Le coalescent de Kingman

Quentin Berger et Vincent Fall
sous la direction de Jean Bertoin

Table des matières

1 Introduction

Introduction :
Le coalescent de Kingman est un objet mathématique défini sur l'ensemble des relations d'équivalence de $\{1, ..., n\}$ ou $\mathbb{N}$ qui décrit la fusion de classes d'équivalence que nous appelerons ici blocs. Nous le présenterons tout d'abord pour un n fini puis grâce à quelques notions développées dans une deuxième partie nous pourrons le construire sur $\mathbb{N}$. Enfin, nous aborderons quelques applications du coalescent en biologie.

1.1 Introduction aux chaînes de Markov à temps continu

1.1.1 Q-matrice et première définition

Pour définir un n-coalescent nous avons besoin de quelques notions préliminaires concernant les chaînes de Markov à temps continu.

Définition
Soit I un ensemble dénombrable. Une Q-matrice $Q = (q_{ij})_{(i,j)\in I^2}$ sur I est une matrice qui satisfait les propriétés suivantes :

1. $\forall i \in I, 0 \leq -q_{ii} < \infty$
2. $\forall (i,j) \in I^2, i \neq j \Rightarrow q_{ij} \geq 0$ (les "taux de saut")
3. $\forall i \in I,\ \sum_{j\in I} q_{ij} = 0$

On peut citer ici un théorème qui relie Q-matrice et matrice stochastique :

Théorème 1.1. *Une matrice Q sur I est une Q-matrice si et seulement si $P(t) = exp(tQ)$ est une matrice stochastique pour tout $t \geq 0$.*

Définition :
Pour une Q-matrice Q, on pose

$$\mathbf{P} = (p_{ij}(t))_{i,j} = \exp(tQ)$$

Un processus $(X(t))_{t\geq 0}$ est *une chaîne de Markov* associé à la Q-matrice Q si :

$$\forall n \in \mathbb{N},\ \forall (t_1, ..., t_{n+1}) \in \mathbb{R}_+,\ \forall (i_0, ..., i_n) \in I^{n+1},$$

$$P(X_{t_{n+1}} = i_{n+1} \mid X_{t_n} = i_n, ..., X_{t_0} = i_0) = p_{i_n i_{n+1}}(t_{n+1} - t_n)$$

On peut cependant définir les chaînes de Markov de façon plus explicite.

1.1.2 Temps de séjour et chaîne de saut

Soit $(X(t))_{t\geq 0}$ une chaîne de Markov. On associe à ce processus deux autres processus (à temps discret cette fois) qui le caractérise.

Temps de séjour : On définit les temps durant lesquels le processus $(X(t))$ reste dans le même état
On définit les temps de saut T_k : $T_0 = 0$ et $T_{n+1} = inf\{(t \geq T_n / X_t \neq X_{T_n})\}$
T_k est le temps auquel le processus change d'état pour la $k^{\text{ième}}$ fois
On définit les temps de séjour τ_k : $\tau_k = T_{k+1} - T_k$

Chaîne de saut :
On définit la chaîne de saut $(Y_n)_{n\in\mathbb{N}}$ associée à (X_t) (les états successifs dans lesquels passe le processus (X_t)) :

$$\left\{ \begin{array}{c} Y_0 = X_0 \\ Y_n = X_{T_n} \end{array} \right.$$

Nouvelle définition de la chaîne de Markov :
A partir d'une Q-matrice, on va définir une matrice de transition pour la chaîne de saut, et une loi pour les temps de séjour.
. La matrice de saut Π est définie à partir de la Q-matrice Q (en notant $q_i = q_{ii}$) par :

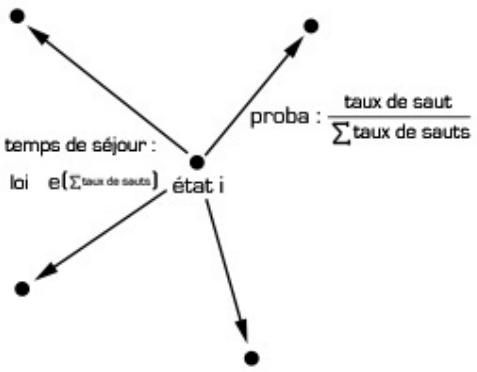
$$\Pi_{ij} = \left\{ \begin{array}{lll} \frac{q_{ij}}{q_i} & \text{si} & i \neq j \text{ et } q_i \neq 0 \\ 0 & \text{si} & i \neq j \text{ et } q_i = 0 \end{array} \right. \quad et \quad \Pi_{ii} = \left\{ \begin{array}{l} 1 \text{ si } q_i = 0 \\ 0 \text{ si } q_i \neq 0 \end{array} \right.$$

. Les temps de séjours dans l'état k sont donnés par une loi exponentielle de paramètre $-q_k$ (notée $\mathbf{e}(-q_k)$).

Définition :
Le processus $(X_t)_{t\geq 0}$ associée à la Q-matrice Q est une *chaîne de Markov* si et seulement si :

1. La chaîne de saut $(Y_n)_{n\in\mathbb{N}}$ est une chaîne de Markov à temps discret de matrice de transition Π.
2. Conditionnellement à $Y_0 = i_0, ..., Y_{n-1} = i_{n-1}$, les temps de séjour $\tau_1, ..., \tau_k, ..., \tau_n$ sont indépendants, de loi $\mathbf{e}(-q_k)$.

Donnons une intuition à la définition :
Le processus (X_t) reste un temps T dans un état i (T indépendant du passé, de loi exponentielle de paramètre dépendant de i, donné par la Q-matrice), puis va dans un autre état j indépendamment du passé (avec une probabilité donné par les taux de sauts de la Q-matrice).



Cette définition est équivalente à celle du 1.1.1. On dispose désormais de tous les outils pour définir un n-coalescent, et étudier ses premières propriétés.

1.2 Le n-coalescent

1.2.1 Définitions

Soit $\mathcal{E}_n$ l'ensemble fini des relations d'équivalence sur {1,...,n} (on le voit aussi comme l'ensemble des partitions de {1,...,n}). On note $|\xi|$ le nombre de blocs de la partition ξ

Définition :
Un *n-coalescent* est une chaîne de Markov $(R_t)_{t\geq 0}$ sur $\mathcal{E}_n$ telle que :
- $R_0 = \Lambda_n = \{(i,i), i \in \{1,...,n\}\}$ (la partition avec n blocs)
- Les taux de saut de la Q-matrice Q associée à (R_t) sont :

$$q_{\xi\eta} = \begin{cases} 1 & \text{si} \quad \xi \prec \eta \\ 0 & \text{sinon} \end{cases}$$

avec $\xi \prec \eta \Leftrightarrow \ \xi \subset \eta$ et $|\xi| = |\eta| + 1$ (en pratique, on fusionne deux blocs de ξ).

Remarque :
Les n-coalescents ont tous la même loi. En effet, les coalescents ont la même loi concernant leurs temps de séjour et la même matrice de transition. Ils sont donc issus de la même Q-matrice, et ont la même loi.

On a donc un processus sur $\mathcal{E}_n$ dont la chaîne de saut est une suite de partitions de {1,...,n} décroissante, et où l'on fusionne deux blocs d'une partition à vitesse 1 pour obtenir la suivante (la fusion de deux blocs s'appelle la coalescence).

Si on pose

$$q_\xi = \sum_{\eta \neq \xi} q_{\xi\eta}$$

Alors on a

$$q_\xi = \frac{|\xi|(|\xi|-1)}{2}$$

(car il y a $\frac{k(k-1)}{2}$ façon de fusionner deux blocs d'une partition à k blocs).
Le temps de séjour (calculé à partir de la Q-matrice du n-coalescent)dans l'état ξ avec $|\xi| = k$ est donc de loi $\mathbf{e}(d_k)$ avec

$$d_k = \frac{k(k-1)}{2} \tag{1}$$

.
Ensuite, on définit le processus

$$D_t = |R_t|$$

qui est une chaîne de Markov à valeur dans {1,...,n}, qui décroit de 1 en 1, et qui reste un temps τ_k de loi $\mathbf{e}(d_k)$ dans l'état $D_t = k$.

1.2.2 Premières propriétés

On note $\rho_{nN} : \mathcal{E}_N \to \mathcal{E}_n$ la restriction des partitions de {1,...,N} à {1,...,n}. (par la suite, on notera aussi ρ_n la restriction des partitions de $\mathbb{N}$ à {1,...,n})

Proposition 1.2. *Si N>n, alors un N-coalescent est un n-coalescent*

Démonstration. :
On fixe un N-coalescent $(R_t)_{t\geq 0}$. On considère $(\rho_n R_t)_{t\geq 0}$ le N-coalescent restreint à $\mathcal{E}_n$.
Alors $(\rho_{nN} R_t)_0 = \Lambda_n$. De plus, on a :

$$q_{\xi\eta} = \begin{cases} 1 \text{ si } \xi \prec \eta \\ 0 \text{ sinon} \end{cases}$$

et enfin $(\rho_n R_t)_{t\geq 0}$ est une chaîne de Markov à temps continu de Q-matrice la matrice Q (de (R_t)) restreinte aux $|\mathcal{E}_n|$ premières lignes et $|\mathcal{E}_n|$ premières colonnes en ordonnant correctement les indices des matrices correspondantes. Cette matrice est bien la Q-matrice d'un coalescent. □

1.3 Détermination de la loi du n-coalescent

On va maintenant déterminer la loi d'un n-coalescent (qui est la même pour tous les n-coalescents). Pour cela nous avons besoin d'un premier théorème :

Théorème 1.3. *Dans un n-coalescent, le processus de mort $(D_t)_{t\geq 0}$ et la chaîne de saut $(\mathcal{R}_k)_{k=n,n-1,...,1}$ sont indépendants et*

$$\forall t \geq 0, R_t = \mathcal{R}_{D_t} \tag{2}$$

Les probabilités de transition de la chaîne de Markov $(\mathcal{R}_k)$ sont données par

$$\mathbb{P}(\mathcal{R}_{k-1} = \eta \mid \mathcal{R}_k = \xi) = \begin{cases} \frac{2}{k(k-1)} & si \quad \xi \prec \eta \\ 0 & sinon \end{cases} \tag{3}$$

avec $\xi \in \mathcal{E}_n$, $|\xi| = k$ et $2 \leq k \leq n$.
La probabilité totale est donnée par :

$$\mathbb{P}(\mathcal{R}_k = \xi) = \frac{(n-k)!k!(k-1)!}{n!(n-1)!}\lambda_1!\lambda_2!...\lambda_k! \tag{4}$$

où $\lambda_1, ..., \lambda_k$ sont les tailles des blocs de ξ

Démonstration. :

Les probabilités de transition sont de la forme $q_{\xi\eta}/q_\xi$. Donc tant que $q_\xi \neq 0$ (à la fin, on arrive à un état où $q_\xi = 0$) les temps de séjour τ_ξ dans l'état ξ sont indépendants conditionnellement à la chaîne de saut, de loi $\mathbf{e}(q_\xi)$. (3) se déduit alors de la définition du n-coalescent.
Si $\mathcal{R}_k = \xi$ alors $q_\xi = d_k$. Donc la loi conditionnelle de (D_t) sachant $(\mathcal{R}_k)$ est la même que la loi non conditionnée. Et donc on a l'indépendance des deux familles et la première relation.

Ensuite, on prouve (4) par une récurrence descendante.
. Par définition, $\mathcal{R}_n = R_0$ et $R_0 = \Lambda_n$. Donc :

$$\text{si } |\xi| = n,\ \mathbb{P}(\mathcal{R}_n = \xi) = 1$$

Car la seule partition à n blocs est Λ_n, et $\lambda_1 = ... = \lambda_n = 1$, la formule est donc vérifiée pou k=n.

. On suppose désormais le résultat vrai au rang k. On pose $p_k(\xi) = \mathbb{P}(\mathcal{R}_k = \xi)$ avec $|\xi| = k$ et $\xi \in \mathcal{E}_n$. Par (3), on a

$$p_{k-1}(\nu) = \sum_{\xi \prec \nu} \frac{2p_k(\xi)}{k(k-1)}$$

Si $\lambda_1, ..., \lambda_{k-1}$ sont les blocs de ν, alors $\lambda_1, ..., \lambda_{l-1}, \nu, \lambda_l - \nu, \lambda_{l+1}, .., \lambda_{k-1}$ pour $1 \leq l \leq k-1$ et $1 \leq \nu \leq \lambda_l - 1$ sont les classes de ξ.
Par hypothèse de récurrence, on a :

$$\begin{aligned}
p_{k-1}(\nu) &= \sum_{\xi<\nu} \frac{2(n-k)!k!(k-1)!\lambda_1!...\lambda_{l-1}!\nu!(\lambda_l-\nu)!...\lambda_{k-1}!}{n!(n-1)!} \\
&= \sum_{l=1}^{k-1}\sum_{\nu=1}^{\lambda_l-1} \frac{2(n-k)!k!(k-1)!\lambda_1!...\lambda_{l-1}!\nu!(\lambda_l-\nu)!...\lambda_{k-1}!}{n!(n-1)!2}\binom{\lambda_l}{\mu} \\
&= \frac{(n-k)!(k-1)!(k-2)!\lambda_1!\lambda_2!...\lambda_{k-1}!}{n!(n-1)!}\sum_{l=1}^{k-1}\sum_{\nu=1}^{\lambda_l-1} 1 \\
&= \frac{(n-k)!(k-1)!(k-2)!\lambda_1!\lambda_2!...\lambda_{k-1}!}{n!(n-1)!}(n-(k-1))
\end{aligned}$$

ce qui conduit au résultat par récurrence. □

Le théorème nous permet ainsi de déterminer la loi du coalescent.
En effet :

$$\mathbb{P}(R_t = \xi) = \mathbb{P}(D_t = k \cap \mathbb{R}_k = \xi)$$

Par indépendance de $\{D_t = k\}$ et de $\{\mathcal{R}_k = \xi\}$ on a

$$\mathbb{P}(R_t = \xi) = \mathbb{P}(D_t = k)\mathbb{P}(\mathcal{R}_k = \xi)$$

De plus,

$$\mathbb{P}(D_t = k) = \mathbb{P}(\sum_{r=k+1}^{n} \tau_r \leq t) - \mathbb{P}(\sum_{r=k}^{n} \tau_r \leq t)$$

avec les τ_k des variables de loi exponentielle et de paramètres connus. On peut donc déterminer complètement la loi de D_t, et comme on a déterminé la loi de $\mathcal{R}_k$, on peut connaître entièrement la loi du n-coalescent.

2 Partitions aléatoires de {1,2,...} échangeables

Dans cette section, on va travailler sur des partitions échangeables de $\mathbb{N}$. Grâce à leurs propriétés, on pourra alors construire un coalescent à valeur dans l'ensemble des partitions échangeables.

2.1 Partitions aléatoires échangeables

2.1.1 Définitions

On note désormais $\mathcal{E}$ l'ensemble des partitions sur $\mathbb{N}$
Si π est une permutation $\pi : \mathbb{N} \to \mathbb{N}$, elle induit une bijection $\hat{\pi} : \mathcal{E} \to \mathcal{E}$ par

$$\hat{\pi}R = \{(\pi i, \pi j); (i,j) \in R\}$$

Définition :
Une mesure de probabilité R sur $\mathcal{E}$ est dite *échangeable* si, pour toute permutation $\pi : \mathbb{N} \to \mathbb{N}$,

$$\hat{\pi}R \overset{(\text{loi})}{=} R$$

Remarque : un processus $Y = (Y_i)_{i \geq 1}$ est *échangeable* si $\pi Y \overset{(\text{loi})}{=} Y$, où $\pi Y = (Y_{\pi(1)}, Y_{\pi(2)}, ...)$

2.1.2 Processus des "boîtes de peinture"

On va construire dans ce paragraphe un processus aléatoire dit des "boîtes de peinture" qui est un exemple de partition échangeable :

Le processus des "boîtes de peinture" se décrit de la manière suivante :
On définit

$$\Delta = \{ \ (x_0, x_1, ...) \ / \ x_r \geq 0 \ et \ \sum x_r = 1 \ \}$$

On prend $x = (x_0, x_1, ...)$ dans Δ, puis $Y_1, Y_2, ...$ i.i.d. de loi $\mathbb{P}(Y_i = r) = x_r$

On définit alors le processus R de "boîte de peinture" de loi P^x par

$$\forall \omega \in \Omega \ , \ les \ blocs \ de \ R(\omega) \ sont \ les \ \beta_r = \{i \ ; \ Y_i(\omega) = r\}$$

$$Autrement \ dit, \ (i,j) \in R(\omega) \iff Y_i(\omega) = Y_j(\omega) \geq 1$$

Remarque : si x est aléatoire de loi μ sur Δ, R est de loi $P^\mu = \int P^x \mu(dx)$

On peut visualiser la construction de cette partition aléatoire de la manière suivante : on a des couleurs $C_1, C_2, ...$ toutes différentes, et on assigne la couleur C_r à un élément k avec probabilité x_r. On a alors une partition de $\mathbb{N}$ constituée des blocs d'éléments de même couleur (ou bien $(i,j) \in R \iff$ i et j ont même couleur).

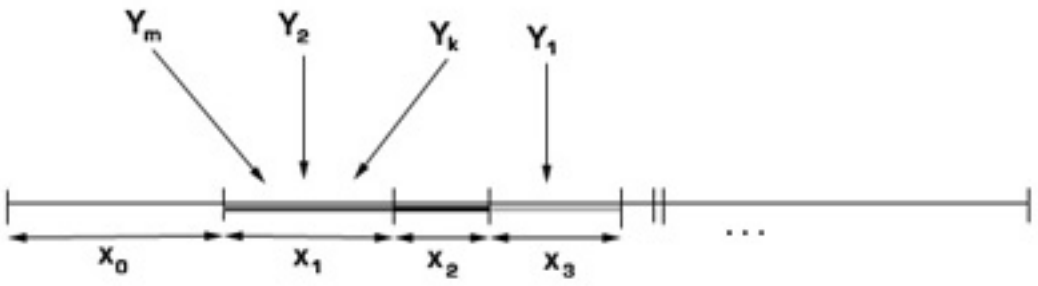
Construction équivalente : On remarque que $x_0, x_1, ...$ "découpe" [0,1] en intervalles de taille x_i :

$$[0, x_0[, [x_0, x_0 + x_1[, [x_0 + x_1, x_0 + x_1 + x_2[...$$

On peut alors considérer $Z_1, Z_2, ...$ i.i.d de loi uniforme sur [0,1], et construire les blocs de R ainsi :

$$\beta_r = \left\{ i \ ; \ Z_i(\omega) \in \left[\sum_{k=0}^{r-1} x_k, \sum_{k=0}^{r-1} x_k + x_r \right[\right\}$$

qui donne la même loi pour R



Ce procédé définit une partition échangeable de {1,2,...}.
En effet, les Y_i sont i.i.d, ce qui implique que $(Y_i)_{i\geq 1}$ est un processus échangeable. Donc "$Y_i = Y_j$" a la même loi que "$Y_{\pi(i)} = Y_{\pi(j)}$".

De plus, P^x reste inchangé si on permute les x_r, $r \geq 1$ (car ce qui est important, c'est que les Y_i et Y_j soient dans le même intervalle, l'intervalle importe peu). On réordonne donc les x_r, et on pourra désormais considérer que $x_1 \geq x_2 \geq ...$ (par exemple si l'on souhaite que les couleurs soient ordonnées, de la plus représentée à la moins représentée).

On note

$$\nabla = \{(x_1, x_2, ...), \ x_1 \geq x_2 \geq ... \geq 0 \ et \sum_{k\geq 1} x_k \leq 1\}$$

La loi de R un processus de "boîte de peinture" est entièrement défini par un élément de ∇. (R est de loi $P^{(1-\sum x_k, x_1, x_2, ...)}$ pour $(x_1, x_2, ...) \in \nabla$). Si x $\in \nabla$, on notera aussi P^x la loi de R.

On a donc un exemple de partition aléatoire échangeable, et la partie suivante démontre que toute partition échangeable de {1,2,...} est de cette forme.

2.2 Théorème des partitions échangeables

Nous avons maintenant besoin de quelques outils supplémentaires, pour comprendre le théorème de de Finetti, dont une variante est le théorème qui nous intéresse :

Définition : Mélange de variables aléatoires i.i.d

Soit α une mesure de probabilité sur $\mathcal{M}(\mathbb{R})$ (l'ensemble des mesures de probabilité sur $\mathbb{R}$) : pour tout $\omega \in \Omega,\ \alpha(\omega)$ est une probabilité sur $\mathbb{R}$.
On définit alors $\forall \omega \in \Omega,\ X(\omega) = (X_1, X_2, ...)$ où les X_i sont i.i.d de loi $\alpha(\omega)$
On dit que X est un *mélange de v.a i.i.d* dirigé par α

2.2.1 Théorème de de Finetti

Théorème 2.1. *Tout processus aléatoire échangeable $Z = (Z_i)_{i\geq 1}$ est un mélange de v.a i.i.d*

Il est clair qu'un mélange de v.a i.i.d est un processus échangeable, ce théorème nous apprend donc que tout processus échangeable est de cette forme.

Remarque : on peut retrouver la mesure directrice de Z grâce à l'approximation empirique :

$$\Lambda_n(Z_1, ..., Z_n) = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i} \ (où\ \delta_Z(A) = 1_{\{Z\in A\}})$$

$$\text{on note } \Lambda(Z) = \begin{cases} \lim_{n\to\infty} \Lambda_n(Z_1, ..., Z_n) \text{ si la limite existe} \\ \delta_0 \text{ si la limite n'existe pas} \end{cases}$$

La loi forte des grands nombres montre que si Z est un mélange de v.a i.i.d dirigé par μ, Alors

$$\Lambda(Z) \overset{p.s}{=} \mu$$

2.2.2 Variante du théorème de de Finetti pour les partitions échangables

Théorème 2.2. *Si R est une partition échangeable de {1,2,...}, et si $\lambda_r(n)$ désigne le $r^{ième}$ plus grand bloc de $\rho_n R$, Alors*

1. $\frac{1}{n}(\lambda_0(n), \lambda_1(n), ...) \xrightarrow[n\to\infty]{p.s} (D_1, D_2, ...) = D\ où\ D \in \nabla$
2. *R est un processus de "boîte de peinture" de loi P^D*

Démonstration. :

1. Pour un processus de "boîte de peinture de loi P^x (x dans ∇), on a

   $$\frac{1}{n}(\lambda_0(n), \lambda_1(n), ...) \xrightarrow[n\to\infty]{p.s} (x_1, x_2, ...) = x$$

   Car si $(Y_i)_{1\leq i\leq n}$ est i.i.d de loi uniforme sur [0,1], la proportion de Y_i qui tombent dans l'intervalle $\left[\sum_{k=0}^{r-1} x_k, \sum_{k=0}^{r-1} x_k + x_r\right[$ est en moyenne la taille de l'intervalle, c'est-à-dire x_r. Si les x_k sont rangés dans l'ordre décroissant, les i tels que Y_i tombent cet intervalle sont dans le $r^{\text{ième}}$ plus grand bloc de $\rho_n R$ (au moins pour n grand). Et on a

   $$\frac{1}{n}\lambda_r(n) \xrightarrow[n\to\infty]{p.s} \text{"proportion de } Y_i \text{ qui tombent dans l'intervalle } \left[\sum_{k=0}^{r-1} x_k, \sum_{k=0}^{r-1} x_k + x_r\right[\text{"}$$

   $$\xrightarrow[n\to\infty]{p.s} \mathbb{P}\left(Y_i \in \left[\sum_{k=0}^{r-1} x_k, \sum_{k=0}^{r-1} x_k + x_r\right[\right) = x_r$$

   Il suffit donc de vérifier que R est un processus de boîte de peinture.

2. Soit (ξ_i) des v.a i.i.d, de loi uniforme sur [0,1], indépendantes de R.
Les $\xi_i(\omega)$ sont tous distincts (à un ensemble de mesure nulle près)
On définit :

$$F_i = \min\{\text{j, i et j sont dans le même bloc de } R(\omega)\}$$
$$Z_i = \xi_{F_i}$$

Ainsi, pout tout ω, la partition R(ω) est exactement la partition de blocs {i, $Z_i(\omega) = z$}

Montrons que (Z_i) est échangeable :
On a $(Z_i) = g((\xi_i), R)$ pour une certaine fonction g, et alors $(Z_{\pi(i)}) = g((\xi_{\pi(i)}), \hat{\pi}R)$. Or $((\xi_{\pi(i)}), \hat{\pi}R) \overset{(loi)}{=} ((\xi_i), R)$ par échangeabilité et indépendance de R et (ξ_i). Donc finalement $((\pi Z_i) \overset{(loi)}{=} (Z_i))$
Par le théoème de de Finetti, Z est un mélange de v.a i.i.d de mesure directrice α. Alors, conditionnellement à $\alpha = \mu$, (Z_i) est i.i.d de loi μ. Alors les blocs de R sont les {i, $Z_i(\omega) = z$} et R est un processus de "boîte de peinture". En effet, si x_k sont les atomes de μ et si (Y_i) est i.i.d de loi $\mathbb{P}[Y_i = r] = \mu(x_r) (= \mathbb{P}[Z_i = x_r])$, alors R est bien un processus de boîte de peinture de loi P^D, où $D = (\mu(x1), \mu(x2), ...) \in \nabla$ (on peut ranger les atomes dans l'ordre de poids décroissant d'après la remarque du 2.1.2)

□

2.3 Partitions échangeables issue de la "boîte de peinture uniforme"

On s'intéresse ici au cas où R possède moins de k blocs p.s, et où μ est proportionnelle à la mesure de Lebesgue sur le simplex

$$\Delta_k = \{(x_1, x_2, ..., x_k) \ ; \ x_r \geq 0, \sum_{r=1}^{k} x_k = 1\}$$

On note $\mathcal{P}_k$ la mesure de probabilité sur $\mathcal{E}$ correspondante (en choisissant $x_0 = 0$)

$$\mathcal{P}_k = \int_{\Delta_k} \cdots \int P^{(0,x_1,...,x_k,0,...)}(k-1)! dx_1 ... dx_{k-1}$$

On suppose maintenant que R est de loi $\mathcal{P}_k$ (On utilisera ces partitions aléatoires, pour $k \leq 1$, afin de construire la chaîne de saut du coalescent)

On peut calculer la loi de sa restriction $\rho_n R$:
Si $\xi \in \mathcal{E}_n, |\xi| = l \leq k$, on a

$$\mathbb{P}[\rho_n R = \xi] = \frac{k!(k-1)!\lambda_1!\lambda_2!...\lambda_l!}{(k-l)!(k-1+n)!}$$

(Détails des calculs en appendice)

Donc

$$\mathbb{P}\left[\rho_n R = \xi\right] = \mathcal{P}_k\left(\rho_n^{-1}(\xi)\right) = \frac{k!(k-1)!\lambda_1!\lambda_2!...\lambda_l!}{(k-l)!(k-1+n)!}$$

On peut aussi calculer la probabilité que le nombre de blocs de $\rho_n R$ soit k sous $\mathcal{P}_k$.

$$\mathbb{P}\left[|\rho_n R| = k\right] = \frac{n!(n-1)!}{(n+k-1)!(n-k)!} \text{ sous } \mathcal{P}_k$$

et ainsi,

$$\mathbb{P}\left[\mathcal{R}_k = \xi\right] = \mathbb{P}\left[\rho_n R = \xi \ \mid \ |\rho_n R| = k\right]$$

Et comme, sous $\mathcal{P}_k$,

$$\mathbb{P}\left[|\rho_n R| = k\right] \xrightarrow[n\to\infty]{} 1$$

On peut raisonnablement penser que $\mathcal{P}_k$ est la "forme limite" de la loi de $\mathcal{R}_k$ quand n tend vers $+\infty$, car on a vraisemblablement

$$\mathbb{P}\left[\mathcal{R}_k = \xi\right] \sim_{n\to\infty} \mathcal{P}_k\left[\rho_n^{-1}(\xi)\ \right]$$

Autrement dit, lorsque l'on va construire le coalescent, il est raisonnable d'utiliser une chaîne de saut $(\mathcal{R}_k)$, $\mathcal{R}_k$ de loi $\mathcal{P}_k$ (supposée forme limite de la loi des $\mathcal{R}_k$ pour un n-coalescent)

3 Le coalescent

3.1 Construction

Avant d'envisager la construction d'un coalescent, rappelons sa définition.
Définition : Un processus sur $\mathcal{E}$ $(\mathcal{R}_t, t \geq 0)$ est appelé un coalescent si :
$\forall n \in \mathbb{N}, (\rho_n \mathcal{R}_t)$ est un n-coalescent.
On va maintenant construire un coalescent.

Théorème 3.1. *Il existe une chaîne de Markov $(R_k, k = 1, .., n, ..)$ où les valeurs possibles de $\mathcal{R}_k$ sont les relations de $\mathcal{E}$ avec exactement k blocs, telles que $\mathcal{R}_k$ a la loi $\mathcal{P}_k$ et*

$$\mathbb{P}(\mathcal{R}_{k-1} = \nu \mid \mathcal{R}_k = \xi) = \begin{cases} \frac{2}{k(k-1)} \text{ si } \xi \prec \nu \\ 0 \text{ sinon} \end{cases}$$

avec $\xi \in \mathcal{E}, |\xi| = k$. Si $(D_t, t > 0)$ est un processus de mort avec le taux de mort $d_k = \frac{1}{2}k(k-1)$ indépendant de $\mathcal{R}_k$ et tel que $D_t \xrightarrow[t \to 0]{} \infty$, alors

$$R_t = \begin{cases} \Lambda \text{ si } t = 0 \\ \mathcal{R}_{D_t} \text{ si } t > 0 \end{cases}$$

définit une chaîne de Markov sur $\mathcal{E}$ pour lequel

$$\forall n \in \mathbb{N}, (\rho_n R_t, t \geq 0)$$

est un n-coalescent.

Remarque : Un tel processus de mort existe. On peut le définir pour $n = \infty$ car la série $\sum_{k \geq 0} \frac{1}{d_k}$ converge. La transition de k à k-1 se fait au temps $\sum_{r=k}^{\infty} \tau_r$ où les τ_r sont indépendants de loi exponentielle de paramètre d_r et ainsi le temps de transition a une espérance finie : $\sum_{r=k}^{\infty} \frac{1}{d_r} = \frac{2}{k-1}$.

Démonstration. :
Si $(X_1, X_2, ..., X_k)$ est uniformément distribué sur le simplex Δ_k, alors on peut obtenir $(X_1, ..., X_{l-1}, X', X_{l+2}, ..., X_k)$ uniformément distribué sur le simplex Δ_{k-1} en fusionnant deux blocs de manière aléatoire. En appliquant ce résultat aux fréquences X_r des boîtes de peinture de loi $\mathcal{P}_k$ et en choisissant R aléatoirement, X' est la fréquence correspondante à une relation R' obtenue à partir de R en combinant deux blocs aléatoirement. Cela entraîne que R' est de loi $\mathcal{P}_{k-1}$ car on décrit le même phénomène à l'échelon inférieur. On a ainsi par construction les probabilités de transition désirées et k blocs à chaque étape. On a alors l'existence de la chaîne de Markov.

On prend alors R_t comme défini dans le théorème. On s'intéresse pour $s > 0$ à la loi de $(R_{s+t, t \geq 0})$ conditionnellement à $(R_u, u \leq s)$. R_{s+t} ne dépend que de D_{t+s} et de $(\mathcal{R}_k, k \leq D_s)$ car R_{s+t} ne dépend que du nombre de blocs et l'état des blocs pour des $t \leq 0$. Or, D et $\mathcal{R}$ obéissent à des propriétés de Markov. Donc R_{s+t} ne dépend plus que de D_s et $\mathcal{R}_{D_s}$, et donc seulement de R_s. Ainsi R est une chaîne de Markov car elle possède la propriété de Markov.

On va maintenant vérifier que (R_t) est un coalescent. Par l'échangeabilité de $\mathcal{R}_k$ (cf construction), on a

$$\forall i \neq j, \mathbb{P}((i,j) \in \mathcal{R}_k) = \mathbb{P}((1,2) \in \mathcal{R}_k) = \mathbb{P}(\rho_2 \mathcal{R}_k = \rho_2 \theta)$$

avec θ la relation à un seul bloc.
Or, $\mathbb{P}(\rho_n \mathcal{R}_k = \xi) = \frac{k!(k-1)!}{(k-l)!(n+k-1)!} \lambda_1 ! ... \lambda_l !$.
Comme ici n=2 et l=1, on obtient :

$$\mathbb{P}((i,j) \in \mathcal{R}_k) = \frac{2}{k+1}$$

De plus, par indépendance,

$$\begin{aligned}
\mathbb{P}((i,j) \in R_t) &= \sum_k \mathbb{P}(R_t = \mathcal{R}_k) \mathbb{P}((i,j) \in \mathcal{R}_k) \\
&= \sum_k \frac{2}{k+1} \mathbb{P}(D_t = k) \\
&= \mathbb{E}\left[\frac{2}{D_t + 1}\right]
\end{aligned}$$

Donc comme $D_t \xrightarrow[t \to 0]{} \infty$ on a par convergence dominée :

$$\mathbb{E}\left[\frac{2}{D_t + 1}\right] \xrightarrow[t \to 0]{} 0$$

D'où

$$R_t \xrightarrow[t \to 0]{} \Lambda$$

car pour $i \neq j$ $\mathbb{P}((i,j) \in \rho_n R_t) \xrightarrow[t \to 0]{} 0$ par continuité des applications ρ_n. Ainsi, pour t assez petit on a : $\rho_n R_t = \Lambda_n$.

Soit $N \in \mathbb{N}$. On choisit $C_1, C_2, ..., C_N$ les classes d'équivalences de $\mathcal{R}_N$, ordonnées telles que le plus petit élément de chaque classe C_r soit c_r et que $c_1 \leq c_2 \leq ... \leq c_N$. Grâce au résultat fraîchement démontré on peut en déduire que :

$$\forall n \in \mathbb{N},\ \exists n_0 \in \mathbb{N} \ / \ \forall N \geq n_0,\ r \in C_r$$

avec $r \in \{1, .., n\}$

On définit ensuite $R_t^{(N)} \in \mathcal{E}_N$ la relation telle que $i \sim j$ si C_i et C_j sont dans un même bloc de $R_{T(N)+t}$ avec $T(N) = \sum_{r=N+1}^{\infty} \tau_r$ (l'instant où D_t arrive à N). On regarde ainsi le processus au moment où il rentre dans ∇_N.

Par la première partie du théorème, $(\mathcal{R}_k^{(N)}, k = N, N-1, ..., 1)$ est une chaîne de Markov, indépendante du processus de mort (en remarquant que $|R_t^{(N)}| = D_{t+T(N)}$) , avec les probabilités de transition de la forme énoncée dans le théorème. Alors, $(R_t^{(N)}, t \geq 0)$ est un N-coalescent. Par

la propriété (1.3.2) sur les N-coalescents, $(\rho_{nN} R_t^{(N)})$ est un n-coalescent.
De plus, pour un n fixé, on a un N suffisamment grand tel que :

$$\rho_{nN} R_t^{(N)} = \rho_n R_{t+T(N)}$$

. En effet,

$$\begin{aligned}(i,j) \in \rho_{nN} R_t^{(N)} &\iff C_i \text{ et } C_j \text{ sont dans le même bloc} \\ &\iff (i,j) \in R_{T(N)+t} \text{ et } (i,j) \in \mathcal{E}_n \\ &\iff (i,j) \in \rho_n R_{T(N)+t}\end{aligned}$$

et donc

$$\rho_n R_{t+T(N)} \xrightarrow[N\to\infty]{} \rho_n R_t$$

par continuité à droite du coalescent construit.
Donc $(\rho_n R_t)$ est un n-coalescent comme limite presque sûre de n-coalescents.

□

3.2 Propriétés

3.2.1 Théorème

On choisit dans cette section un coalescent continu à droite et ayant une limite à gauche. Les coalescents continus à droite ont la même loi. On va donner l'intuition de ce résultat : en effet, tous les coalescents ont les mêmes lois de dimension finie : pour un n fixé,

$$\forall t_1, ..., t_k,\ \forall A_1, ..., A_n \subseteq \mathcal{E},\ \mathbb{P}(R_{t_1} \in A_1, ..., R_{t_n} \in A_n)$$

est la même pour tous les coalescents (ceci est vrai pour les n-coalescent et la propriété de Stone-Weierstrass permet de l'obtenir pour les coalescents en faisant n$\to +\infty$). Alors toutes les probabilités liées au coalescent peuvent être obtenues à partir des lois de dimension finie en utilisant la continuité à droite et l'existence d'une limite à gauche du coalescent.

Théorème 3.2. *Soit $(R_t, t \geq 0)$ un tel coalescent. Alors (R_t) est une chaîne de Markov avec $\forall t > 0, P(R_t \in \mathcal{E}*) = 1$ avec $\mathcal{E}*$ l'ensemble des relations d'équivalence sur les entiers naturels tel que les blocs soient en nombre fini mais de taille infinie. Alors le processus de mort $D_t = |R_t|$ est un processus de mort avec taux de mort $d_k = \frac{1}{2}k(k-1)$ qui satisfait $D_t \xrightarrow[t\to 0]{} \infty$*

On peut toujours définir la chaîne de saut qui est indépendante du processus de mort. La chaîne de saut est de Markov et les probabilités de transition sont conservées. En particulier pour tout E dans $\mathcal{E}$,

$$\mathbb{P}(R_t \in E) = \sum_{k=1}^{\infty} \mathbb{P}(D_t = k)\mathcal{P}(E)$$

3.2.2 Application

On peut calculer la probabilité de $R_{t+s}, t > 0$ conditionnellement à $R_s = \xi \in \mathcal{E}*$. On sait que $R_{s+t} \in \{\nu \ / \ \nu \ \supseteq \xi\}$. On peut donc décrire R_{t+s} par la relation $\frac{\nu}{\xi}$ c'est-à-dire celle induite par ν sur les blocs de ξ. Alors en posant $R_t^{\geq s} = \frac{R_{s+t}}{R_s}$ on obtient un R_s-coalescent. En pratique, cela revient à dire que l'on considère les blocs de R_s comme des individus. Avec le théorème de la première partie on peut donc calculer $\mathbb{P}(R_t^{\geq s} = \xi)$ pour tout $\xi \in \mathcal{E}_n$. Finalement on a

$$\mathbb{P}(R_{s+t}^{\geq s} = \nu \mid R_s = \xi) = \begin{cases} 0 \text{ si } \xi \subset \nu \\ \mathbb{P}(D_t = |\frac{\nu}{\xi}|)\mathbb{P}(\mathcal{R}_{|\frac{\nu}{\xi}|} = \frac{\nu}{\xi}) \text{ sinon} \end{cases}$$

4 Applications du coalescent en Biologie

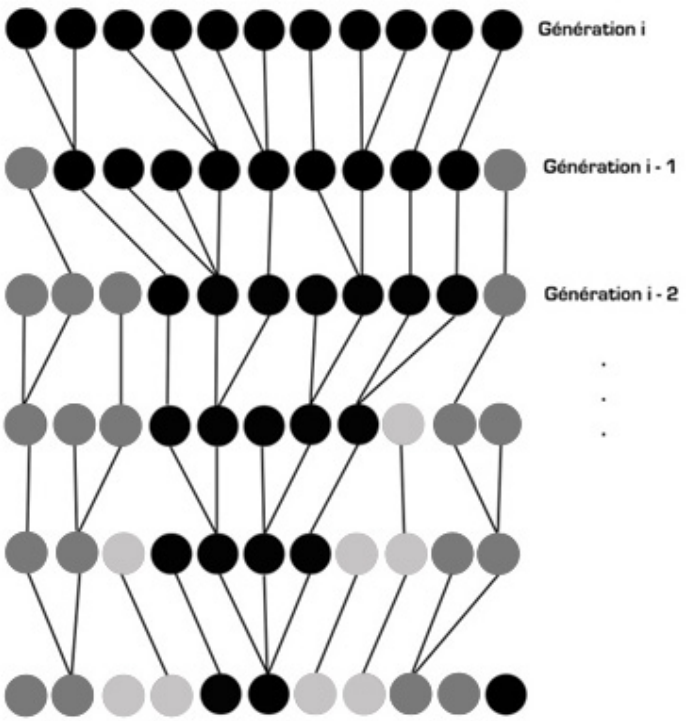
4.1 Modèle de Wright-Fisher

4.1.1 Présentation

Le modèle de Wright-Fisher concerne l'étude de la génétique des populations haploïdes (c'est à dire avec un seul allèle, et un individu n'a donc qu'un seul parent). Il représente l'évolution d'un allèle à l'intérieur d'une population. Il fait les hypothèses simplifiées suivantes :

1. La population est de taille constante N
2. Les générations sont disjointes (la population est renouvelée en entier à chaque cycle)
3. Il n'y a aucune sélection : tous les allèles ont la même chance de survie, la même chance d'avoir des descendants
4. Le nombre d'enfants est aléatoire, et est indépendant des autres individus.

Remarque : on peut facilement étudier le cas diploïde (avec deux allèles, donc avec deux parents), en considérant une population de taille 2N (on étudie en fait la population des allèles, qui se comporte comme une population haploïde).
On peut simuler facilement ce modèle grâce à cet algorithme simple : pour chaque individu de la génération, on choisit un parent au hasard parmi la génération précédente, chaque individu choisissant son parent indépendamment des autres individus.



4.1.2 Simulation du modèle

On peut commencer à appréhender ce modèle par le calcul de quelques probabilités :
. Si on prend deux individus i et j, la probabilité qu'ils descendent du même parent est $\frac{1}{N}$ (i choisit son parent, puis j doit choisir le même, avec probabilité $\frac{1}{N}$).
. Si T_2 est le nombre de générations qu'il faut remonter pour trouver l'ancêtre commun à i et j, on a

$$\mathbb{P}\left[T_2 \geq k\right] = \left(1 - \frac{1}{N}\right)^k$$

(la probabilité que, pour chacune des k générations précédentes, les ancêtres de i et j n'aient pas le même parent)

On caractérise ainsi un processus $(\widetilde{R}_k^N)_{k\geq 0}$ à valeur dans les partitions de {1,...,N} :

- Au départ, tous les individus sont distincts : $\widetilde{R}_0^N = \{(i,i), / i \in \{1,...,N\}\ \} = \Lambda_N$ (avec les notations du 1.)
- Ensuite, on regroupe à chaque génération les individus ayant les mêmes parents : $\widetilde{R}_{k+1}^N = \{(i,j),\ (i,j) \in \widetilde{R}_k^N$ ou bien i et j ont le même parent$\}$ *(on parle alors de coalescence)*

On retrouve bien l'idée du N-coalescent.

Pour montrer que le coalescent est une limite du modèle de Wright-Fisher, il est commode de s'intéresser au probabilités de transition du processus $(\widetilde{R}_k^N)$ (qui est une chaîne de Markov : la propriété de Markov est intuitive).

Si $\widetilde{R}_k^N$ possède i bloc, on s'intéresse aux coalescences possibles de i blocs :

- Les événements impliquant la coalescence simultanée de p individus, ou p coalescences de plusieurs individus ($p \geq 2$) sont peu probables : probabilité en $\circ\left(\frac{1}{N}\right)$
- La probabilité pour que deux blocs donnés coalescent est $\frac{1}{N} + \circ\left(\frac{1}{N}\right)$
- Et la probabilité pour qu'il n'y ait aucune coalescence est $1 - \frac{1}{N}\frac{i(i-1)}{2} + \circ\left(\frac{1}{N}\right)$ (car il y a $\frac{i(i-1)}{2}$ façon de coalescer deux blocs)

On a donc finalement la matrice de transition $\mathbf{P}_N$ de $(R_k^N)_{k\geq 0}$:

$$\mathbf{P}_N\left[\widetilde{R}_k^N = \eta \ \mid\ \widetilde{R}_{k-1}^N = \xi\right] = \begin{cases} 0 & \text{si} \quad \xi \not\subseteq \eta \\ 1 - \dfrac{i(i-1)}{2N} + \circ\left(\dfrac{1}{N}\right) & \text{si} \quad \xi = \eta \\ \dfrac{1}{N} + \circ\left(\dfrac{1}{N}\right) & \text{si} \quad \xi \prec \eta \\ \circ\left(\dfrac{1}{N}\right) & \text{sinon} \end{cases}$$

4.2 Le coalescent comme limite du modèle de Wright-Fisher

4.2.1 Approximation continue du modèle de Wright-Fisher

On souhaite trouver une échelle de temps appropriée pour étudier le modèle de Wright-Fisher (on veut par exemple que les coalescences aient lieu à vitesse 1 en moyenne)

Théorème 4.1. *Si $(\widetilde{R}_k^N)_{k\geq 0}$ est le processus de Markov décrit en 4.1 (modèle de Wright-Fisher), on a :*

$$\left(\widetilde{R}^N_{\lfloor tN\rfloor}\right)_{t\geq 0} \xrightarrow[N\to\infty]{(loi)} (R_t)_{t\geq 0}$$

où $(R_t)_{t\geq 0}$ est le coalescent de Kingman

Démonstration. :
On va montrer que pour tout $n \geq m \geq 1$,

$$\rho_m \widetilde{R}^N_{\lfloor tN\rfloor} \xrightarrow[N\to\infty]{(loi)} \rho_m R_t$$

où $(\rho_m R_t)$ est un m-coalescent. Etudions $\rho_m \widetilde{R}^N_k$, qui est la chaîne de Markov décrite dans 4.1 (modèle de Wright-Fisher), de matrice de transition $\rho_m \mathbb{P}_N$ (où ρ_m désigne la restriction de $\mathbf{P}_N$ à $(\mathbf{P}_N)_{1\leq i,j\leq m}$).
On remarque que

$$\rho_m \mathbf{P}_N = I_m + \frac{Q_m}{n} + \circ\left(\frac{1}{N}\right)$$

où Q_m est la matrice directrice du m-coalescent (voir la forme de $\mathbf{P}_N$ dans 4.1).

$$\begin{aligned}\mathbb{P}\left[\rho_m \widetilde{R}^N_{\lfloor t_2 N\rfloor} = \eta \;\middle|\; \rho_m \widetilde{R}^N_{\lfloor t_1 N\rfloor} = \xi\right] &= (\rho_m \mathbf{P}_N)^{\lfloor t_2 N\rfloor - \lfloor t_1 N\rfloor}(\xi,\eta) \\ &= \left(I_m + \frac{Q_m}{N} + \circ\left(\frac{1}{N}\right)\right)^{\lfloor t_2 N\rfloor - \lfloor t_1 N\rfloor}(\xi,\eta)\end{aligned}$$

Or, pour tous $t_1,\ t_2$ positifs, on a

$$\left(I_m + \frac{Q_m}{N} + \circ\left(\frac{1}{N}\right)\right)^{\lfloor t_2 N\rfloor - \lfloor t_1 N\rfloor} \xrightarrow[N\to\infty]{} e^{(t_1-t_2)Q_m}$$

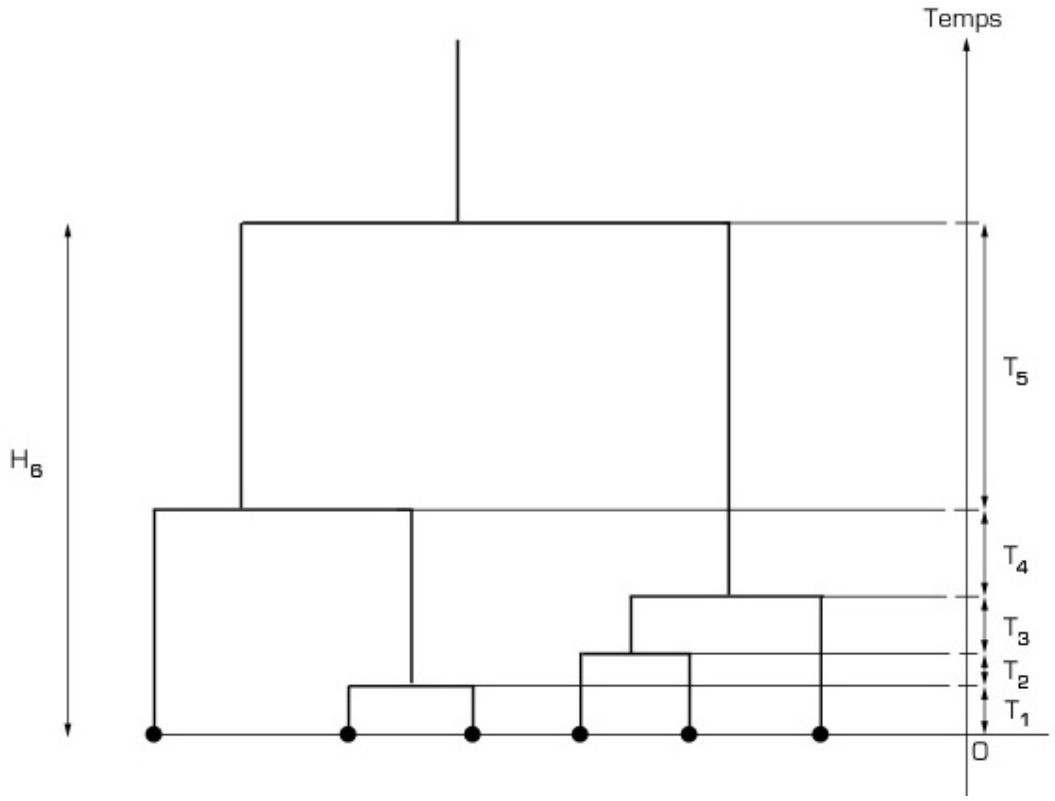
Ceci signifie, que quand n tend vers $+\infty$, les probabilités $\mathbb{P}\left[\rho_m \widetilde{R}^N_{\lfloor t_2 N\rfloor} = \eta \;\middle|\; \rho_m \widetilde{R}^N_{\lfloor t_1 N\rfloor} = \xi\right]$ tendent vers $\mathbb{P}\left[\rho_m R_{t_2} = \eta \;\middle|\; \rho_m R^n_{t_1} = \xi\right]$ où $(\rho_m R_t)$ est un m-coalescent. Ceci est vrai pour tout $t_1,\ t_2$ positifs, donc $(\rho_m \widetilde{R}^N_{\lfloor tN\rfloor})$ tend en loi vers un m-coalescent. □

4.2.2 Temps du plus récent ancêtre commun

On étudie maintenant le temps auquel il faut remonter pour trouver l'ancêtre commun d'une population de n personnes. On regarde donc un n-coalescent (on se restreint à une population de n individus, on a donc un n-coalescent), et on étudie le temps moyen auquel on rentre dans l'état Θ

(où il n'y a plus qu'un seul bloc). On note H_n ce temps (qui peut être considéré comme la hauteur de l'arbre généalogique) :

$$H_n = \sum_{k=0}^{n} \tau_k$$



On a facilement

$$\begin{aligned} \mathbb{E}\left[H_n\right] &= \sum_{k=2}^{n} \mathbb{E}\left[\tau_k\right] \\ &= \sum_{k=2}^{n} \frac{2}{k(k-1)} \\ &= 2 \times \left(1 - \frac{1}{n}\right) \end{aligned}$$

Il faut donc en moyenne remonter $2\left(1 - \frac{1}{n}\right)$ générations pour trouver l'ancêtre commun de n individus (rappel : on a changé l'échelle de temps, une unité de temps correspond à n générations). De plus, on sait que $\mathbb{E}[\tau_2] = 1$, donc pour deux individus fixés, il faut remonter en moyenne n

générations. Ainsi, la dernière coalescence est responsable, en moyenne, de plus de la moitié du temps total avant le plus récent ancêtre commun. De plus, $\mathbb{E}(H_n) \xrightarrow[n\to\infty]{} 2$, ce qui signifie que pour un échantillon de quelques k personnes prises au hasard, la moyenne de H_k est peu différente de celle de H_n (pour la population prise dans son entier) : cela veut dire que le plus récent ancêtre commun de cet échantillon est proche de celui de la population.
Remarque : par contre, la variance est assez élevée

$$\begin{aligned} Var\,(H_n) &= \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} \\ &\xrightarrow[n\to\infty]{} 4 \times \left(\frac{\pi^2}{3} - 3\right) \approx 1,16 \end{aligned}$$

4.2.3 Longeur totale des branches

Lorsque l'on s'intéresse aux mutations, on suppose qu'elles se produisent à taux constant (le nombre de mutation ne dépend que du temps écoulé, c'est-à-dire de la taille de la branche). Il faut donc étudier la longueur totale des branches, afin de connaître le nombre total de mutations.

Soit

$$L_n = \sum_{k=2}^{n} k\tau_k$$

qui représente la taille totale des branches (on a k branches de longueur τ_k).
Il est facile de calculer son espérance :

$$\begin{aligned} \mathbb{E}\,[L_n] &= \sum_{k=2}^{n} k\mathbb{E}\,[\tau_k] \\ &= 2\sum_{k=1}^{n} \frac{1}{k} \end{aligned}$$

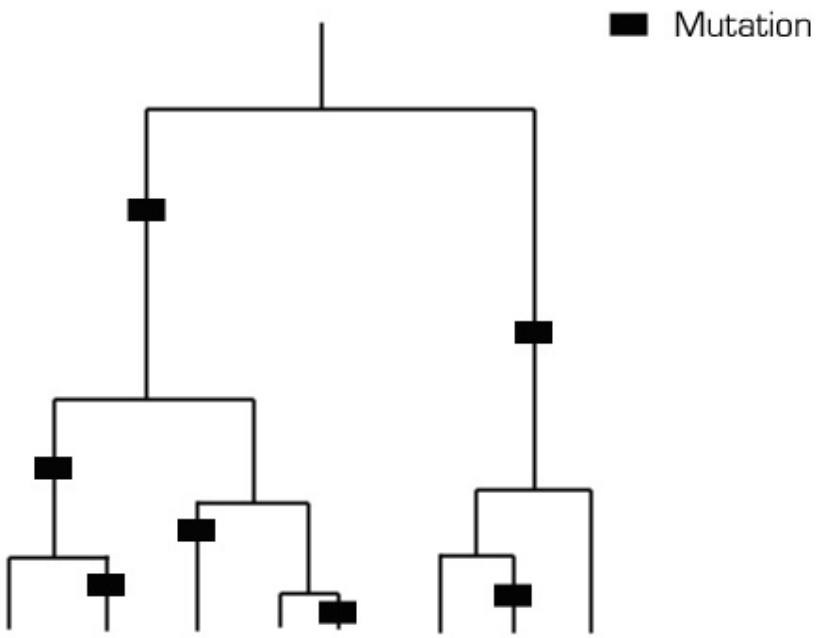
4.2.4 Un autre résultat intéressant

Un unique ancêtre

Dans la section 3. on a démontré que, dans le processus de coalescence, pour tout $t > 0$, il ne reste qu'un nombre fini de blocs, tous infinis. On dit que le coalescent de Kingman "descend de l'infini", ce qui permet d'affirmer que l'ensemble de la population (aussi grande soit-elle), descend d'un unique ancêtre commun (il n'y a qu'un nombre fini d'ancêtres pout tout temps $t > 0$)

4.3 Coalescent avec mutation

On suppose tout d'abord que les mutations sont neutres (il n'y a pas de sélection, et elles n'affectent pas la généalogie). En d'autres termes, le processus de mutation est indépendant du processus généalogique. Il suffit en fait, pour obtenir le modèle souhaité, de considérer les généalogies, et d'y ajouter ensuite des mutations.



On admet :
- que la fréquence de mutation par locus et par génération est faible
- que le nombre de générations considéré est grand
- que les mutations se font à taux constant μ : la probabilité qu'il y ait une mutation dans un temps t est μt

Sachant μ le taux de mutation, le nombre k de mutations après t générations suit une loi de Poisson :

$$\mathbb{P}(k|t) = e^{-\mu t}\frac{(\mu t)^k}{k!}$$

(Calculs en appendice)

Estimation du nombre de différences dans les séquences d'ADN :
On note S_n le nombre de différences dans les séquences d'ADN de n individus. On fait l'hypothèse que les mutations n'apportent que de nouveaux allèles (on ne retombe pas sur un allèle existant), ainsi, le nombre de différences dans les séquences d'ADN est égal au nombre de mutations survenues dans la généalogie. Rappelons qu'une unité de temps du coalescent correspond à N générations du modèle de Wright-Fisher (N est la taille de la population totale).
Calculons l'espérance de S_n :

$$\begin{aligned}\mathbb{E}[S_n] &= \mathbb{E}\left[\mathbb{E}[S_n \mid L_n]\right] \\ &= \mathbb{E}[N\mu L_n] \quad \text{(car on a } \mu \text{ mutation par unité de temps, donc } N\mu L_n \text{ mutations au total)} \\ &= 2N\mu \sum_{k=0}^{n-1} \frac{1}{k}\end{aligned}$$

On pose $\theta = 2N\mu$, qui correspond à la diversité de la population (c'est le nombre moyen de différences entre deux individus pris au hasard dans la population)
On peut calculer complètement la loi de S_2, le nombre de différences entre deux individus.

$$\begin{aligned}
\mathbb{P}[S_2 = k] &= \int_0^\infty \mathbb{P}[S = k \mid L_2 = t]\frac{e^{-t/2}}{2}dt \quad (\text{car } L_2 = 2\tau_2, \text{ avec } \tau_2 \text{ de loi } \mathbf{e}(1), \text{ donc } L_2\text{est de loi } \mathbf{e}(1/2)) \\
&= \int_0^\infty \frac{\left(\frac{\theta t}{2}\right)^k e^{-\frac{\theta t}{2}}}{k!}\frac{e^{-t/2}}{2}dt \quad (\text{car } \mathbb{P}[S = k \mid L_2 = t] \text{ suit une loi de Poisson de paramètre } \frac{\theta t}{2}) \\
&= \frac{\theta^k}{2^{k+1}}\int_0^\infty \frac{e^{-\frac{(\theta+1)t}{2}}}{k!}dt \\
&= \frac{\theta^k}{(\theta+1)^{k+1}} \qquad (\text{l'intégrale se calcule par IPP successives})
\end{aligned}$$

On a donc une loi géométrique de paramètre $\frac{1}{\theta+1}$. Ce résultat s'avère assez utile en pratique, car il permet d'estimer le taux de mutation μ, ou de voir si le taux que l'on suppose est vraisemblable.

On peut donner ici un résultat intéressant pouvant être obtenu à partir de l'étude du coalescent : on cherche à savoir combien d'allèles différents on est censé observer parmi n individus choisis au hasard, et avec quelle fréquence ils apparaissent. Si Π_θ est la partition dont les blocs sont les individus possédant les mêmes allèles, on peut donner la distribution exacte de Π_θ : on prend ξ une partition de {1,...,n} avec m_1 blocs de taille 1, ... , m_n blocs de taille n, et k blocs en tout, on a alors

$$\mathbb{P}[\Pi_\theta = \xi] = \frac{\theta^k}{\theta(\theta+1)...(\theta+n-1)}\prod_{i=1}^{k-1}(m_i - 1)!$$

C'est la formule d'échantillonnage d'Ewens, qui permet, en observant la fréquence des allèles de connaître le taux de mutation μ avec une bonne précision.

Conclusion

Nous avons donc étudié dans ce mémoire le processus de coalescent de Kingman, dont les valeurs sont les partitions de $\mathbb{N}$. Nous nous sommes alors intéressés au partitions aléatoires échangeables, et notamment au fait que toute partition échangeable est en loi égal à un processus dit de "boîte de peinture". Nous avons alors été en mesure de construire un tel coalescent. Ce processus aléatoire possède certaines applications intéressantes en biologie (notamment en généalogie) : il permet en effet d'estimer l"âge" moyen de l'ancêtre commun d'un échantillon de la population, et le taux de mutation à partir des différences entre les séquences d'ADN, ou à partir des fréquences des allèles rencontrés. Mais il ne faut pas oublier que nous avons fait quelques hypotheses simplificatrices, et si l'on veut pouvoir s'en passer, il faudra encore un peu travailler pour trouver des modèles et les étudier.

5 Appendice

5.1 Calcul de la loi de la "boîte de peinture uniforme"

On suppose maintenant que R est de loi $\mathcal{P}_k$, et on veut calculer la loi de sa restriction $\rho_n R$: Si $\xi \in \mathcal{E}, |\xi| = l \leq k$, on a

$$\begin{aligned}\mathbb{P}\left[\rho_n R = \xi\right] &= \int_{\Delta_k} \cdots \int P^x \left[\rho_n R = \xi\right] (k-1)! dx_1 ... dx_{k-1} \\ &= \int_{\Delta_k} \cdots \int \mathbb{P}\left[Y_i = Y_j \Leftrightarrow (i,j) \in \xi, 1 \leq i,j \leq n\right] (k-1)! dx_1 ... dx_{k-1}\end{aligned}$$

où (Y_i) est i.i.d de loi $\mathbb{P}\left[Y_i = k\right] = x_k$ donne la "boîte de peinture" de loi $P^{(0,x_1,...,x_k,0,...)}$.
De plus

$$\begin{aligned}\mathbb{P}\left[Y_i = Y_j \Leftrightarrow (i,j) \in \xi, 1 \leq i,j \leq n\right] &= \sum_{\substack{(r_1,r_2,...,r_n)\in\{1,...,k\} \\ r_i = r_j \Leftrightarrow (i,j)\in\xi}} \mathbb{P}\left[Y_1 = r_1, Y_2 = r_2, ..., Y_n = r_n\right] \\ &= \sum_{\substack{(r_1,r_2,...,r_n)\in\{1,...,k\} \\ r_i = r_j \Leftrightarrow (i,j)\in\xi}} x_{r_1} x_{r_2} ... x_{r_n}\end{aligned}$$

on notera par la suite $\sum^{\xi}$ pour alléger l'indice de sommation, et $\lambda_1, \lambda_2, ..., \lambda_l$ les tailles des blocs de ξ. Le calcul donne donc

$$\begin{aligned}\mathbb{P}\left[\rho_n R = \xi\right] &= \int_{\Delta_k} \cdots \int \sum^{\xi} x_{r_1} x_{r_2} ... \ x_{r_n} (k-1)! dx_1 ... dx_{k-1} \\ &= (k-1)! \times \sum^{\xi} \int_{\Delta_k} \cdots \int x_{r_1} x_{r_2} ... x_{r_n} dx_1 ... dx_{k-1}\end{aligned}$$

Calculons l'intégrale du membre de droite.

$$\begin{aligned}&\int_{\Delta_k} \cdots \int x_{r_1} x_{r_2} ... x_{r_n} dx_1 ... dx_{k-1} \\ &= \int_{\Delta_k} \cdots \int x_{r_{i_1}}^{\lambda_{i_1}} x_{r_{i_2}}^{\lambda_{i_2}} ... \ x_{r_{i_l}}^{\lambda_{i_l}} dx_1 ... dx_{k-1} \\ &= \int_{\Delta_k} \cdots \int x_1^{\lambda_1} x_2^{\lambda_2} ... \ x_l^{\lambda_l} dx_1 ... dx_{k-1} \\ &\quad \text{(Par Fubini-Tonelli, les } x_{r_i} \text{ jouent des rôles symétriques)} \\ &= \int_0^1 \int_0^{1-x_1} \int_0^{1-(x_1+x_2)} \cdots \int_0^{1-(\sum_{i=0} k-2x_i)} x_1^{\lambda_1} x_2^{\lambda_2} ... \ x_l^{\lambda_l} dx_1 ... dx_{k-1}\end{aligned}$$

On va montrer (par induction) que

$$\int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=0}^{k-2} x_i} x_1^{\lambda_1} x_2^{\lambda_2} ... \, x_l^{\lambda_l} dx_1 ... dx_{k-1} = \frac{\lambda_1! \lambda_2! ... \lambda_l!}{(k-1+\lambda_1+...+\lambda_l)!}$$

Notons que

$$\int_0^{1-y} x^\alpha (y-x)^\beta dx = \frac{\alpha! \beta!}{(\beta+\alpha+1)!} y^{\beta+\alpha+1}$$

Un peu de calcul...

$$\begin{aligned}
&\int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=0} k-2x_i} x_1^{\lambda_1} x_2^{\lambda_2} ... \, x_l^{\lambda_l} dx_1 ... dx_{k-1} \\
&= \int_0^1 \cdots \int_0^{1-\sum_{i=0}^{k-3} x_i} x_1^{\lambda_1} ... \, x_l^{\lambda_l} \left((1 - \sum_{i=0}^{k-2} x_i) - x_{k-2} \right) dx_1 ... dx_{k-2} \\
&= \frac{1}{2} \int_0^1 \cdots \int_0^{1-(\sum_{i=0}^{k-4} x_i)} x_1^{\lambda_1} ... \, x_l^{\lambda_l} \left(1 - \sum_{i=0}^{k-3} x_i - x_{k-2} \right)^2 dx_1 ... dx_{k-3} \\
&= \frac{1}{3!} \int_0^1 \cdots \int_0^{1-(\sum_{i=0}^{k-5} x_i)} x_1^{\lambda_1} ... \, x_l^{\lambda_l} \left(1 - \sum_{i=0}^{k-4} x_i - x_{k-3} \right)^2 dx_1 ... dx_{k-4} \\
&\vdots \\
&= \frac{1}{(k-1-l)!} \int_0^1 \cdots \int_0^{1-(\sum_{i=0}^{l-1} x_i)} x_1^{\lambda_1} ... \, x_l^{\lambda_l} \left(1 - \sum_{i=0}^{l-1} x_i - x_l \right)^{k-l} dx_1 ... dx_l \\
&= \frac{\lambda_l!}{(k-l+\lambda_l+1)!} \int_0^1 \cdots \int_0^{1-(\sum_{i=0}^{l-2} x_i)} x_1^{\lambda_1} ... \, x_{l-1}^{\lambda_{l-1}} \left((1 - \sum_{i=0}^{l-2} x_i) - x_{l-1} \right)^{k-l+\lambda_l+1} dx_1 ... dx_{l-1} \\
&= \frac{\lambda_l! \lambda_{l-1}!}{(k-l+\lambda_l+\lambda_{l-1}+2)!} \int_0^1 \cdots \int_0^{1-(\sum_{i=0}^{l-3} x_i)} x_1^{\lambda_1} ... \, x_{l-2}^{\lambda_{l-2}} \left((1 - \sum_{i=0}^{l-3} x_i) - x_{l-2} \right)^{k-l+\lambda_l+\lambda_{l-1}+2} dx_1 ... dx_{l-3} \\
&\vdots
\end{aligned}$$

qui donne le résultat voulu par induction. Récapitulons :

$$\begin{aligned}
\mathbb{P}\left[\rho_n R = \xi\right] &= (k-1)! \sum^{\xi} \frac{\lambda_1! \lambda_2! ... \lambda_l!}{(k-1+\lambda_1+...+\lambda_l)!} \\
&= N(\xi) \times (k-1)! \frac{\lambda_1! \lambda_2! ... \lambda_l!}{(k-1+n)!}
\end{aligned}$$

Où $N(\xi)$ est le nombre de $(r_1, r_2, ..., r_n) \in \{1, ..., k\}$ tels que $r_i = r_j \Leftrightarrow (i,j) \in \xi$. On choisit les l valeurs que prennent les r_i (il y a $\binom{k}{l}$ possibilités), puis on choisit quelle valeur (parmi ces l) on

assigne à un représentant de chaque classe (l! possibilités). Ainsi

$$\mathbb{P}\left[\rho_n R = \xi\right] = \mathcal{P}_k\left(\rho_n^{-1}(\xi)\right) = \frac{k!(k-1)!\lambda_1!\lambda_2!...\lambda_l!}{(k-l)!(k-1+n)!}$$

On peut aussi calculer la probabilité que le nombre de blocs de $\rho_n R$ soit k sous $\mathcal{P}_k$.
Rappel : pour le processus de saut $(\mathcal{R}_k)_{k\leq n}$ d'un n-coalescent

$$\mathbb{P}\left[\mathcal{R}_k = \xi\right] = \frac{(n-k)!k!(k-1)!}{n!(n-1)!}$$

Ce qui donne

$$\begin{aligned}
\mathbb{P}\left[|\mathcal{R}_k| = k\right] &= 1 \\
&= \sum_{|\xi|=k} \mathbb{P}\left[\mathcal{R}_k = \xi\right] \\
&= \frac{(n-k)!k!(k-1)!}{n!(n-1)!} \times \sum_{|\xi|=k} \lambda_1!\lambda_2!...\lambda_k! \\
\text{d'où } \sum_{|\xi|=k} \lambda_1!\lambda_2!...\lambda_k! &= \frac{n!(n-1)!}{(n-k)!k!(k-1)!}
\end{aligned}$$

On a donc

$$\begin{aligned}
\mathbb{P}\left[|\rho_n R| = k\right] &= \sum_{|\xi|=k} \mathbb{P}\left[|\rho_n R| = k\right] \\
&= \frac{k!(k-1)!}{(k-1+n)!} \sum_{|\xi|=k} \lambda_1!\lambda_2!...\lambda_l! \\
&= \frac{n!(n-1)!}{(n+k-1)!(n-k)!} \text{ sous } \mathcal{P}_k
\end{aligned}$$

$$\mathbb{P}\left[\mathcal{R}_k = \xi\right] = \frac{\mathbb{P}\left[\rho_n R = \xi\right]}{\mathbb{P}\left[|\rho_n R| = k\right]} \text{(avec les formules explicites de chaque probabilité)}$$

Ainsi, comme l'ensemble des valeurs possibles est fini, on a, pour tout $n \geq k$

$$\mathbb{P}\left[\mathcal{R}_k = \xi\right] = \mathbb{P}\left[\rho_n R = \xi \;\middle|\; |\rho_n R| = k\right]$$

5.2 La loi de Poisson pour les mutations

En supposant que les mutations se font à taux constant, on obtient une équation à partir de la réflexion suivante :
pour avoir k mutations à l'instant t+h (h suffisamment petit),

1. Soit on avait k-1 mutation à l'instant t, et il se produit une mutation dans un temps h

2. Soit on avait déjà k mutations, et il n'y a pas de mutation dans un temps h

On a donc

$$\begin{aligned} \mathbb{P}(k|t+h) &= \mu h \mathbb{P}(k-1|t) + (1-\mu h)\mathbb{P}(k|t) \\ \text{d'où } \frac{\mathbb{P}(k|t+h) - \mathbb{P}(k|t)}{h} &= \mu \mathbb{P}(k-1|t) - \mu \mathbb{P}(k|t) \end{aligned}$$

D'où, en faisant $h \to 0$, on obtient l'équation diférentielle

$$\frac{d\mathbb{P}(k|t)}{dt} = \mu \mathbb{P}(k-1|t) - \mu \mathbb{P}(k|t)$$

Or, $\mathbb{P}(k|t) = e^{-\mu t}\frac{(\mu t)^k}{k!}$ convient, donc on a le résultat (il suffit d'appliquer une récurrence : on détermine $\mathbb{P}(0|t)$, puis $\mathbb{P}(1|t)$, etc... et on a l'unicité de la solution C^1 par le théorème de Cauchy-Lipschitz,).

Donc le nombre k de mutations après un temps t ($\lfloor tN \rfloor$ générations) suit une loi de Poisson :

$$\mathbb{P}(k|t) = e^{-\mu t}\frac{(\mu t)^k}{k!}$$

Références

[1] J.F.C. Kingman. *The coalescent*. Stochastic Process. Appl. 13 (1982), no.3,235-248

[2] J.R. Norris. *Markov Chains*. Publications Cmbridge University Press, 1997

[3] D.J. Aldous, I.A. Ibragimov, J. Jacob. *Ecole d'été de probabilités de Saint-Flour : XIII - 1983*. Édition Springer-Verlag, 1985

[4] Sites Internet :

1. http ://wwwabi.snv.jussieu.fr/jompo/Public/OBI/OBI4/coalescence.pdf
2. http ://newton.mat.ulaval.ca/theses/JH-Smith-Lacroix_05.pdf