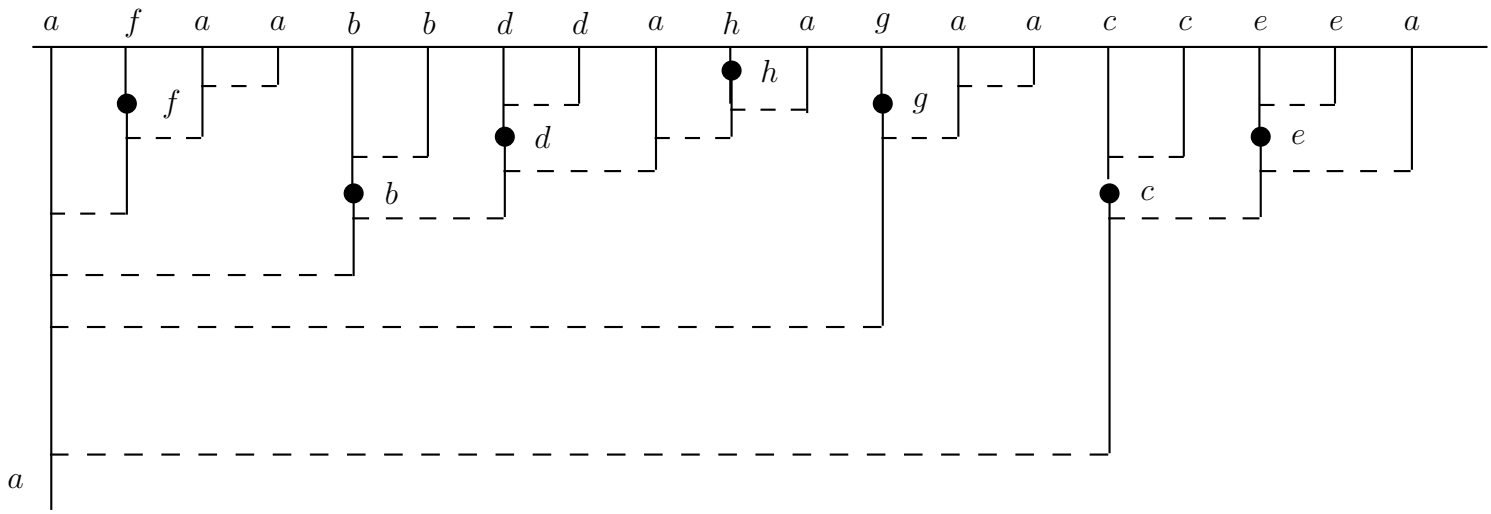


FIGURE 16 – PPC avec mutation

PPC ancestral

Nous allons voir que si on ne s'intéresse qu'aux individus ayant gardé le type ancestral (type 0), on retrouve à nouveau un PPC qui nous sera utile par la suite.

Définition 3.26 (PPC clonal)**.** *On définit par récurrence la suite* $(c_n)_{n\in[\![0,C]\!]}$ *(avec* $C = \#\{i \mid \mathcal{T}(i) = 0\}$*) des individus clonaux dans le PPC par (ordonnée par ordre croissant)*

$$\begin{cases} c_0 = \inf\{i \in [\![0, N_t - 1]\!] \mid \mathcal{T}(i) = 0\}, \\ c_n = \inf\{i \in [\![c_{n-1}, N_t - 1]\!] \mid \mathcal{T}(i) = 0\} \qquad si\ n \in]\!]0, C]\!]. \end{cases}$$

Alors la famille

$$\begin{cases} H_0^\star = t, \\ H_n^\star = \sup\{H_i \mid i \in]\!]c_{n-1}, c_n]\!]\} \qquad si\ n \in]\!]0, C[\![, \end{cases}$$

est appelée PPC clonal du PPC $P^{(t)}$ *.*

Remarque 3.27.
Une construction alternative de PPC clonal se fait en partant directement du splitting tree. Dans ce cas, les individus sont tués dès qu'une mutation apparait (Voir figure 17). On montre alors que l'arbre obtenu est également un splitting tree.

Théorème 3.28 (Lambert)**.** *Le PPC clonal est également un PPC de fonction caractéristique* W_θ *donnée par*

$$\begin{cases} \dfrac{d}{dx}W_\theta(x) = e^{-\theta x}\dfrac{d}{dx}W(x), \\ W_\theta(0) = 1. \end{cases}$$

La figure 18 illustre le PPC associé au processus de contour du sous-arbre clonal.

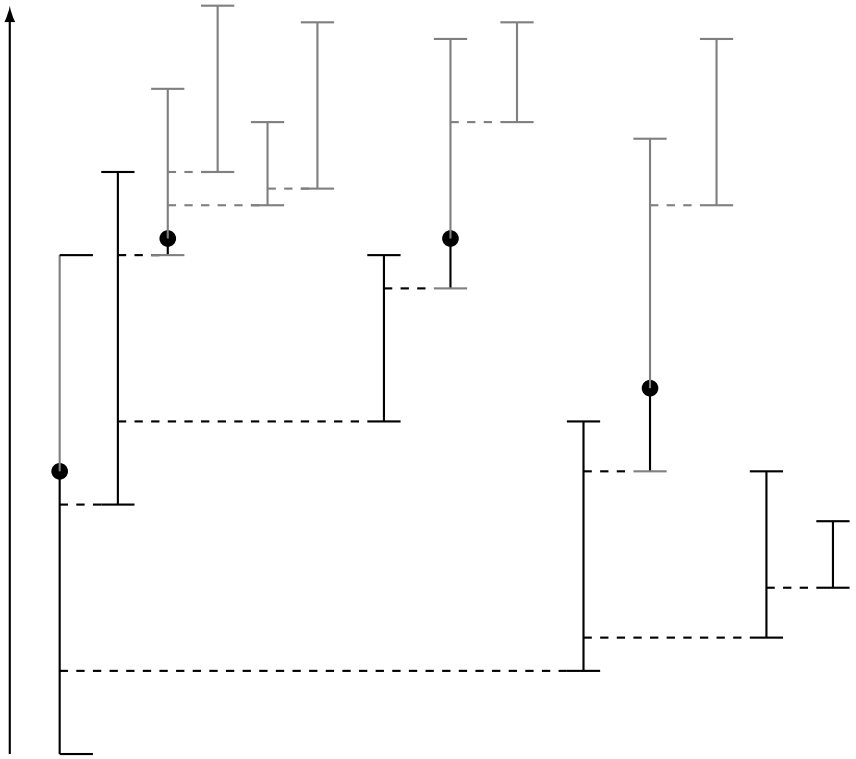


FIGURE 17 – Splitting tree clonal

3.4.3 Autour du spectre de fréquence

Remarque 3.29.
Attention : Dans toute la suite, et sauf mention contraire, on se place désormais sous la probabilité

$$\mathbb{P}\left(\bullet \mid N_t > 0\right).$$

Dans un souci d'alléger les notations, nous la noterons simplement $\mathbb{P}$ *comme avant. La notation de l'espérance sous cette nouvelle proba reste, elle aussi, inchangée.*

Le modèle introduit ci-dessus a déjà été étudié par Champagnat et Lambert dans [3], [5] et [4]. Nous allons citer uniquement les résultats de [3] et [5] qui nous serviront par la suite, mais d'abord nous introduisons certaines notations :

- $Z_0(t)$ le nombre d'individus clonaux dans la population vivante à la date t.
- $A(k,t)$ le nombre de types *non clonaux* représentés dans la population à la date t par k individus.
  (Spectre de fréquence)
- Z_t le nombre de type non clonaux dans la population $\left(Z(t) = \sum_{k\geq 1} A(k,t)\right)$.

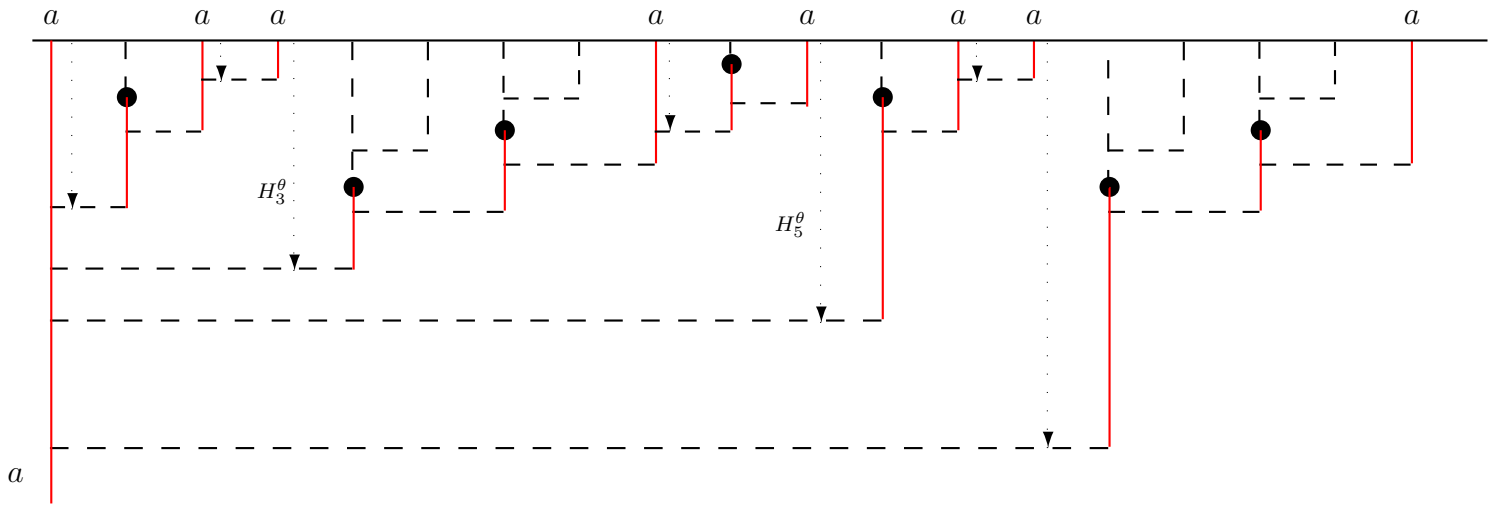


FIGURE 18 – PPC clonal

On a dans [3],

Théorème 3.30 ([3])**.**
Pour tout $k \in \mathbb{N}^*$

$$\mathbb{E}A(k,t) = W(t) \int_0^t ds\ \theta e^{-\theta s} \frac{1}{W_\theta(s)^2} \left(1 - \frac{1}{W_\theta(s)}\right)^{k-1}$$

et

$$\mathbb{P}(Z_0(t) = k) = W(t) \frac{e^{-\theta t}}{W_\theta(t)^2} \left(1 - \frac{1}{W_\theta(t)}\right)^{k-1}.$$

En particulier

$$\mathbb{E}(Z_t) = W(t) \int_0^t ds\ \frac{\theta e^{-\theta s}}{W_\theta(s)}.$$

Dans le cadre du stage, nous avons obtenu des résultats plus précis sur Z_t. Ainsi le but de la section à suivre est d'étudier plus précisément les moments factoriels de $(Z_t, t \geq 0)$. L'objectif est le résultat suivant, qui permet ensuite de calculer les moments factoriels par récurrence.

Proposition 3.31 (Moments factoriels de Z_t)**.**
On a pour tout $n \geq 2$,

$$\mathbb{E}\left[(Z_t)_{(n)}\right] = n! \int_0^t da\ \theta \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} \sum_{m=1}^{k} \left(1 - \frac{W(a)}{W(t)}\right)^{k-1} \left(\frac{W(t)}{W(a)}\right)^k \mathbb{E}\left[\mathbb{1}_{Z_0(a)>0} \binom{Z_a}{n_m}\right] \prod_{\substack{l=1 \\ l\neq m}}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right]$$

$$+n! \int_0^t da\ \theta \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} (k+1) \left(1 - \frac{W(a)}{W(t)}\right)^k \left(\frac{W(t)}{W(a)}\right)^{k+1} \mathbb{P}(Z_0(a) > 0) \prod_{l=1}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right].$$

Proposition 3.32. *De plus les moments du type :* $v_k(t) := \mathbb{E}\left[\mathbb{1}_{Z_0(t)>0} \binom{Z_t}{k}\right]$ *vérifient*

$$\partial_t v_k(t) = (p(t) - \theta) v(t) + \sum_{l=k-1}^{k} \binom{k}{l} \mathbb{E}\left[\binom{Z_t}{l} \mathbb{1}_{Z_0(t)>0}\right] \theta$$

$$+ \sum_{l=1}^{k-1} \binom{k}{l} \mathbb{E}\left[\binom{Z_t}{l} \mathbb{1}_{Z_0(t)>0}\right] \mathbb{E}\left[\binom{Z_t}{k-l} \mathbb{1}_{Z_0(t)>0}\right] p(t)$$

et

$$v_0(t) = \mathbb{P}(Z_0(t) > 0).$$

De tel manière que tout est connu récursivement.

Avant de faire la preuve de ce résultat, nous allons faire l'ébauche du cas $n = 2$, afin d'en faire ressortir les idées. La procédure est la suivante (voir figure 19) :

– On parcourt l'arbre en démarrant du temps le plus ancien (qui est $-t$ en considérant 0 comme le présent).

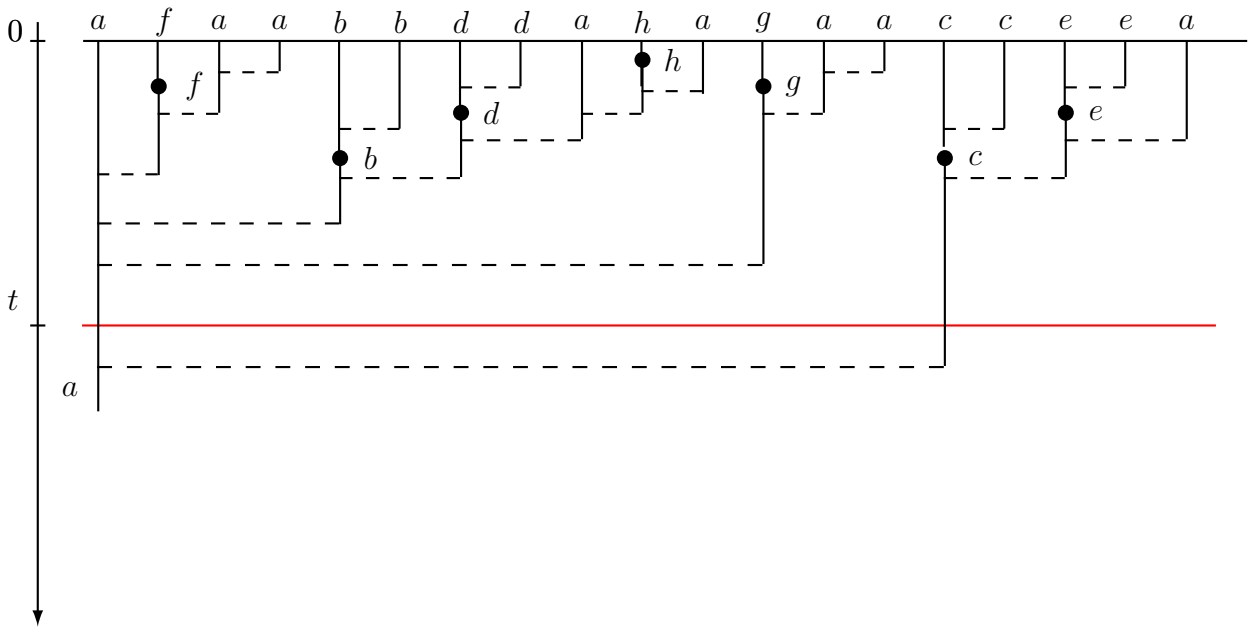


FIGURE 19 – Parcours de l'arbre à la recherche de mutations

- A chaque mutation rencontrée ayant un représentant vivant, on compte le nombre de mutations plus jeunes encore représentées en 0.
- On obtient à la fin $\binom{Z_t}{2}$ en faisant la somme de ces nombres.

Dans l'exemple de la figure 20, on suppose la mutation c trouvée à l'instant s. Alors, à chaque branche du PPC dépassant le niveau s correspond un sous-arbre qui est également un PPC de hauteur s.

De plus, tous les sous-arbres sont indépendants et si tant est qu'ils ne portent pas la mutation c, ils sont indépendants des individus portant le type c. Ainsi, pour compter le nombre de types non ancestraux d'apparition postérieure à c encore représentés à la date 0 :

- on compte le nombre de types non ancestraux encore représentés à la date 0 dans chaque sous-arbre ne portant pas c,

- et pour le sous-arbre portant c, on pense c comme le type ancestral et on compte le nombre de types non ancestraux sur l'événement "le type ancestral a encore des représentants vivants".

La preuve de la proposition repose donc sur le lemme suivant, dont le but est de donner une construction alternative des PPC par concaténation de "sous-PPC".

Construction alternative du processus ponctuel de coalescence par concaténation de sous-arbres.

L'objectif de cette section est de proposer une construction alternative des processus de coalescence afin de compter plus facilement les mutations en exploitant l'indépendance des différentes parties des arbres.

Soit $(A^i)_{i\geq 0}$ une famille *i.i.d.* de processus ponctuel de coalescence de loi $\mathbb{P}(H \in dx)$ au temps a. Ainsi, d'après la définition 3.21 on peut associer à chaque PPC A^i une suite i.i.d. $(H^i_j)_{j\geq 1}$ de loi

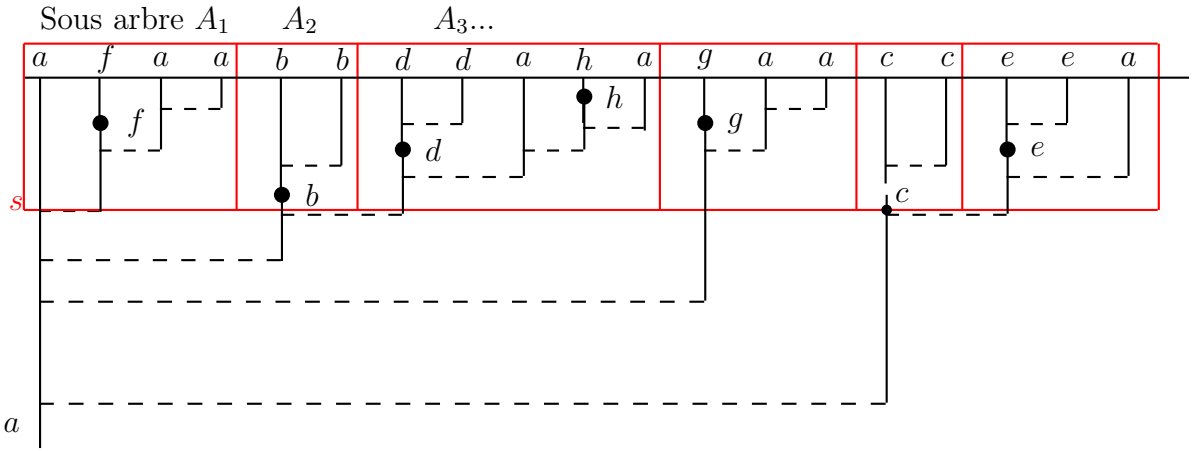


FIGURE 20 – Découpage de l'arbre en autant de sous-arbres

$\mathbb{P}(H \in dx)$. On définit pour tout i

$$N_a^i = \inf\left\{j \geq 0 | H_j^i \geq a\right\}.$$

Soit également $(H_j^*)_{j\geq 1}$, une suite *i.i.d.* de loi $\mathbb{P}(H \in dx | H > a)$. Alors l'arbre construit par la concaténation des parties vivantes de chaque arbre (arrêté en a) recollées par des H_i^* est un processus ponctuel de coalescence arrêté en t. Plus précisément, soit $t > a$.

Proposition 3.33. *Le vecteur aléatoire (de longueur aléatoire) X défini par*

$$X(k) = \sum_{i\geq 1} \left(A_a^i(k - S_{i-1})\mathbb{1}_{k\in]\!]S_{i-1}, S_i[\![} + \mathbb{1}_{S_i=k} H_i^*\right) \qquad si\ k < S_{J_t},$$

où

$$\begin{cases} S_i = \sum_{l=0}^{i} N_a^k, \\ S_{-1} = 0, \\ J_t = \inf\left\{j \geq 0 | H_j^\star \geq t\right\}, \end{cases}$$

est un PPC de loi $\mathbb{P}(H \in dx)$ au temps t.

Démonstration.
Pour montrer que le processus X est un processus ponctuel de coalescence, il suffit de montrer que la suite de variables aléatoires $(X(k))_{k\geq 0}$ est une suite *i.i.d.* Pour l'indépendance, il est facile de voir que pour toutes fonctions f, g continue bornée

$$\mathbb{E}\left[f\left(X\left(k\right)\right) g\left(X\left(k+1\right)\right)\right] = \mathbb{E}\left[f\left(X\left(k\right)\right)\right] \mathbb{E}\left[g\left(X\left(k+1\right)\right)\right]$$

On conclut alors par récurrence.

Soit $h \geq 0$ et $k \geq 0$
On a si $h \leq a$:

$$\begin{aligned}\mathbb{P}\left(X(k) \leq h \mid k \in]\!]S_{i-1}, S_i[\![\right) &= \mathbb{P}\left(A_a^i(k - S_i) \leq s | k - S_i < N_a^i\right) \\ &= \frac{\mathbb{P}\left(A_a^i\left(k - S_{i-1}\right) \leq h, k - S_{i-1} \leq k \mid S_{i-1}\right)}{\mathbb{P}\left(k - S_{i-1} \leq N^i\left(a\right) \mid S_{i-1}\right)} \\ &= \frac{\mathbb{P}\left(H \leq a\right)^{k - S_{i-1} - 1} \mathbb{P}\left(H \leq h\right)}{\mathbb{P}\left(H \leq a\right)^{k - S_{i-1}}} = \mathbb{P}\left(H \leq h \mid H \leq a\right).\end{aligned}$$

La formule reste cependant vrai si $h > a$. D'autre part, on a pour tout i et $h > a$

$$\mathbb{P}\left(X_t(k) \leq s | S_i = k\right) = \mathbb{P}\left(H_k^* \leq s | S_i = k\right) = \mathbb{P}\left(H_k^* \leq s\right) \quad \text{par indépendance de H}^* \text{ et des } S_I,$$

et qui reste la encore vrai si $h \leq a$. Alors

$$\begin{aligned}&\mathbb{P}\left(\sum_{i \geq 0} A_a^i(k - S_{i-1}) \mathbb{1}_{k \in]\!]S_{i-1}, S_i[\![} + \mathbb{1}_{S_i = k} H_i^* \leq h\right) \\ &\qquad = \sum_{i \geq 0} \mathbb{P}\left(A_a^i(k - S_{i-1}) \leq h, k \in]\!]S_{i-1}, S_i[\![\right) + \sum_{i \geq 0} \mathbb{P}\left(S_i = i, H_i^* \leq h\right) \\ &\qquad = \sum_{i \geq 0} \mathbb{P}\left(X(k) \leq h, k \in]\!]S_{i-1}, S_i[\![\right) \\ &\qquad\qquad + \mathbb{P}\left(H_i^* \leq h\right)\left(1 - \mathbb{P}\left(\bigcup_{i \geq 1} \{k \in]\!]S_{i-1}, S_i[\![\}\right)\right) \\ &\qquad = \sum_{i \geq 0} \mathbb{P}\left(k \in]\!]S_{i-1}, S_i[\![\right) \mathbb{P}\left(H \leq h | H < a\right) \\ &\qquad\qquad + \mathbb{P}\left(H_i^* \leq h\right)\left(1 - \mathbb{P}\left(\bigcup_{i \geq 1} \{k \in]\!]S_{i-1}, S_i[\![\}\right)\right) \\ &\qquad = \mathbb{P}\left(\bigcup_{i \geq 1} \{k \in]\!]S_{i-1}, S_i[\![\}\right) \mathbb{P}\left(H \leq h | H < a\right) \\ &\qquad\qquad + \mathbb{P}\left(H \leq h \mid H > a\right)\left(1 - \mathbb{P}\left(\bigcup_{i \geq 1} \{k \in]\!]S_{i-1}, S_i[\![\}\right)\right).\end{aligned}$$

Il reste juste à montrer que

$$\mathbb{P}\left(\bigcup_{i \geq 1} \{k \in]\!]S_{i-1}, S_i[\![\}\right) = \mathbb{P}\left(H < a\right),$$

pour achever la preuve. En effet, soit $\gamma \in (0,1)$.

$$\begin{aligned}\sum_{k\geq 0}\gamma^k \mathbb{P}\left(\exists i\ S_i = k\right) =& \mathbb{E}\left[\sum_{k,i\geq 0}\gamma^{S_i}\mathbb{1}_{k=S_i}\right]\\ =& \mathbb{E}\left[\sum_{i\geq 0}\gamma^{S_i}\right] = \sum_{i\geq 0}\mathbb{E}\left[\gamma^{N_a^1}\right]^i\\ =& \frac{1}{1-\mathbb{E}\left[\gamma^{N_a^1}\right]} = \frac{1-\mathbb{P}\left(H<a\right)\gamma}{1-\gamma}\\ =& \sum_{k\geq 0}\mathbb{P}\left(H\geq a\right)\gamma^k.\end{aligned}$$

□

Etude du moment factoriel d'ordre 2

Muni de ce résultat, nous allons compter les allèles de la manière suivante : pour chaque mutation rencontrée sur l'arbre admettant un représentant vivant à la date 0, nous allons compter le nombre de mutations plus jeunes, se trouvant dans un sous-arbre particulier et admettant également un représentant vivant à la date 0.

On note pour cela

- $\forall i \geq 0, \quad C_i\left(a, da\right)$ ="Il existe une mutation d'âge appartenant à $[a, a+da)$ sur la branche i admettant un représentant vivant à la date 0",
- $S_j^{(1)}\left(a\right)$= #{Mutations dans le sous-arbre j admettant au moins un représentant vivant à la date 0}.

Alors on a

$$\binom{Z_t}{2} = \int_0^t \sum_{i,j\geq 0}\mathbb{1}_{C_i(a,da)}\ S_j^{(1)}\left(a\right).$$

En passant à l'espérance et en conditionnant par $i \vee j < J$, on obtient

$$\mathbb{E}\binom{Z_t}{2} = \int_0^t \sum_{i,j\geq 1}\mathbb{P}\left(i\vee j<J\right)\mathbb{E}\left(\mathbb{1}_{C_i(a,da)}\ S_j^{(1)}\left(a\right)\Big| i\vee j<J\right).$$

Or, conditionnellement à être en vie (<J) les sous-arbres sont indépendants, on obtient alors

$$\begin{aligned}\mathbb{E}\binom{Z_t}{2} =& \int_0^t \sum_{\substack{i,j\geq 1\\ i\neq j}}\mathbb{P}\left(i\vee j<J\right)\mathbb{P}\left(C_i\left(a,da\right)\Big| i<J\right)\ \mathbb{E}\left(S_j^{(1)}\left(a\right)\Big| j<J\right)\\ &+\int_0^t \sum_{i\geq 1}\mathbb{P}\left(i<J\right)\mathbb{E}\left(\mathbb{1}_{C_i(a,da)}\ S_i^{(1)}\left(a\right)\Big| i<J\right).\end{aligned}$$

Or

$$\mathbb{P}\left(C_i\left(a,da\right)\mid i<J\right) = \mathbb{P}\left(Z_0\left(a\right)>0\right)$$

et

$$\mathbb{E}\left(S_j^{(1)}\left(a\right)\Big| j<J\right) = \mathbb{E}\left[Z_a\right],$$

d'où

$$\begin{aligned}
\mathbb{E}\binom{Z_t}{2} &= \int_0^t da\ 2\theta \sum_{j\geq 2}\sum_{i=1}^{j}\left(1-\frac{W(a)}{W(t)}\right)^{j-1}\mathbb{P}(Z_0(a)>0)\,\mathbb{E}(Z_a) \\
&\qquad + \int_0^t da\ \theta \sum_{i\geq 1}\left(1-\frac{W(a)}{W(t)}\right)^{i-1}\mathbb{E}\left(\mathbb{1}_{Z_0(a)>0}Z_a\right) \\
&= \int_0^t da\ 2\theta\left(1-\frac{W(a)}{W(t)}\right)\sum_{j\geq 1} j\ \left(1-\frac{W(a)}{W(t)}\right)^{j-1}\mathbb{P}(Z_0(a)>0)\,\mathbb{E}(Z_a) \\
&\qquad + \int_0^t da\ \theta \sum_{i\geq 1}\left(1-\frac{W(a)}{W(t)}\right)^{i-1}\mathbb{E}\left(\mathbb{1}_{Z_0(a)>0}Z_a\right) \\
&= \int_0^t da\ 2\theta\left(1-\frac{W(a)}{W(t)}\right)\frac{W(t)^2}{W(a)^2}\mathbb{P}(Z_0(a)>0)\,\mathbb{E}(Z_a) \\
&\qquad + \int_0^t da\ \theta\frac{W(t)}{W(a)}\mathbb{E}\left(\mathbb{1}_{Z_0(a)>0}Z_a\right).
\end{aligned}$$

On obtient donc l'expression désirée :

$$\begin{aligned}
\mathbb{E}(Z_t(Z_t-1)) &= 4\theta\int_0^t da\ \left(1-\frac{W(a)}{W(t)}\right)\frac{W(t)^2}{W(a)^2}\mathbb{P}(Z_0(a)>0)\,\mathbb{E}(Z_a) \\
&\qquad + 2\theta\int_0^t da\ \frac{W(t)}{W(a)}\mathbb{E}\left(\mathbb{1}_{Z_0(a)>0}Z_a\right).
\end{aligned}$$

Etude du moment factoriel d'ordre n

Nous allons ici itérer la méthode utilisée pour le moment factoriel d'ordre 2. Nous comptons de la manière suivante :

Pour chaque mutation d'âge compris entre a et $a+da$ représenté à la date 0, on compte
- le nombre de $(n-1)$-uplets de mutations représentées par des individus vivants dans chaque sous-arbre,
- le nombre de $(n-1)$-uplets de mutations représentées par des individus vivants dont $n-2$ éléments appartiennent à un sous-arbre donné et le dernier à un autre sous-arbre,
- etc....

A cette fin, on note
- $\forall i \geq 0, \quad C_i(a,da)$ ="Il existe une mutation d'âge appartenant a (a,a+da) sur la branche i admettant un représentant vivant",

- $\forall n \geq 1,\ \forall j \geq 1, \quad S_j^{(n)}(a) = \#\{n$-uplets de mutation dans le sous-arbre j admettant au moins un représentant vivant à la date $0\}$.

Introduisons tout d'abord les notations dont nous aurons besoin. On note $\forall k \in \mathbb{N}$,

$$I_k = \left\{ x \in \bigcup_{l \geq 1} (\mathbb{N}^\star)^l \Big| \sum x_i = k \right\}.$$

Et $\forall x \in I_k$, $|x|$ le nombre d'éléments distincts dans x notés $\bar{x}_1, .., \bar{x}_{|x|}$ (ordre arbitraire) et $n_1, \ldots, n_{|x|}$ leur multiplicité respective. On note également

$$||x|| = \sum_{i=1}^{d} n_i$$

et

$$\mathcal{S}^x = \left\{ \sigma \in \mathcal{S}_{|x|} \big| \sigma(x) \neq x \right\}.$$

Ainsi

$$|\mathcal{S}^x| = \frac{||x||!}{n_1!..n_{|x|}!}.$$

Ainsi $\binom{Z_t}{n}$ s'écrit

$$\binom{Z_t}{n} = \int_0^t \sum_{i \geq 1} \mathbb{1}_{C_i(a,da)} \sum_{x \in I_{n-1}} \frac{1}{|S^x|} \frac{1}{n_1! \ldots n_{|x|}!} \sum_{\substack{j_1, \ldots, j_{|x|} \geq 1 \\ \forall i \neq j \;\; j_i \neq j_l}} \prod_{m=1}^{|x|} S_{j_m}^{(x_m)}.$$

Nous allons faire quelques remarques sur cette expression afin de la rendre plus intelligible :

$$\binom{Z_t}{n} = \int_0^t \sum_{i \geq 1} \mathbb{1}_{C_i(a,da)} \sum_{\bar{x} \in I_{n-1}} \frac{1}{|\mathcal{S}^{\bar{x}}|} \frac{1}{n_1! \ldots n_{|x|}!} \sum_{\substack{j_1, \ldots, j_{|x|} \geq 1 \\ \forall i \neq j \;\; j_i \neq j_l}} \prod_{m=1}^{|x|} S_{j_m}^{(x_l)}.$$

- La partie rouge correspond à la recherche d'une première mutation à partir de laquelle le PPC sera découpé en sous -arbres : c'est le premier élément de notre n-uplet de mutations. Il s'agit d'énumérer tous les $(n-1)$-uplets possibles complétants le n-uplet

- La partie bleu correspond aux différentes manières de partager le $n-1$-uplet entre les sous-arbres. Par exemple, cette somme comprend la recherche du nombre de $n-1$-uplets dans chaque sous-arbre, ou encore le nombre de $n-2$-uplets dans chaque sous-arbre multiplié par le nombre de mutation unique dans les autres sous-arbres (ce qui correspond à une autre manière de construire un $n-1$-uplet), etc...

- Le terme en jaune est un terme de renormalisation. En effet, la somme bleu contient des termes doublons : si $x \in I_{n-1}$ alors pour toute permutation distinguable $\sigma \in \mathcal{S}^x$, $\sigma(x) \in I_{n-1}$, mais dans le compte total, ces termes ne devraient être comptés qu'une fois, c'est pourquoi on renormalise par le nombre de termes.

- Une fois le découpage du $(n-1)$-uplet choisi, il reste à choisir quels sous-arbres participent à quelles hauteurs à "l'uplet", puis de comptabiliser le nombre de $(n-1)$-uplets possibles correspondant à notre "découpage" : la somme violette correspond à la façon de répartir la charge dans chaque sous arbre pour chaque élément dans x, on en fait ensuite le produit.

– Le terme vert est également un terme de renormalisation : en effet lorsqu'il y a plusieurs termes identiques dans x il faut éviter de compter plusieurs fois certains sous-arbres.

On peut cependant aisément simplifier la formule précédente

$$\binom{Z_t}{n} = \int_0^t \sum_{i\geq 1} \mathbb{1}_{C_i(a,da)} \sum_{x\in I_{n-1}} \frac{1}{||x||!} \sum_{\substack{j_1,\dots,j_{|x|}\geq 1 \\ \forall i\neq j \ j_i\neq j_l}} \prod_{m=1}^{|x|} S_{j_m}^{(x_m)}.$$

On obtient finalement l'expression suivante :

$$\binom{Z_t}{n} = \int_0^t \sum_{i\geq 1} \mathbb{1}_{C_i(a,da)} \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} \frac{1}{k!} \sum_{\substack{j_{1:k}\geq 0 \\ \forall m\neq l \ j_l\neq j_m}} \prod_{l=1}^{k} S_{j_l}^{(n_l)}.$$

En passant à l'espérance, on obtient

$$\begin{aligned}
\mathbb{E}\left[\binom{Z_t}{n}\right] &= \mathbb{E}\left[\int_0^t \sum_{i\geq 1}\sum_{k=1}^{n-1} \frac{1}{k!} \sum_{n_1+..+n_k=n-1} \sum_{\substack{j_{1:k}\geq 0 \\ \forall m\neq l \ j_l\neq j_m}} \prod_{l=1}^{k} \mathbb{1}_{C_i(a,da)} S_{j_l}^{(n_l)}\right] \\
&= \int_0^t \sum_{k=1}^{n-1} \frac{1}{k!} \sum_{n_1+..+n_k=n-1} \sum_{i\geq 1} \sum_{\substack{j_{1:k}\geq 0 \\ \forall m\neq l \ j_l\neq j_m}} \mathbb{P}\left(\vee_l j_l \vee i < J\right) \mathbb{E}\left[\mathbb{1}_{C_i(a,da)} \prod_{l=1}^{k} S_{j_l}^{(n_l)} \middle| \vee_l j_l \vee i < J\right] \\
&= \int_0^t \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} \frac{1}{k!} \sum_{m=1}^{k} \sum_{\substack{j_{1:k}\geq 0 \\ \forall m\neq l \ j_l\neq j_m}} \mathbb{P}\left(\vee_l j_l < J\right) \mathbb{E}\left[\mathbb{1}_{C_{j_m}(a,da)} S_{j_m}^{n_m} \prod_{\substack{l=1 \\ l\neq m}}^{k} S_{j_l}^{(n_l)} \middle| \vee_l j_l < J\right] \\
&\quad + \int_0^t \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} \frac{1}{k!} \sum_{\substack{i,j_{1:k}\geq 0 \\ \forall m\neq l \ j_l\neq j_m, \forall m \ i\neq j_m}} \mathbb{P}\left(\vee_l j_l \vee i < J\right) \mathbb{E}\left[\mathbb{1}_{C_i(a,da)} \prod_{l=1}^{k} S_{j_l}^{(n_l)} \middle| \vee_l j_l \vee i < J\right].
\end{aligned}$$

En utilisant à nouveau l'indépendance des sous-arbres conditionnellement à l'événement $\{\vee_l j_l \vee i < J\}$, il suit que

$$\begin{aligned}
\mathbb{E}\left[\binom{Z_t}{n}\right] &= \int_0^t da\ \theta \sum_{k=1}^{n-1} \frac{1}{k!} \sum_{n_1+..+n_k=n-1} \sum_{m=1}^{k} \sum_{\substack{j_{1:k}\geq 0 \\ \forall m\neq l \ j_l\neq j_m}} \mathbb{P}\left(\vee_l j_l < J\right) \mathbb{E}\left[\mathbb{1}_{Z_0(a)>0} \binom{Z_a}{n_m}\right] \prod_{\substack{l=1 \\ l\neq m}}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right] \\
&\quad + \int_0^t da\ \theta \sum_{k=1}^{n-1} \frac{1}{k!} \sum_{n_1+..+n_k=n-1} \sum_{\substack{i,j_{1:k}\geq 0 \\ \forall m\neq l \ j_l\neq j_m, \forall m \ i\neq j_m}} \mathbb{P}\left(\vee_l j_l \vee i < J\right) \mathbb{P}\left(Z_0(a) > 0\right) \prod_{l=1}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right] \\
&= \int_0^t da\ \theta \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} \sum_{0\leq j_1<..<j_k} \left(1-\frac{W(a)}{W(t)}\right)^{j_k} \sum_{m=1}^{k} \mathbb{E}\left[\mathbb{1}_{Z_0(a)>0} \binom{Z_a}{n_m}\right] \prod_{\substack{l=1 \\ l\neq m}}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right] \\
&\quad + \int_0^t da\ \theta \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} (k+1) \sum_{0\leq j_1<..<j_k<i} \left(1-\frac{W(a)}{W(t)}\right)^{i} \mathbb{P}\left(Z_0(a) > 0\right) \prod_{l=1}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right].
\end{aligned}$$

Puis en utilisant le fait que pour tout $0 \leq p < 1$, $\sum_{0\leq j_1<..<j_n} p^{j_n} = \frac{p^{n-1}}{(1-p)^n}$, on obtient finalement

$$\mathbb{E}\left[(Z_t)_{(n)}\right] = n! \int_0^t da\ \theta \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} \sum_{m=1}^{k} \left(1 - \frac{W(a)}{W(t)}\right)^{k-1} \left(\frac{W(t)}{W(a)}\right)^{k} \mathbb{E}\left[\mathbb{1}_{Z_0(a)>0} \binom{Z_a}{n_m}\right] \prod_{\substack{l=1 \\ l\neq m}}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right]$$

$$+n! \int_0^t da\ \theta \sum_{k=1}^{n-1} \sum_{n_1+..+n_k=n-1} (k+1) \left(1 - \frac{W(a)}{W(t)}\right)^{k} \left(\frac{W(t)}{W(a)}\right)^{k+1} \mathbb{P}(Z_0(a) > 0) \prod_{l=1}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right].$$

Ce qui achève la preuve.

Lemme 3.34.
Pour tout $0 < p < 1$, $\sum_{0\leq j_1<..<j_n} p^{j_n} = \frac{p^{n-1}}{(1-p)^n}$.

Démonstration.

$$\begin{aligned}
\sum_{0\leq j_1<\cdots<j_n} p^{j_n} &= \sum_{j_n=n-1} p^{j_n} \sum_{0\leq j_1<\cdots<j_{n-1}\leq j_n-1} 1 \\
&= \sum_{j_n=n-1} p^{j_n}\ (n-1)! \sum_{j_{1:(n-1)}=0}^{j_n-1} 1 \\
&= (n-1)! p^{n-1} \sum_{j_n=n-1} p^{j_n-n+1}\ (j_n)_{(n-1)} \\
&= (n-1)! p^{n-1}\ \frac{d^{n-1}}{dx}\left(\frac{1}{1-x}\right) = \frac{p^{n-1}}{(1-p)^n}
\end{aligned}$$

□

Il reste à prouver la dernière partie de la proposition.

Démonstration. Proposition 3.32

Remarquons qu'à partir d'une suite $H_0 = \infty$, et $(H_i)_{i\geq 1}$ i.i.d. on peut coupler tous les PPC à des instants différents. Ce couplage ne correspond pas à celui des PPC associés à un même splitting tree, mais il permet de faire des calculs de loi marginale (à un temps fixé).

Si on étudie la dynamique du processus $(Z_t, Z_0(t))_t$, en considérant les PPC construits à partir de la même suite $(H_i)_{i\geq 1}$ mais aux temps t et $t + dt$, on observe

- $(Z_{t+dt}, Z_0(t+dt)) = (Z_t, Z_0(t))_t$ avec probabilité $1 - (p(t) + \theta)\, dt$,
- $(Z_{t+dt}, Z_0(t+dt)) = \left(Z_t + \mathbb{1}_{Z_0(t)>0}, 0\right)_t$ avec probabilité θdt,
- $(Z_{t+dt}, Z_0(t+dt)) = (Z_t + Z'_t, Z_0(t) + Z'_0(t))_t$ avec probabilité $p(t)\, dt$, avec $p(t)\, dt = \mathbb{P}(H \leq t + dt \mid H > t)$.

Ici, Z'_t et $Z'_0(t)$ sont construits à partir de la suite $(H_i)_{i\geq N_t}$.

Ainsi, Z'_0 est de même loi que Z_0 et indépendant de Z_0 et Z. D'autre part, Z' est de même loi que Z et indépendant de Z_0 et Z

Donc, si on note $L(t,u) = \mathbb{E}\left[\binom{Z_t}{k} u^{Z_0(t)}\right]$, on a

$$\begin{aligned}
L(t+dt,u) &= (1-(p(t)+\theta)\,dt)\,L(t,u) + \mathbb{E}\left[\binom{Z_t+\mathbb{1}_{Z_0(t)>0}}{k}\right]\theta dt \\
&\quad + p(t)\,dt\;\mathbb{E}\left[\binom{Z_t+Z'_t}{k} u^{Z_0(t)+Z'(0)_t}\right] \\
&= (1-(p(t)+\theta)\,dt)\,L(t,u) + \sum_{l=k-1}^{k}\binom{k}{l}\mathbb{E}\left[\binom{Z_t}{l}\mathbb{1}_{Z_0(t)>0}\right]\theta dt \\
&\quad + \sum_{l=0}^{k}\binom{k}{l}\mathbb{E}\left[\binom{Z_t}{l}u^{Z_0(t)}\right]\mathbb{E}\left[\binom{Z'_t}{k-l}u^{Z'_0(t)}\right]p(t)\,dt \\
&= ((1-\theta dt)+p(t)\,dt)\,L(t,u) + \sum_{l=k-1}^{k}\binom{k}{l}\mathbb{E}\left[\binom{Z_t}{l}\mathbb{1}_{Z_0(t)>0}\right]\theta dt \\
&\quad + \sum_{l=1}^{k-1}\binom{k}{l}\mathbb{E}\left[\binom{Z_t}{l}u^{Z_0(t)}\right]\mathbb{E}\left[\binom{Z_t}{k-l}u^{Z_0(t)}\right]p(t)\,dt.
\end{aligned}$$

Ainsi $L(u,t)$ vérifie

$$\begin{aligned}
\partial_t L(t,u) &= (p(t)-\theta)\,L(t,u) + \sum_{l=k-1}^{k}\binom{k}{l}\mathbb{E}\left[\binom{Z_t}{l}\mathbb{1}_{Z_0(t)>0}\right]\theta \\
&\quad + \sum_{l=1}^{k-1}\binom{k}{l}\mathbb{E}\left[\binom{Z_t}{l}u^{Z_0(t)}\right]\mathbb{E}\left[\binom{Z_t}{k-l}u^{Z_0(t)}\right]p(t).
\end{aligned}$$

En prenant $u=0$, on obtient finalement

$$\begin{aligned}
\partial_t\mathbb{E}\left[\binom{Z_t}{k}\mathbb{1}_{Z_0(t)>0}\right] &= (p(t)-\theta)\,\mathbb{E}\left[\binom{Z_t}{k}\mathbb{1}_{Z_0(t)>0}\right] + \sum_{l=k-1}^{k}\binom{k}{l}\mathbb{E}\left[\binom{Z_t}{l}\mathbb{1}_{Z_0(t)>0}\right]\theta \\
&\quad + \sum_{l=1}^{k-1}\binom{k}{l}\mathbb{E}\left[\binom{Z_t}{l}\mathbb{1}_{Z_0(t)>0(t)}\right]\mathbb{E}\left[\binom{Z_t}{k-l}\mathbb{1}_{Z_0(t)>0(t)}\right]p(t).
\end{aligned}$$

□

Corollaire 3.35.
Dans le cas surcritique $\left(\int_{(0,\infty]} r\Lambda(dr) > 1\right)$ *et tel que* $\alpha > \theta$, *on a*

$$\forall n \in \mathbb{N}^* \qquad \mathbb{E}\,Z(t)^n = n!\,W(t)^n\left(\int_0^t dx\;\frac{\theta e^{-\theta x}}{W_\theta(x)}\right)^n + o_\infty\left(W(t)^n\right).$$

Avant d'attaquer la preuve, nous avons besoin du lemme suivant :

Lemme 3.36 (Comportement asymptotique de $W(t)$ et $W_\theta(t)$)**.**
Dans le cas surcritique $\left(\int_{(0,\infty]} r\Lambda(dr) > 1\right)$, *on a*

$$W(t) \underset{\infty}{\sim} e^{\alpha t}.$$

D'autre part, si $\alpha > \theta$, on a également

$$W_\theta(t) \underset{\infty}{\sim} e^{(\alpha-\theta)t}.$$

Démonstration. On a vu que $W(t)$ satisfait la relation

$$\begin{cases} W(x) = 0 \ \forall x < 0, \\ \int_0^\infty dr\ e^{-xr} W(r) = \dfrac{1}{\psi(x)} \quad \forall x > \alpha, \end{cases}$$

où α est la plus grande racine de ψ et

$$\psi(x) = x - \int_{(0,\infty]} (1 - e^{-rx}) \Lambda(dr).$$

Alors le résultat s'obtient en appliquant le théorème Tauberiens pour la transformée de Laplace 5.13 de [8]. Puis la seconde partie du lemme s'obtient en passant par le théorème 3.2. □

Nous pouvons maintenant passer à la preuve du corollaire.

Démonstration. La preuve se fait par récurrence sur n.

Cas n=1

D'après le lemme 3.1 et $(\star)$, $t \to \frac{e^{-\theta t}}{W_\theta(t)}$ est intégrable sur $\mathbb{R}_+$. Il est alors clair que le cas $n = 1$ vérifie le corollaire.

Dans un souci de clarté, nous donnons la preuve du cas $n = 2$, même si elle est comprise dans le cas général.

Cas n=2

On a vu que

$$\begin{aligned} \mathbb{E}\left(Z_t\left(Z_t - 1\right)\right) &= 4\theta \int_0^t da\ \left(1 - \frac{W(a)}{W(t)}\right) \frac{W(t)^2}{W(a)^2} \mathbb{P}\left(Z_0(a) > 0\right) \mathbb{E}\left(Z_a\right) \\ &\quad + 2\theta \int_0^t da\ \frac{W(t)}{W(a)} \mathbb{E}\left(\mathbb{1}_{Z_0(a)>0} Z_a\right) \\ &= 4\theta W(t)^2 \int_0^t da\ \frac{1}{W(a)^2} \mathbb{P}\left(Z_0(a) > 0\right) \mathbb{E}\left(Z_a\right) - 4\theta W(t) \int_0^t da\ \frac{1}{W(a)} \mathbb{P}\left(Z_0(a) > 0\right) \mathbb{E}\left(Z_a\right) \\ &\quad + 2\theta W(t) \int_0^t da\ \frac{1}{W(a)} \mathbb{E}\left(\mathbb{1}_{Z_0(a)>0} Z_a\right). \end{aligned}$$

D'après le cas $n = 1$, il existe $C \in \mathbb{R}_+$ tel que $\mathbb{E}Z_t \sim C\ W(t)$, d'où

$$\frac{1}{W(a)} \mathbb{P}\left(Z_0(a) > 0\right) \mathbb{E}\left(Z_a\right) \sim C.$$

Or la fonction $a \to C$ n'est pas intégrale sur $[0, \infty)$, donc

$$-4\theta W(t) \int_0^t da \ \frac{1}{W(a)} \mathbb{P}(Z_0(a) > 0) \mathbb{E}(Z_a) \sim -4\theta t \ W(t) = o_\infty\left(W(t)^2\right).$$

Le cas du terme

$$2\theta W(t) \int_0^t da \ \frac{1}{W(a)} \mathbb{E}\left(\mathbb{1}_{Z_0(a)>0} Z_a\right)$$

se traite de la même manière, en utilisant la majoration $\mathbb{E}\left(\mathbb{1}_{Z_0(a)>0} Z_a\right) \leq \mathbb{E}(Z_a)$. On réduit alors l'expression à

$$\mathbb{E}(Z_t(Z_t - 1)) = 4\theta W(t)^2 \int_0^t da \ \frac{1}{W(a)^2} \mathbb{P}(Z_0(a) > 0) \mathbb{E}(Z_a) + o_\infty\left(W(t)^2\right).$$

Et finalement, comme

$$\theta \int_0^t da \ \frac{1}{W(a)^2} \mathbb{P}(Z_0(a) > 0) \mathbb{E}(Z_a) = \theta \int_0^t da \ \frac{e^{-\theta a}}{W_\theta(a)} \int_0^a ds \ \frac{\theta e^{-\theta s}}{W_\theta(s)} \tag{2}$$

$$= \frac{1}{2}\left(\int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)}\right)^2, \tag{3}$$

on obtient que

$$\mathbb{E}(Z_t(Z_t - 1)) \sim 2W(t)^2 \left(\int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)}\right)^2.$$

Cas général $n \in \mathbb{N}_*$

Supposons le corollaire vrai pour tout $k < n$, nous séparons la preuve en 2 lemmes :

Lemme 3.37.
$\forall k \in [\![1, n-2]\!]$, $\forall \bar{n} = (n_1, .., n_k) \in \mathbb{N}_*^k$, *tels que* $\sum_{i=1}^k n_i = n - 1$,

$$\int_0^t da \ \theta(k+1) \left(1 - \frac{W(a)}{W(t)}\right)^k \left(\frac{W(t)}{W(a)}\right)^{k+1} \mathbb{P}(Z_0(a) > 0) \prod_{l=1}^k \mathbb{E}\left[\binom{Z_a}{n_l}\right] = o_\infty\left(W(t)^n\right).$$

Démonstration. On a

$$\int_0^t da \ \left(1 - \frac{W(a)}{W(t)}\right)^k \left(\frac{W(t)}{W(a)}\right)^{k+1} \mathbb{P}(Z_0(a) > 0) \prod_{l=1}^k \mathbb{E}\left[\binom{Z_a}{n_l}\right] \tag{4}$$

$$= \sum_{j=0}^k \binom{k}{j} (-1)^j \int_0^t da \frac{W(t)^{k+1-j}}{W(a)^{k+1-j}} \mathbb{P}(Z_0(a) > 0) \prod_{l=1}^k \mathbb{E}\left[\binom{Z_a}{n_l}\right]. \tag{5}$$

Or, pour tout $j \in [\![0, k]\!]$, on a par l'hypothèse de récurrence

$$\frac{1}{W(a)^{k+1-j}} \mathbb{P}(Z_0(a) > 0) \prod_{l=1}^k \mathbb{E}\left[\binom{Z_a}{n_l}\right] \sim \frac{1}{W(a)^{k-n-j+2}}.$$

Et pour tout $j \in [\![0,k]\!]$, $k-n-j+1 \leq -1$ donc $a \to \dfrac{1}{W(a)^{k-n-j+1}}$ n'est jamais intégrale, d'où pour tout j

$$W(t)^{k+1-j} \int_0^t da \frac{1}{W(a)^{k+1-j}} \mathbb{P}(Z_0(a)>0) \prod_{l=1}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right] \sim W(t)^{n-1} = o_\infty(W(t)^n) \qquad si\ k < n-2,$$

$$W(t)^{k+1-j} \int_0^t da \frac{1}{W(a)^{k+1-j}} \mathbb{P}(Z_0(a)>0) \prod_{l=1}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right] \sim t\, W(t)^{n-1} = o_\infty(W(t)^n) \qquad si\ k = n-2.$$

Ce qui achève la preuve. □

Lemme 3.38.
$\forall k \in [\![1,n-1]\!]$, $\forall \bar{n} = (n_1,..,n_k) \in \mathbb{N}_*^k$, *tels que* $\sum_{i=1}^k n_i = n-1$, $\forall u \in [\![1,p]\!]$,

$$\int_0^t da\ \theta \left(1-\frac{W(a)}{W(t)}\right)^{k-1} \left(\frac{W(t)}{W(a)}\right)^{k} \mathbb{E}\left[\mathbb{1}_{Z_0(a)>0}\binom{Z_a}{n_u}\right] \prod_{\substack{l=1 \\ l\neq u}}^{k} \mathbb{E}\left[\binom{Z_a}{n_l}\right] = o_\infty(W(t)^n).$$

Démonstration. La preuve de ce lemme est très similaire à celle du précédent en utilisant la majoration

$$\mathbb{E}\left[\mathbb{1}_{Z_0(a)>0}\binom{Z_a}{n_u}\right] \leq \mathbb{E}\left[\binom{Z_a}{n_u}\right].$$

□

Retour à la preuve du Corrolaire

Dans l'expression de $\mathbb{E}\left[\binom{Z_t}{n}\right]$ seul un terme sort du cadre des deux lemmes précédents, c'est le cas $(1,..,1) \in \mathbb{N}^{n-1}$. On a alors

$$\begin{aligned}
\mathbb{E}\left[\binom{Z_t}{n}\right] &= \int_0^t da\ \theta\ n \left(1-\frac{W(a)}{W(t)}\right)^{n-1} \left(\frac{W(t)}{W(a)}\right)^{n} \mathbb{P}(Z_0(a)>0) \prod_{l=1}^{n-1} \mathbb{E}\left[\binom{Z_a}{1}\right] + o_\infty(W(t)^n) \\
&= \sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j \int_0^t da\ \theta\ n \left(\frac{W(t)}{W(a)}\right)^{n-j} \mathbb{P}(Z_0(a)>0)\, \mathbb{E}\left[\binom{Z_a}{1}\right]^{n-1} + o_\infty(W(t)^n) \\
&= \int_0^t da\ \theta\ n \left(\frac{W(t)}{W(a)}\right)^{n} \mathbb{P}(Z_0(a)>0)\, \mathbb{E}\left[\binom{Z_a}{1}\right]^{n-1} + o_\infty(W(t)^n),
\end{aligned}$$

où la dernière ligne a été obtenue en éliminant les autres termes de la somme par les même méthodes que précédemment. Finalement

$$\begin{aligned}\mathbb{E}\left[\binom{Z_t}{n}\right] &= nW(t)^n \int_0^t da\ \theta \frac{1}{W(a)^n} \mathbb{P}(Z_0(a) > 0)\, \mathbb{E}[Z_a]^{n-1} + o_\infty(W(t)^n) \\ &= nW(t)^n \int_0^t da\ \theta \frac{e^{-\theta a}}{W_\theta(a)} \left(\int_0^a ds\ \frac{\theta e^{-\theta s}}{W_\theta(s)}\right)^{n-1} + o_\infty(W(t)^n) \\ &= W(t)^n \left(\int_0^t da\ \frac{\theta e^{-\theta a}}{W_\theta(a)}\right)^n + o_\infty(W(t)^n).\end{aligned}$$

Ce qui achève la preuve. □

Proposition 3.39. *(Convergence du nombre d'allèles non clonaux renormalisés)*
Dans le cas surcritique et pour $\theta < \alpha$.
Soit $(Z_t, t \geq 0)$ le nombre d'allèles non clonaux dans le processus ponctuel de coalescence avec mutation tué en t, alors $\frac{Z_t}{W(t)}$ converge en loi, quand $t \to \infty$ vers une variable aléatoire de loi exponentielle de paramètre $\left(\int_0^\infty da \frac{\theta e^{-\theta a}}{W_\theta(a)}\right)^{-1}$, conditionnellement à la non-extinction.

3.5 Inférence sur les processus ponctuels de coalescence

Remarque 3.40. *A partir d'ici, nous ne nous plaçons plus sous $\mathbb{P}(. \mid N_t > 0)$.*

Nous nous replaçons comme dans la section 2.4 sous le modèle de mutation à une infinité de sites. Nous devons alors redéfinir le type des individus en reprenant les notations de la section 3.4.2.

Définition 3.41 (Type sous IMSM)**.**
Soit $i < N_t$, on note $\mathcal{T}(i)$ le type de i définit par

$$\mathcal{T}(i) = \sum_{\substack{l \geq 1 \\ T_l \in M(i)}} \delta_l.$$

On peut alors redéfinir la configuration de la population à l'instant présent.

Définition 3.42 (Configuration de la population)**.**
Soit un PPC avec mutation, on note

$$\mathcal{T}_t = (\mathcal{T}(1), ..., \mathcal{T}(N_t - 1)),$$

la configuration de type de la population.

Avant d'aller plus loin, nous allons nous placer dans un cas plus simple que le cas général. Ainsi, nous allons supposer que $\Lambda(du) = \lambda \delta_\infty(du)$ $(\lambda \in \mathbb{R}_+)$. Ceci correspond à un splitting tree coupé sous t qui est un arbre de Yule de taux λ. Dans ce cas là, les H_i suivent des lois exponentielles de paramètre λ.

Notations préliminaires

Commençons par introduire quelques notations pour simplifier les preuves à venir. Leur but est de redéfinir les ancêtres et les types afin de rendre explicite la dépendance en $(H_i)_{i\geq 0}$ et en $(T_i, J_i)_{i\geq 0}$.

On note $\forall u = (u_n)_{n\geq 0} \subset \mathbb{R}_+$,

$$\begin{array}{rccc} f^u & : & \mathbb{N} & \to & \mathbb{N} \\ & & n & \mapsto & \sup\{0 \leq j < n \mid u_j > u_n\}. \end{array}$$

De sorte que $\forall i \in [\![1, N_t - 1]\!]$, $A(i) = f^{(H_i)}(i)$.

En note d'autre part

$$\begin{array}{rccc} g^u & : & \mathbb{N} \times (\mathbb{R} \times \mathbb{N})^{\mathbb{N}} & \to & \Pi \\ & & \left(k, (v_n, w_n)_{n\geq 0}\right) & \mapsto & \sum_{\substack{l\geq 1 \\ v_l \in M}} \delta_l, \end{array}$$

avec

$$M = \left\{v_n \mid n \in \mathbb{N} \text{ et } \exists l \in \mathbb{N}\ w_n = A^{(l)}(k)\ \textit{ et } u_{A^{(l-1)}(k)} < v_l < u_{w_l}\right\}.$$

Là encore, de manière à ce que

$$\forall i \in [\![1, N_t - 1]\!] \quad \mathcal{T}(i) = g^{(H_i)}\left(i, (T_n, J_n)_{n\geq 0}\right).$$

Et note finalement, pour tout $n \in \mathbb{N}$,

$$\begin{array}{rccc} F_n^u & : & (\mathbb{R} \times \mathbb{N})^{\mathbb{N}} & \to & \Pi^n \\ & & (v_n, w_n)_{n\geq 0} & \mapsto & \left(g^u\left(0, (v_n, w_n)_{n\geq 0}\right), ..., g^u\left(n-1, (v_n, w_n)_{n\geq 0}\right)\right). \end{array}$$

Remarque 3.43. *On notera que pour un n donné, F_n^u ne dépend que des n premiers termes de la suite u.*

De façon à ce que

$$\mathcal{T}_t = F_{N_t}^{(H_i)}\left((T_n, J_n)_{n\geq 0}\right).$$

Nous allons alors comme dans le cadre du coalescent de Kingman essayer de construire une relation récursive pour la probabilité d'obtenir une configuration donnée.

Remarque 3.44. *Soit $(t_1, .., t_l) \in \Pi^l$, alors on peut réécrire*

$$\{\mathcal{T}_t = (t_0, .., t_{l-1})\} = \left\{\forall i \in [\![1, l-1]\!]\ H_i < t, H_l \geq t, F_l^{(H_i)}\left((T_n, J_n)_{n\geq 0}\right) = (t_0, .., t_{l-1})\right\}.$$

Passons au résultat principale de cette section.

Proposition 3.45 (Formulation récursive de la probabilité d'une configuration)**.**
Soit $(t_0, ..., t_n) \in \Pi^{n+1}$, alors

$$\mathbb{P}\big(\mathcal{T}_t = (t_0...t_n)\big) = \sum_{\substack{i\in[\![1,n]\!] \\ t_i = t_{i-1}}} \lambda \int_0^t ds\ e^{-(n+1)(\lambda+\theta)s} \mathbb{P}\left(\mathcal{T}_{t-s} = (t_0, ..t_{i-1}, t_{i+1}..., t_n)\right) \tag{6}$$

$$+ \sum_{\substack{i\in[\![0,n]\!] \\ \forall j\neq i\ \ t_j \neq t_i}} \theta \int_0^t ds\ e^{-(n+1)(\lambda+\theta)s} \mathbb{P}\left(\mathcal{T}_{t-s} = (t_0, ..\Delta t_i..., t_n)\right), \tag{7}$$

avec Δ défini comme dans la section 2.4.2.

Démonstration. Comme dans le cadre de la formule récursive pour le coalescent de Kingman, nous allons conditionner notre probabilité aux derniers événements ayant eu lieu (dans le sens normal du temps). On note

- E^i_{mut}="le dernier événement est une mutation sur la branche i",
- E^i_{coal}="le dernier événement est la coalescence de la branche i".

Remarque 3.46. *Si le dernier événement observé est la coalescence de la branche i alors, de par le forme du modèle, celle-ci a forcement lieu avec la branche qui se trouve directement à sa gauche.*

On a alors

$$\mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n)\right) = \sum_{i \in [\![1,n]\!]} \mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{coal}\right) + \sum_{i \in [\![0,n]\!]} \mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{mut}\right).$$

Mais, d'une part les probabilités du type

$$\mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{coal}\right),$$

sont nulles si i et $i-1$ ne sont pas du même type, sinon il doit y avoir une mutation avant leur coalescence.

D'autre part, comme chaque mutation est unique, les probas du type

$$\mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{mut}\right),$$

sont nulle si l'individu i n'est pas le seul de son espèce, car sinon les "jumeaux" (enfin le groupe des individus partageant son type) doivent coalescer avant rencontrer une mutation.

Ceci conduit à la formule suivante :

$$\mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{coal}\right) = \sum_{\substack{i \in [\![1,n]\!] \\ t_i = t_{i-1}}} \mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{coal}\right) \tag{8}$$

$$+ \sum_{\substack{i \in [\![0,n]\!] \\ \forall j \neq i \ t_j \neq t_i}} \mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{mut}\right). \tag{9}$$

Nous allons maintenant étudier les différents termes de la somme.

Terme en coalescence :

Soit $i \in [\![1, n]\!]$ tel que $t_i = t_{i-1}$.

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{coal}\right) \\
&= \mathbb{P}\Big(F_n^{(H_j)_{j=1..n}}\left((T_n, J_n)_{n\geq 0}\right) = (t_0, ..., t_n), \forall j \in [\![1,n]\!]\ H_j < t, H_{n+1} \geq t, \forall j \in [\![1,n]\!]\ i \neq j,\ H_j > H_i,\ T_1 > H_i\Big) \\
&= \mathbb{P}\Big(F_{n-1}^{(H_j - H_i)_{j=1.,i-1,i+1..n}}\left((T_n - H_i, J_n)_{n\geq 0}\right) = (t_0, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n), \\
&\qquad \forall j \in [\![1,n]\!], i \neq j,\ H_j > H_i,\ T_1 > H_i \mid N_t = n+1\Big)\mathbb{P}\left(N_t = n+1\right).
\end{aligned}$$

Car $\{N_t = k\} = \{\forall j \in [\![1, k-1]\!]\ H_j < t,\ H_k \geq t\}$.

On pose désormais $\tilde{\mathbb{P}} = \mathbb{P}(. \mid N_t = n)$. Sous cette nouvelle probabilité les variables aléatoires $(H_j)_{j\geq 0}$ sont indépendantes et indépendantes de T_1. Il suit que

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{coal}\right) \\
&= \int_{\mathbb{R}} d\tilde{\mathbb{P}}_{H_i}(s)\ \tilde{\mathbb{P}}\Big(F_{n-1}^{(H_j-s)_{j=1.,i-1,i+1..n}}\left((T_n - s, J_n)_{n\geq 0}\right) = (t_0, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n), \\
&\quad \forall j \in [\![1, n]\!], i \neq j\ H_j > s,\ T_1 > s,\ H_{n+1} \geq s\Big)\mathbb{P}(N_t = n+1) \\
&= \int_{\mathbb{R}} d\mathbb{P}_{H_i | H_i < t}(s)\ \tilde{\mathbb{P}}\Big(F_{n-1}^{(H_j-s)_{j=1.,i-1,i+1..n}}\left((T_n - s, J_n)_{n\geq 0}\right) = (t_0, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n) \\
&\qquad\qquad \mid \forall j \in [\![1, n]\!], i \neq j\ H_j > s,\ T_1 > s,\ H_{n+1} \geq s\Big) \\
&\qquad \tilde{\mathbb{P}}(\forall j \in [\![1, n]\!], i \neq j\ H_j > s,\ T_1 > s,\ H_{n+1} \geq s)\,\mathbb{P}(N_t = n+1).
\end{aligned}$$

Or les variables aléatoires $(H_j)_{j\geq 0}$ sont indépendantes sous $\tilde{\mathbb{P}}$ et indépendantes de T_1. Ce qui donne

$$\begin{aligned}
\mathbb{P}\left(\mathcal{T}_t = (t_0, ..., t_n), E^i_{coal}\right) = \int_{\mathbb{R}} d\mathbb{P}_{H_i | H_i < t}(s)\ \tilde{\mathbb{P}}\Big(&F_{n-1}^{(H_j-s)_{j=1.,i-1,i+1..n}}\left((T_n - s, J_n)_{n\geq 0}\right) = (t_0, .t_{i-1}, t_{i+1}, .., t_n), \\
&\mid \forall j \in [\![1, n]\!], i \neq j\ H_j > s,\ T_1 > s,\ H_{n+1} \geq s\Big) \\
&\tilde{\mathbb{P}}(H_j > s)^{n-1}\,\tilde{\mathbb{P}}(H_{n+1} \geq s)\,\tilde{\mathbb{P}}(T_1 > s)\,\mathbb{P}(N_t = n+1).
\end{aligned}$$

Par ailleurs

$$\mathcal{L}_{\tilde{\mathbb{P}}}(H_j - s \mid H_j > s) = \mathcal{L}_{\mathbb{P}}(H_j \mid H_j < t - s).$$

De plus, la suite $(T_n)_{n\geq 0}$ a, conditionnellement à l'événement $N_t = n+1$, la loi des temps sauts d'un processus de Poisson de paramètre $(n+1)\,\theta$. Ce qui implique, par la propriété de Markov, que

$$\mathcal{L}_{\tilde{P}}\left((T_n - s)_{n\geq 1} \mid T_1 > s\right) = \mathcal{L}_{\tilde{P}}\left((T_n)_{n\geq 1}\right).$$

On en déduit que

$$\begin{aligned}
&\tilde{\mathbb{P}}\Big(F_{n-1}^{(H_j-s)_{j=1.,i-1,i+1..n}}\left((T_n - s, J_n)_{n\geq 0}\right) = (t_0, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n) \mid \forall j \neq i \leq n\ H_j > s,\ T_1 > s,\ H_{n+1} \geq s\Big) \\
&= \mathbb{P}\Big(F_{n-1}^{(H_j)_{j=1.,i-1,i+1..n}}\left((T_n, J_n)_{n\geq 0}\right) = (t_0, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n) \mid N_{t-s} = n\Big).
\end{aligned}$$

Puis, comme

$$\tilde{\mathbb{P}}(H_j > s)^{n-1} = \left(\frac{e^{-\lambda s} - e^{-\lambda t}}{1 - e^{-\lambda t}}\right)^{n-1},$$

$$\tilde{\mathbb{P}}(H_{n+1} \geq s) = 1,$$

$$\tilde{\mathbb{P}}(T_1 > s) = e^{-(n+1)\theta s},$$

on obtient

$$\begin{aligned}
\mathbb{P}\left(\mathcal{T}_t=(t_0,...,t_n),E^i_{coal}\right) &= \int_{\mathbb{R}} d\mathbb{P}_{H_i|H_i<t}(s)\ \mathbb{P}\left(\mathcal{T}_{t-s}=(t_0,\dots,t_{i-1},t_{i+1},\dots,t_n)\right)\mathbb{P}\left(N_{t-s}=n\right)^{-1}\\
&\qquad \left(\frac{e^{-\lambda s}-e^{-\lambda t}}{1-e^{-\lambda t}}\right)^{n-1} e^{-(n+1)\theta s}\mathbb{P}\left(N_t=n+1\right)\\
&= \int_{\mathbb{R}} d\mathbb{P}_{H_i|H_i<t}(s)\ \frac{\mathbb{P}\left(\mathcal{T}_{t-s}=(t_0,\dots,t_{i-1},t_{i+1},\dots,t_n)\right)}{e^{-\lambda(t-s)}\left(1-e^{-\lambda(t-s)}\right)^{n-1}}\\
&\qquad \left(\frac{e^{-\lambda s}-e^{-\lambda t}}{1-e^{-\lambda t}}\right)^{n-1} e^{-(n+1)\theta s}e^{-\lambda t}\left(1-e^{-\lambda t}\right)^{n}\\
&= \int_{\mathbb{R}} d\mathbb{P}_{H_i|H_i<t}(s)\ \mathbb{P}\left(\mathcal{T}_{t-s}=(t_0,\dots,t_{i-1},t_{i+1},\dots,t_n)\right)e^{-n\lambda s}\left(1-e^{-\lambda t}\right)\\
&= \int_0^t ds\ \lambda e^{-(n+1)\lambda s}\ e^{-(n+1)\theta s}\mathbb{P}\left(T_{t-s}=(t_0,..t_{i-1},t_{i+1}...,t_n)\right).
\end{aligned}$$

Terme en mutation :

Soit $i\in[\![1,n]\!]$ tel que pour tout $j\neq i,\ t_i\neq t_j$.

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{T}_t=(t_0,...,t_l),E^i_{mut}\right)\\
&=\mathbb{P}\left(F_n^{(H_j)_{j=1..n}}\left((T_n,J_n)_{n\geq 0}\right)=(t_0,...,t_n),\forall j\in[\![1,n]\!]\ H_j<t,H_{n+1}\geq t,\ \forall j\in[\![1,n]\!],\ H_j>T_1,J_1=i\right).
\end{aligned}$$

On utilise à nouveau la probabilité $\tilde{\mathbb{P}}$ introduite au dessus.

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{T}_t=(t_0,...,t_l),E^i_{mut}\right)\\
&\quad=\tilde{\mathbb{P}}\left(F_n^{(H_j)_{j=1..n}}\left((T_n,J_n)_{n\geq 1}\right)=(t_0,...,t_n),\ \forall j\in[\![1,n]\!],\ H_j>T_1,J_1=i\right)\\
&\quad=\tilde{\mathbb{P}}\left(F_n^{(H_j-T_1)_{j=1..n}}\left((T_n-T_1,J_n)_{n\geq 2}\right)=(t_0,\dots,\delta t_i,\dots,t_n),\ \forall j\in[\![1,n]\!],\ H_j>T_1,J_1=i\right)\mathbb{P}\left(N_t=n+1\right).
\end{aligned}$$

Or sous $\tilde{\mathbb{P}}$ on a que $\mathcal{L}_{\tilde{\mathbb{P}}}(T_1)=\mathcal{E}\left((n+1)\,\theta\right)$, alors

$$\begin{aligned}
&\mathbb{P}\left(T_t=(t_0,...,t_l),E^i_{mut}\right)\\
&=\int_0^t ds\ \ (n+1)\,\theta e^{-(n+1)\theta s}\tilde{\mathbb{P}}\left(F_n^{(H_j-s)_{j=1..n}}\left((T_n-s,J_n)_{n\geq 2}\right)=(t_0,.,\delta t_i..,t_n),\ \forall j\in[\![1,n]\!],\ H_j>s,J_1=i\right)\\
&\qquad\mathbb{P}\left(N_t=n+1\right)\\
&=\int_0^t ds\ \ (n+1)\,\theta e^{-(n+1)\theta s}\tilde{\mathbb{P}}\left(F_n^{(H_j-s)_{j=1..n}}\left((T_n-s,J_n)_{n\geq 2}\right)=(t_0,.,\delta t_i..,t_n),\ |\ \forall j\in[\![1,n]\!],\ H_j>s\right)\\
&\qquad\tilde{\mathbb{P}}\left(J_1=i\right)\tilde{\mathbb{P}}\left(\forall j\in[\![1,n]\!],\ H_j>s\right)\mathbb{P}\left(N_t=n+1\right).
\end{aligned}$$

Puis on utilise le fait que $\mathcal{L}_{\tilde{\mathbb{P}}}(J_1) = \mathcal{U}_{[\![0,n]\!]}$,

$$\begin{aligned}
&\mathbb{P}\left(T_t = (t_0, ..., t_l), E^i_{mut}\right) \\
&= \int_0^t ds\ \theta e^{-(n+1)\theta s} \mathbb{P}\left(\mathcal{T}_{t-s} = (t_0, ..., \delta t_i, ..t_n)\right) \mathbb{P}\left(N_{t-s} = n+1\right)^{-1} \tilde{\mathbb{P}}\left(\forall j \in [\![1,n]\!],\ H_j > s\right) \mathbb{P}\left(N_t = n+1\right) \\
&= \int_0^t ds\ \theta e^{-(n+1)\theta s} \mathbb{P}\left(\mathcal{T}_{t-s} = (t_0, ..., \delta t_i, ..t_n)\right)\ \frac{e^{-\lambda t}\left(1-e^{-\lambda t}\right)^n}{e^{-\lambda(t-s)}\left(1-e^{-\lambda(t-s)}\right)^n} \left(\frac{e^{-\lambda s} - e^{-\lambda t}}{1-e^{-\lambda t}}\right)^n \\
&= \int_0^t ds\ \theta e^{-(n+1)\theta s} \mathbb{P}\left(\mathcal{T}_{t-s} = (t_0, ..., \delta t_i, ..t_n)\right)\ e^{-(n+1)\lambda s}.
\end{aligned}$$

On obtient finalement la formule désirée. □

Cette formule pourrait être utilisée directement pour calculer la probabilité d'obtenir une configuration. Cependant, le coût de calcul augmente très rapidement avec la taille de la population. C'est pourquoi nous allons utiliser une méthode adaptée de [13].

3.5.1 Calcul par méthode MCMC

Pour cela, commençons par introduire le

Lemme 3.47. *Soit $X = (X_n)_{n\geq 0}$ une chaine de Markov homogène à valeur dans $(E, \mathcal{E})$ métrique complet. On note $P : E \times \mathcal{E} \to [0,1]$ la probabilité de transition de X, et définie par*

$$\forall A \in \mathcal{E} \quad \mathbb{P}\left(X_1 \in A \mid X_0\right) = P\left(X_0, A\right).$$

En particulier

$$\forall x \in E \quad A \in \mathcal{E} \to P(x, A) \text{ est une mesure sur } E.$$

Soient f une fonction définie sur E, $A \in \mathcal{E}$, et ν_A le temps d'atteinte de A.
Sous l'hypothèse que $\nu_A < \infty$ p.s., en posant

$$\forall x \in E, \quad u_f(x) = \mathbb{E}_x\left[\prod_{i=0}^{\nu_A} f\left(X_n\right)\right],$$

on a alors

$$u_f(x) = f(x)\ \int_E u_f(y)\, P(x, dy).$$

En particulier

$$\forall x \in A, \quad u_f(x) = f(x).$$

Démonstration. Soit $x \in E$,

$$\begin{aligned}
\mathbb{E}_x\left[\prod_{i=0}^{\nu_A} f\left(X_n\right)\right] &= f(x)\, \mathbb{E}_x\left[\mathbb{E}\left[\prod_{i=0}^{\nu_A} f\left(X_n\right) \circ \theta_1 \mid \sigma\left(X_0, X_1\right)\right]\right] \\
&= f(x)\ \mathbb{E}_x\left[u_f(X_1)\right] \\
&= f(x) \int_E u_f(y)\ P(x, dy).
\end{aligned}$$

□

Nous allons maintenant construire une chaîne de Markov et une fonction f sur l'espace des configurations, de façon à obtenir la formule récursive de la proposition 3.45. Soit f définit par

$$\begin{array}{rccl} f & : & \Pi & \to & \mathbb{R} \\ & & (t;t_0,\dots,t_n) & \mapsto & \begin{cases} \frac{\left(1-e^{-(n+1)(\lambda+\theta)t}\right)(\lambda\#\{i\in[\![1,n]\!]|t_i=t_{i-1}\}+\theta\#\{i\in[\![0,n]\!]|\forall j\neq i\ t_j\neq t_i\})}{(n+1)(\theta+\lambda)} & \text{si } n>1, \\ \mathbb{P}\left(T_t=(t_0,t_1)\right) & \text{si } n=1. \end{cases} \end{array}$$

Nous allons maintenant définir la probabilité de transition de la chaine de Markov : pour tout $t\in\mathbb{R}$, pour tout $(t_0,\dots,t_n)$ (avec $n>2$), soit $P\left(t,(t_0,\dots,t_n),\bullet\right)$ l'unique mesure sur $\mathbb{R}\times\Pi$ telle que

$$\forall A\in\mathcal{B}\left(\mathbb{R}\right),\ \forall\left(t'_0,\dots,t'_m\right)\in\Pi,$$

$$P\left(t,(t_0,\dots,t_n);A,(t'_0,\dots,t'_m)\right)=\begin{cases} \frac{1}{f(t_0,\dots,t_n)}\int_A ds\ \lambda e^{-(n+1)(\lambda+\theta)(t-s)}\mathbb{1}_{[0,t]}, \\ \qquad \text{si } \left(t'_0,\dots,t'_m\right)=\left(t_0,\dots t_{i-1},t_{i+1},\dots t_n\right), \\ \frac{1}{f(t_0,\dots,t_n)}\int_A ds\ \theta e^{-(n+1)(\lambda+\theta)(t-s)}\mathbb{1}_{[0,t]}, \\ \qquad \text{si } \left(t'_0,\dots,t'_m\right)=\left(t_0,\dots \delta t_i,\dots t_n\right), \\ 0 \qquad \text{sinon.} \end{cases}$$

Soit $X=\left(X_n\right)_{n\geq 0}$ une chaîne de Markov de transition P et $\nu=\inf\left\{n\geq 0\mid X_n^{(1)}\text{ de taille }2\right\}$. Alors, par le lemme précèdent

$$u_f\left(t,(t_0,\dots,t_n)\right)=f\left(t,(t_0,\dots,t_n)\right)\int_{\mathbb{R}\times\Pi}u_f\left(s\right)P\left(t,(t_0,\dots,t_n);ds\right) \tag{10}$$

$$=\sum_{\substack{i\in[\![1,n]\!]\\ t_i=t_{i-1}}}\int_0^t ds\ \lambda e^{-(n+1)(\lambda+\theta)(t-s)}u_f\left(s,(t_0,\dots t_{i-1},t_{i+1},\dots t_n)\right) \tag{11}$$

$$+\sum_{\substack{i\in[\![0,n]\!]\\ \forall j\neq i\ t_j\neq t_i}}\int_0^t ds\ \theta e^{-(n+1)(\lambda+\theta)(t-s)}u_f\left(s,(t_0,\dots t_{i-1},t_{i+1},\dots t_n)\right). \tag{12}$$

Ainsi, via un changement de variable dans l'intégrale, on voit que u_f vérifie la formule récursive. Par unicité de la solution, on en déduit que

$$\mathbb{P}\left(\mathcal{T}_t=(t_0,\dots,t_n)\right)=\mathbb{E}_{(t,(t_0,\dots,t_n))}\left[\prod_{i=0}^{\nu-1}f\left(X_i\right)f\left(X_\nu\right)\right],$$

avec en particulier $f\left(X_\nu\right)=\mathbb{P}\left(\mathcal{T}_s=(t_0,t_1)\mid X_\nu\right)$ (où $(s,(t_0,t_1))=X_\nu$). Pour pouvoir mettre en application cette méthode, il reste à savoir comment calculer la probabilité d'une configuration de taille 2.

Probabilité d'une configuration de taille 2 :

Soit $t_0,t_1\in\Pi$, on note $n_0=t_0\left(\mathbb{R}\right)$, et $n_1=t_1\left(\mathbb{R}\right)$. Les conditions de compatibilité avec le modèle de mutation impliquent que

$$\exists k^\star\in\mathbb{N},\ \forall n\geq k^\star,\ t_0\left\{n\right\}=t_1\left\{n\right\}.$$

On note $N_{com} = t_0\left([k^\star, \infty)\right)$, correspondant au nombre d'atomes que partagent t_0 et t_1. Ainsi

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{T}_t = (t_0, t_1)\right) \\
&= \mathbb{P}\left(H_1 < t, H_2 \geq t, N\left([0, H_1] \times \{0\}\right) = n_0 - N_{com}, N\left(]H_1, t] \times \{0\}\right) = N_{com}, N\left([0, H_1] \times \{1\}\right) = n_1 - N_{com}\right) \\
&= \mathbb{E}\Bigg[\mathbb{P}\left(N\left([0, H_1] \times \{0\}\right) = n_0 - N_{com} \mid H_1\right)\ \mathbb{P}\left(N\left(]H_1, t] \times \{0\}\right) = N_{com} \mid H_1\right) \\
&\qquad\qquad \mathbb{P}\left(N\left([0, H_1] \times \{1\}\right) = n_1 - N_{com} \mid H_1\right) \mathbb{1}_{H_1 < t}\Bigg] e^{-\lambda t} \\
&= \mathbb{E}\left[e^{-\theta H_1} \frac{\left(\theta H_1\right)^{n_0 - N_{com}}}{(n_0 - N_{com})!}\ e^{-\theta(t - H_1)} \frac{\left(\theta\left(t - H_1\right)\right)^{N_{com}}}{N_{com}!}\ e^{-\theta H_1} \frac{\left(\theta H_1\right)^{n_1 - N_{com}}}{(n_1 - N_{com})!} \mathbb{1}_{H_1 < t}\right] e^{-\lambda t} \\
&= \frac{\theta^{n_0 + n_1 - N_{com}}}{(n_0 - N_{com})!\,(n_1 - N_{com})! N_{com}!} e^{-(\lambda + \theta)t} \int_0^t ds\ e^{-\theta s} s^{(n_0 + n_1 - 2\ N_{com})} \left(t - s\right)^{N_{com}}.
\end{aligned}
$$

3.5.2 Applications numériques

Dans cette section nous allons appliquer les résultats obtenus précédemment sur des configurations assez simples. Nous comparerons d'une part les courbes obtenues via la formule de récurrence 3.33 directement avec celles obtenues grâce à la méthode MCMC. Tous les calculs ont été faits sous *scilab*.

Configuration sans mutation de taille 4

Pour cette section, nous considérons un échantillon de taille 4 en supposant une absence de mutation par rapport à l'individu ancestral (Une telle configuration se note (0 0 0 0)).

La figure 21 représente la courbe de vraisemblance de cette configuration calculée selon les deux méthodes :

- La courbe bleue a été calculée selon la méthode directe via la formule récursive. Pour 20 valeurs de θ, il a fallu plus de 55 secondes de calcul.
- Les croix correspondent aux valeurs calculées avec la méthode MCMC. Pour les 20 mêmes valeurs de θ, le calcul a été cette fois fait en 26 secondes, et en effectuant 1000 simulations par valeur de θ.

Le tout a été calculé avec les paramètres : $\lambda = 1.5$, et $t = 2$.

Configuration du type $(1, 2, 23, 23)$

Pour cette configuration nous considérons 4 individus vivants. On notera bien que la façon dont sont notées les mutations est totalement arbitraire, et seul compte dans le calcul les relations entre les individus (qui partagent quelles mutations avec qui etc...).

La figure 22 représente la courbe de vraisemblance de cette configuration calculé selon les deux méthodes :

- La courbe bleue a été calculée selon la méthode directe via la formule récursive. Pour 20 valeurs de θ, il a fallu plus de 4 heures de calcul.
- Les croix correspondent aux valeurs calculées avec la méthode MCMC. Pour les 20 mêmes valeurs de θ, le calcul a été fait en 15 minutes, et en effectuant 5000 simulations par valeur de θ.

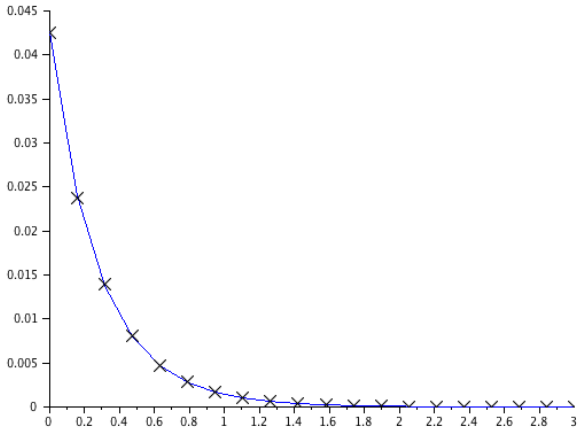


FIGURE 21 – Probabilité de la configuration (en ordonnée en fonction de θ (en abscisse).

Le tout a été calculé avec les paramètres : $\lambda = 1.5$, et $t = 2$.

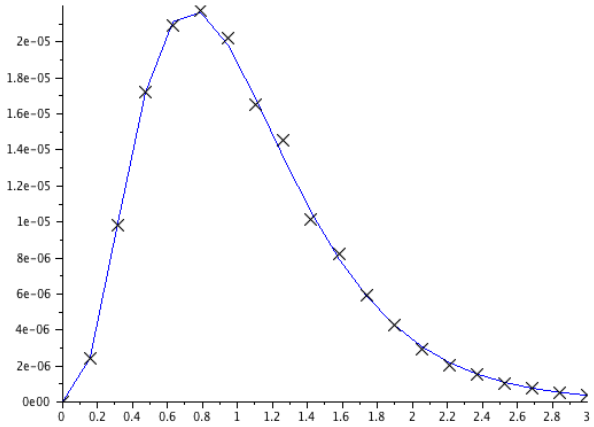


FIGURE 22 – Probabilité de la configuration (en ordonnée en fonction de θ (en abscisse).

4 Conclusion.

Au cours de ce travail nous avons présenté le cadre classique de la génétique des populations, puis étendu ces résultats au cadre plus récent des splitting trees. La principale question restée en suspend est celle de la loi du spectre de fréquence. Un premier progrès a été obtenu dans la proposition 3.31. Bien que celle-ci semble suggérer qu'il y a peu d'espoir d'obtenir des résultats

simples à t quelconque, les corollaires semblent laisser de l'espoir quant à la possibilité d'une loi asymptotique "agréable".

Ces résultats ouvrent donc des perspectives autant d'un point de vue théorique (extension de résultats sur les proportions asymptotiques d'allèles dans des populations branchantes en croissance) que d'un point de vue pratique (étude de l'influence du taux de croissance de la population sur des statistiques employées fréquemment par les biologistes, ex. hétérozygotie). Une question fondamentale est de tester si un gène donné est sélectionné positivement (la population porteuse du gène étant en croissance).

Concernant les méthodes d'inférence ancestrale, nos résultats préliminaires ouvrent des perspectives pour étendre les méthodes existantes aux splitting trees. D'autant que le modèle des splitting trees semble plus naturel du point de vue de la modélisation que le coalescent de Kingman pour étudier l'évolution inter-espèces (phylogénétique).

5 Annexes

5.1 Processus de Markov à espace d'état fini

On se référera notamment au Norris [7] pour tout ce qui concerne les chaînes de Markov à espace d'état fini ou dénombrable. Soit E un ensemble fini.

Définition 5.1 (Q-matrice)**.**
Une Q-matrice A d'éléments de E, est une matrice telle que
- $\forall i \in [\![1,n]\!] \quad A_{ii} < 0$,
- $\forall i,j \in [\![1,n]\!],\ i \neq j \quad A_{ij} \geq 0$,
- $\forall i \in [\![1,n]\!] \quad \sum_{j=1}^{n} A_{ij}$,

avec $n =$Card E.

Définition 5.2 (Chaine de Markov à temps continue)**.**
Un processus $(X_t, t \in \mathbb{R}_+)$ à valeur dans E continue à droite est appelé chaine de Markov à temps continue de Q-matrice A si il vérifie l'une des deux conditions équivalentes suivantes :

(1). $\forall t,h > 0$ X_{t+h} est indépendant de $(X_s, s \in [0,t])$ conditionnellement à X_t et

$$\forall i,j \in E \quad \mathbb{P}\left(X_{t+h} = j \mid X_t = i\right) = \delta_{ij} + A_{ij}h + o\left(h\right) \text{ uniformement par rapport à } t.$$

(2). $\forall n \in \mathbb{N}\ \forall t_1 \leq .. \leq t_{n+1} \in \mathbb{R}_+\ \forall \boldsymbol{i} \in E^n$

$$\mathbb{P}\left(X_{t_{n+1}} = i_{n+1} \mid X_{t_0} = i_0, ..., X_{t_n} = i_n\right) = \left(e^{(t_{n+1}-t_n)Q}\right)_{i_n i_{n+1}}.$$

Théorème 5.3.

5.2 Problème de Martingale

Théorème 5.4 (Problème de martingale en temps discret)**.** *Soit $(X_n)_{n\geq 0}$ un processus à valeur dans I au plus dénombrable, et de filtration naturelle $(\mathcal{F}_n)_{n\geq 0}$. Alors $(X_n)_{n\geq 0}$ est une chaine de Markov de probabilité de transition P* si et seulement si *pour toute fonction $f : I \to \mathbb{R}$ bornée le processus $\left(M_n^f\right)_{n\geq 0}$, définis ci-dessous, est une martingale.*

$$\forall n \quad M_n^f = f\left(X_n\right) - f\left(X_0\right) - \sum_{i=0}^{n-1} \left(P - I\right) f\left(X_i\right)$$

Théorème 5.5 (Problème de martingale pour les diffusions [12])**.**
Soit $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}\right)$ un espace probabilisé filtré. Soit b et σ deux fonctions de $\mathbb{R}$ dans $\mathbb{R}$ vérifiant les conditions suivantes :
- *b est Lipschitzienne.*
- *Il existe une fonction h telle que $h\left(0\right) = 0$, h strictement croissante et $\int_0^\epsilon ds\ h^{-2}\left(s\right) = \infty \quad \forall \epsilon > 0$, telle que $\forall x,y \quad \mid \sigma\left(x\right) - \sigma\left(y\right) \mid \leq h\left(\mid x-y \mid\right)$.*

Sous ces conditions, si $(X_t,\ t \in \mathbb{R}_+)$ un processus, tels que $\forall f \in \mathcal{C}_c^2$, la processus M^f définie par :

$$M_t^f = f\left(X_t\right) - f\left(X_0\right) - \int_0^t f''\left(X_s\right) b\left(X_s\right) ds$$

est une martingale. Alors X a même loi que la solution de l'EDS :

$$dY_t = \sigma\left(Y_s\right) dB_t + b\left(Y_s\right) dt$$

5.3 Critère de tension, et convergence de processus

Voir Billingsley [2].

5.3.1 Espace, et topologie de Skorokhod

Soit $T > 0$, on note $\mathcal{F}([0,T],\mathbb{R})$ l'ensemble des fonctions de $[0,T]$ dans $\mathbb{R}$.

Définition 5.6 (Espace de Skorokhod)**.**
L'espace de Skorokhod, noté $\mathbb{D}([0,T],\mathbb{R})$, est l'ensemble des fonctions de $\mathcal{F}([0,T],\mathbb{R})$ continues à droite et limitées à gauche.

Plus précisément, pour tout $f \in \mathcal{F}([0,T],\mathbb{R})$, $f \in \mathbb{D}([0,T],\mathbb{R})$ si et seulement si pour tout $t \in [0,T]$

- $\lim\limits_{\substack{s\to t\\ s<t}} f(s)$ *existe,*
- $\lim\limits_{\substack{s\to t\\ s>t}} f(s) = f(t)$.

Sur cet espace nous allons avoir besoin de quantifier la régularité des fonctions, c'est-à-dire quantifier leurs oscillations et leurs vitesses de saut.

Définition 5.7. *Soit $x \in \mathbb{D}([0,T],\mathbb{R})$, on note, pour tout intervalle $I \subset [0,T]$,*

$$w(x,I) = \sup |x(t) - x(s)| \ \mid s,t \in I.$$

On pour alors définir un module de continuité par

$$\forall \delta > 0 \quad w'(x,\delta) = \inf_{(t_i)\in\Delta^\delta[0,T]} \sup_i w(x,[t_i,t_{i+1}))\,,$$

avec $\Delta^\delta[0,T]$ l'ensemble des subdivisions de T de pas plus grand que δ.

Nous allons maintenant définir une distance sur $\mathbb{D}([0,T],\mathbb{R})$. Soit Γ l'ensemble des applications λ continue de $[0,T]$ dans $[0,T]$, telle que $\lambda(0) = 0$ et $\lambda(T) = T$.

Définition 5.8 (Métrique de Skorokhod)**.**
Pour tout $x, y \in \mathbb{D}([0,T],\mathbb{R})$, on définit la distance $d(x,y)$ de x à y par

$$d(x,y) = \inf\{\epsilon > 0 \mid \exists \lambda \in \Gamma \ |\lambda(t) - t| \le \epsilon \text{ et } |x(t) - y(\lambda(t))| \le \epsilon \ \forall t \in [0,T]\}\,.$$

Voici maintenant quelques résultats importants, notamment pour appliquer le théorème de Prokorov sur cet espace.

Proposition 5.9 (Résultats sur la topologie de Skorokhod)**.**

- *Il existe une distance d_0, équivalente à d, sous laquelle $\mathbb{D}([0,T],\mathbb{R})$ est un espace Polonais.*
- *La topologie induite (par la topologie de Skorokhod) sur l'espace des fonctions continues est la topologie associée à la norme uniforme.*

Le module de continuité w' permet de caractériser les compacts de $\mathbb{D}$ pour cette topologie.

Théorème 5.10 ([2])**.**
$A \subset \mathbb{D}$ est relativement compact si et seulement si

i. $\sup_{x\in A} \sup_{t\in[0,T]} |x(t)| < \infty$,

ii. $\sup_{x\in A} w'(x,\delta) \underset{\delta\to 0}{\longrightarrow} 0$.

A partir de ce théorème, le théorème de Prokhorov permet de caractériser les compacts de l'ensemble des proba sur $\mathbb{D}$ (loi de processus càdlàg). En particulier on a le résultat suivant :

Théorème 5.11 ([2])**.**
Soit (X^n) une suite de processus stochastiques dans $\mathbb{D}$. Alors X^n converge en loi vers X dans $\mathbb{D}$ ssi

i. $\left(\mathcal{L}\left[\sup_t |X^n(t)|\right]\right)_n$ *est uniformément tendu.*

ii. $\forall \epsilon, \eta > 0\ \exists \delta > 0$ *tel que* $\limsup_n \mathbb{P}(w'(X^n,\delta) > \eta) < \epsilon$.

iii. $\forall (t_1,\dots,t_p) \in [0,T] \setminus \{t : \mathbb{P}(\Delta X_t \neq 0) > 0\} \quad \left(X^n_{t_1},\dots,X^n_{t_p}\right) \xrightarrow{\mathcal{L}} \left(X_{t_1},\dots,X_{t_p}\right)$.

Un critère de convergence vers des chaînes de Markov en temps continu à espace d'état fini

Théorème 5.12. *Critère de convergence*
Soit E un ensemble fini, et $(X_t, t \in \mathbb{R}_+)$ une chaîne de Markov à valeur dans E de $Q-$matrice Q.

Soit $\left(X^N\right)_{N\geq 1}$ une suite de chaînes de Markov à valeur dans E, dont la matrice de transition vérifie

$$\forall x, y \in E, \quad P^N(x,y) = \delta_{x,y} + c_N Q_{x,y} + o(c_N). \quad (c)$$

Alors, si $c_n \underset{N\to\infty}{\to} 0$, et X_0^N converge en loi, on a

$$\left(X^N_{[t/c_N]}, t \in \mathbb{R}_+\right) \xrightarrow{\mathcal{L}} (X_t, t \in \mathbb{R}_+) \quad \text{dans } \mathcal{D}(\mathbb{R}_+, E).$$

Démonstration.
Nous allons appliquer le théorème précédent. i. est évident car E est fini.

On commence par montrer iii. . Par (c), on a que

$$\forall x, y \in E \forall t, s > 0 \quad \mathbb{P}(X^N_{[(t+s)/c_N]} = y \mid X^N_{[t/c_N]} = x) = (I + c_N Q_{x,y} + o(c_N))^{s/c_N}_{xy} \quad (13)$$

$$= (e^{sQ})_{xy} + o(1). \quad (14)$$

Donc

$$\mathbb{P}(X^N_{[(t+s)/c_N]} = y \mid X^N_{[t/c_N]} = x) \underset{N\to\infty}{\longrightarrow} \mathbb{P}(X_{t+s} = y \mid X_t = x).$$

Grâce à la propriété de Markov, on obtient la convergence des lois finies-dimensionnelles.

Pour avoir le résultat, il ne reste plus qu'à montrer la tension de la famille $\left(X^N\right)_N$. Soient $\delta > 0$, et $s, t \in \mathbb{R}$,

$$\mathbb{P}\left(\forall u \in [s,t]\ X^N_u = k \mid X^N_s = k\right) = 1 - \mathbb{P}\left(X^N \text{ saute sur } [s,t] \mid X^N_s = k\right).$$

Or, en posant $i = [s/c_N]$, et $j = [t/c_N]$, on a

$$\begin{aligned}
\mathbb{P}\left(\forall u \in [s,t]\ X_u^N = k \mid X_s^N = k\right) &= \mathbb{P}_k\left(\forall n \in [\![1, j-i]\!]\ X_n^N = k\right) \\
&= \prod_{n=1}^{j-i} \mathbb{P}\left(X_{n+1}^N = k \mid X_n^N = k\right) \\
&= \left(1 + Q_{kk} c_N + o\left(c_N\right)\right)^{j-i} \\
&= e^{(t-s)(Q_{kk}+\epsilon_N)} \\
&\geq e^{\delta Q_{kk}} e^{(t-s)\epsilon_N} \ (\text{car } Q_{kk} < 0),
\end{aligned}$$

avec $\epsilon_N \underset{N\to\infty}{\longrightarrow} 0$. Donc

$$\limsup_{N\to\infty} \mathbb{P}\left(\forall u \in [s,t]\ X_u^N = k \mid X_s^N = k\right) \geq 1 - O\left(t-s\right).$$

On note alors

$$S_{[s,t]}^X = \#\left(\{u \in [s,t] \mid X_{t-} - X_t > 0\}\right).$$

C'est le nombre de sauts de X sur l'intervalle $[s,t]$. Maintenant, on note, sur l'événement $\left\{S_{[s,t]}^X = 2\right\}$="Il y a 2 sauts dans l'intervalle [s,t]", τ_1 et τ_2 les temps des deux premiers sauts.

$$\begin{aligned}
\mathbb{P}\left(S_{[s,t]}^X \geq 2 \mid X_s = k\right) &= \mathbb{P}_k\left(S_{[0,t-s]}^X \geq 2\right) \\
&= \mathbb{E}_k\left[\mathbb{1}_{\tau_1 < t-s}\ \mathbb{P}_{X_{\tau_1}}\left(S_{[0,t-s-\tau_1]}^X \geq 1\right)\right] \\
&\leq \mathbb{E}_k\left[\mathbb{1}_{\tau_1 < t-s}\ \sup_{l\in E} \mathbb{P}_l\left(S_{[0,t-s]}^X \geq 1\right)\right] \\
&\leq C\left(t-s\right)\mathbb{P}_k\left(S_{[0,t-s]}^X \geq 1\right) \\
&\leq \left(C\left(t-s\right)\right)^2.
\end{aligned}$$

Finalement pour $\epsilon < 1$

$$\begin{aligned}
\mathbb{P}\left(w'\left(X_{[./c_N]}^N, \delta\right) \geq \epsilon\right) &= \mathbb{P}\left(\exists s,t \in [0,T],\ |s-t| < \delta, S_{[s,t]}^X \geq 2\right) \\
&\leq \mathbb{P}\left(\exists i \in [\![0, [T/\delta]]\!]\ S_{[(i-1)\delta,(i+1)\delta]}^X \geq 2\right) \\
&\leq \sum_{i=0}^{[T/\delta]-1} \sum_{k\in E} \mathbb{P}\left(S_{[(i-1)\delta,(i+1)\delta]}^X \geq 2 \mid X_{(i-1)\delta} = k\right)\mathbb{P}\left(X_{(i-1)\delta} = k\right) \\
&\leq [T/\delta]^{-1}\ C^2\delta^2.
\end{aligned}$$

Car les sommes ci-dessus sont finies.Alors

$$\limsup_{N\to\infty} \mathbb{P}\left(w'\left(X_{[./c_N]}^N, \delta\right) \geq \epsilon\right) < o\left(1\right).$$

On conclut alors avec le théorème 5.11. □

Références

[1] Jean Bertoin. *Lévy Processes*, volume 121. Cambridge university press, 1998.

[2] Patrick Billingsley. *Convergence of probability measures*, volume 493. Wiley. com, 2009.

[3] Nicolas Champagnat and Amaury Lambert. Splitting trees with neutral poissonian mutations I : Small families. *Stochastic Processes and their Applications*, 122(3) :1003–1033, 2012.

[4] Nicolas Champagnat and Amaury Lambert. Splitting trees with neutral poissonian mutations II : Largest and oldest families. *Stochastic Processes and their Applications*, 2012.

[5] Nicolas Champagnat, Amaury Lambert, and Mathieu Richard. Birth and death processes with neutral mutations. *International Journal of Stochastic Analysis*, 2012, 2012.

[6] Jochen Geiger and Götz Kersting. Depth—first search of random trees, and poisson point processes. In *Classical and modern branching processes*, pages 111–126. Springer, 1997.

[7] Norris J.R. *Markov Chains*. Cambridge University Press, 1997.

[8] Andreas Kyprianou. *Introductory Lectures on Fluctuations of Lévy processes with Applications (Universitext)*. Springer, 2006.

[9] Amaury Lambert. The contour of splitting trees is a lévy process. *The Annals of Probability*, 38(1) :348–395, 2010.

[10] Amaury Lambert. Some aspects of discrete branching processes. *dynamics*, 20 :43, 2010.

[11] Etienne Pardoux. Probabilistic models of population genetics. *Notes of Master 2 lectures*, 2009.

[12] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer, 1999.

[13] Simon Tavaré. Ancestral inference in population genetics. In *Lectures on probability theory and statistics*, volume 1837 of *Lecture Notes in Math.*, pages 1–188. Springer, Berlin, 2004.