Introduction au domaine de recherche : Processus de fragmentation et arbres branchants Markoviens

Jean-Jil Duchamps

25 octobre 2016

Table des matières

1 Introduction

L'étude de la fragmentation aléatoire d'une masse au cours du temps permet d'appréhender de nombreux phénomènes physiques et biologiques. C'est en particulier l'application à la modélisation des mécanismes d'évolution en biologie qui nous intéresse. Dans ce sens, Aldous [2] énonce des hypothèses générales que doivent vérifier les modèles d'arbres phylogénétiques, ce qui donne lieu à une riche littérature au sujet des arbres branchants Markoviens (*Markov branching trees*) [5–7]. Ces modèles se trouvent être liés à des processus dits de fragmentations aléatoires, à valeurs dans l'espace des partitions de $\mathbb{N}$, dont on donne un aperçu ici (on réfère à Bertoin [4] pour une description complète de la théorie). L'hypothèse, naturelle pour les applications, qui fonde l'étude des fragmentations, est une propriété de

branchement : chaque fragment se divise indépendamment des autres et toujours d'une même manière. De telles fragmentations aléatoires sont identifiées à travers de nombreux autres processus stochastiques, en particulier ceux, comme le mouvement Brownien, pouvant être interprétés comme des arbres réels, cf. [1,3,10,11] pour quelques exemples.

On se base en grande partie sur le livre de Bertoin, qui couvre aussi une théorie duale des fragmentations, la théorie des processus de coagulation, très importante en génétique des populations. Dans la section 2, on définira la notion de fragmentations homogènes de partitions des entiers, et l'on évoquera les notions d'échangeabilité et de compatibilité par échantillonnage ; dans la section 3 on montrera le lien qui existe entre les fragmentations et les arbres branchants Markoviens introduits par Aldous, c'est-à-dire les modèles d'arbres aléatoires discrets construits à partir d'hypothèses d'échangeabilité et de compatibilité ; finalement dans la section 4 on donnera un aperçu d'une piste de recherche envisageable qui découle de la théorie des fragmentations.

2 Fragmentations de partitions échangeables

On s'intéresse dans cette section aux fragmentations des partitions des entiers, il nous faut donc aborder la notion de partition aléatoire échangeable due à Kingman [9].

2.1 Partitions échangeables

Pour $n \in \mathbb{N} \cup \{\infty\}$, une partition $\pi = \{A_i, i \in I\}$ est un ensemble de parties de $[n] := \{k \in \mathbb{N},\ 1 \leq k \leq n\}$, non vides et disjointes deux à deux, tel que $\bigcup_{i\in I} A_i = [n]$: Notons $\mathcal{P}_n$ l'ensemble des partitions de $[n]$.

$$\mathcal{P}_n := \{\pi \text{ partition de } [n]\}.$$

Étant fixés $n < m \leq \infty$, on définit la restriction d'une partition $\pi \in \mathcal{P}_m$ à une partition de $[n]$:

$$\pi_{|[n]} = \{A \cap [n],\ A \in \pi, A \cap [n] \neq \emptyset\} \in \mathcal{P}_n.$$

Ceci permet de définir une distance d sur $\mathcal{P}_\infty$, par :

$$d(\pi, \pi') := \left(\sup\{n \geq 1,\ \pi_{|[n]} = \pi'_{|[n]}\}\right)^{-1},$$

avec par convention $(\sup \mathbb{N})^{-1} = \infty^{-1} = 0$. Cette distance fait de $(\mathcal{P}_\infty, d)$ un espace métrique compact, donc en particulier complet et séparable.

On peut donc commencer à considérer les partitions aléatoires, et on définit les partitions échangeables comme celles dont la loi est invariante sous l'action naturelle des permutations. Pour $n \in \mathbb{N} \cup \{\infty\}$, pour $\pi \in \mathcal{P}_n$ et σ une permutation de $[n]$, c'est-à-dire une bijection de $[n]$ dans lui-même, soit $\sigma(\pi)$ la partition définie par

$$\sigma(\pi) := \{\sigma(A),\ A \in \pi\}.$$

Autrement dit, si $\sim^\pi$ est la relation d'équivalence associée à la partition π (i.e. $i \sim^\pi j \iff \exists A \in \pi,\ i,j \in A$), alors $\sigma(\pi)$ est donnée par la relation :

$$i \sim^{\sigma(\pi)} j \iff \sigma^{-1}(i) \sim^\pi \sigma^{-1}(j).$$

On peut enfin définir le concept d'échangeabilité pour les partitions aléatoires.

Définition 1. Une **partition échangeable** est une partition aléatoire Π telle que pour toute permutation σ, on ait l'égalité en loi :

$$\sigma(\Pi) \overset{\mathcal{L}}{=} \Pi.$$

2.1.1 Boîtes de peinture

Un exemple de partition échangeable est une partition donnée par un procédé de boîtes de peinture (*paintbox construction* introduite par Kingman dans la section 3 de [9]). Définissons d'abord l'ensemble des partitions de masse $\mathcal{P}_\mathtt{m}$.

$$\mathcal{P}_\mathtt{m} := \left\{ \mathbf{s} = (s_1, s_2, \ldots) \in [0,1]^\mathbb{N},\ s_1 \geq s_2 \geq \ldots,\ \textstyle\sum_{n\geq 1} s_n \leq 1 \right\}.$$

Un élément $\mathbf{s}$ de $\mathcal{P}_\mathtt{m}$ représente la "fragmentation" de l'intervalle $[0,1]$ en intervalles indistinguables de longueurs $s_1, s_2, \ldots$, de sorte qu'une telle représentation par une suite décroissante est unique. Notons que $\mathcal{P}_\mathtt{m}$ muni de la distance uniforme est compact.

Étant donné une partition de masse $\mathbf{s} = (s_1, s_2, \ldots) \in \mathcal{P}_\mathtt{m}$, on pose $t_i = \sum_{j=1}^i s_j$ pour $i \geq 0$. Notons que $t_\infty := \lim_{i\to\infty} t_i \leq 1$ par définition, l'inégalité pouvant être stricte. Soit maintenant une suite $(U_n)_{n\geq 1}$ *i.i.d.* de variables uniformes sur $[0,1]$. On définit une partition aléatoire π en posant

$$n \sim^\pi m \iff n = m \text{ ou } \exists i \geq 1,\ U_n, U_m \in]t_{i-1}, t_i].$$

On peut vérifier que cette partition est échangeable. Notons que par la loi des grands nombres, chaque bloc de la partition possède une fréquence asymptotique presque sûrement. C'est-à-dire que si l'on note B_1 le bloc de π contenant 1, puis par récurrence, pour $i \geq 1$, B_{i+1} le bloc de π contenant le premier entier qui n'est pas dans $B_1 \cup \ldots \cup B_i$, alors on a pour tout i :

$$|B_i| := \lim_{n\to\infty} \#B_i \cap [n] = \lim_{n\to\infty} \sum_{k=1}^n \mathbb{1}_{k\in B_i} \quad \text{existe } p.s.$$

En particulier chaque bloc est soit infini, avec une fréquence asymptotique égale à l'un des $t_i - t_{i-1} > 0$, soit un singleton, ce qui correspond au cas où U_n tombe dans l'intervalle $[t_\infty, 1]$, ce qui implique que n n'est en relation qu'avec lui-même. De plus, le réordonnement décroissant des fréquences asymptotiques est égal à $\mathbf{s}$ presque sûrement. On appelle $\rho_\mathbf{s}$ la loi d'une partition π créée avec ce procédé de boîtes de peinture. Si $\mathbf{s}$ est en plus également aléatoire (indépendante de la suite $(U_n)_{n\geq 1}$), c'est-à-dire tirée selon une loi ν sur $\mathcal{P}_\mathtt{m}$, alors on note la loi de π ainsi obtenue ρ_ν.

$$\rho_\nu(\,\cdot\,) = \int_{\mathcal{P}_\mathtt{m}} \rho_\mathbf{s}(\,\cdot\,)\, \nu(\mathrm{d}\mathbf{s}).$$

Un théorème de Kingman montre que toute partition échangeable est obtenue avec un procédé de boîtes de peinture comme on vient de le décrire.

Théorème 2.1 (Kingman [9]). *Toute mesure de probabilité sur $\mathcal{P}_\infty$ qui est invariante par l'action des permutations a la forme ρ_ν, pour ν une probabilité sur $\mathcal{P}_{\mathtt{m}}$.*

Une preuve de ce résultat utilise la notion d'échangeabilité pour les suites de variables aléatoires et le théorème de de Finetti.

2.2 Fragmentations échangeables homogènes

2.2.1 Définitions

Pour définir le processus de fragmentation d'une partition échangeable, on définit l'opérateur Frag.

Définition 2. Soit $\pi \in \mathcal{P}_\infty$ une partition, et $\pi^{(\cdot)} = (\pi^{(i)})_{i\geq 1}$ une suite de partitions. On note π_j le j-ième bloc de la partition π. On note alors $\mathrm{Frag}(\pi, \pi^{(\cdot)})$ la partition donnée par les intersections des blocs de π avec les partitions de la suite $\pi^{(\cdot)}$:

$$\mathrm{Frag}(\pi, \pi^{(\cdot)}) := \left\{\pi_i \cap \pi_j^{(i)},\ i, j \geq 1\right\} \setminus \{\emptyset\}.$$

On peut remarquer que si π est une partition aléatoire échangeable, et $\pi^{(\cdot)}$ une suite *i.i.d.* de partitions échangeables et indépendante de π, alors la partition $\mathrm{Frag}(\pi, \pi^{(\cdot)})$ est encore une partition échangeable. Ceci nous permet de définir les processus de fragmentation homogènes.

Définition 3. Un processus de Markov $\Pi = (\Pi(t))_{t\geq 0}$ à trajectoires presque sûrement continues à droite, admettant des limites à gauches (*càdlàg*) à valeurs dans $\mathcal{P}_\infty$ est un **processus de fragmentation homogène** si, pour $t, t' \geq 0$, la loi conditionnelle de $\Pi(t+t')$ sachant $\Pi(t)$ est la loi de $\mathrm{Frag}(\Pi(t), \pi^{(\cdot)})$, où $\pi^{(\cdot)}$ est une suite *i.i.d.* de partitions échangeables dont la loi ne dépend que de t'.

Un processus de fragmentation Π est dit **standard** si $\Pi(0) = \{\mathbb{N}\}$ presque sûrement.

Parmi les propriétés des processus de fragmentation, on remarque en particulier :

1. Si $\Pi^{(\cdot)}$ est une suite *i.i.d.* de copies d'un processus de fragmentation homogène standard, alors pour $\pi \in \mathcal{P}_\infty$, le processus $(\mathrm{Frag}(\pi, \Pi^{(\cdot)}(t)))_{t\geq 0}$ est une version du même processus de fragmentation ayant pour valeur initiale π.
2. Un processus de fragmentation est un processus de Feller, c'est-à-dire que son semi-groupe $(P_t)_{t\geq 0}$, défini sur $\mathcal{C}(\mathcal{P}_\infty) = \{\phi : \mathcal{P}_\infty \to \mathbb{R},\ \phi \text{ continue}\}$ par

   $$P_t\phi(\pi) = \mathbb{E}\left[\phi(\mathrm{Frag}(\pi, \Pi^{(\cdot)}(t)))\right],$$

   avec les notations du point précédent, vérifie :

   $$P_t\phi \in \mathcal{C}(\mathcal{P}_\infty) \quad \text{et} \quad \lim_{t\to 0} \|P_t\phi - \phi\|_{\sup} = 0.$$

   La propriété de Feller implique la propriété de Markov forte, c'est-à-dire que si T est un temps d'arrêt pour la filtration naturelle $(\mathcal{F}_t)_{t\geq 0}$ du processus $(\Pi(t))_{t\geq 0}$, alors la loi conditionnelle du processus $(\Pi(T+t))_{t\geq 0}$ sachant $\mathcal{F}_T$ est la loi du même processus de fragmentation ayant $\Pi(T)$ pour valeur initiale.

2.2.2 Construction par un processus ponctuel

Un résultat central de la théorie des fragmentations est le théorème qui donne une caractérisation des fragmentations homogènes par une mesure échangeable sur $\mathcal{P}_\infty$. Pour le comprendre, on introduit un procédé (déterministe) de construction de fonctions càdlàg à valeurs dans les partitions $\Pi(t)$ à partir d'une mesure ponctuelle (c'est-à-dire un ensemble de points). Soit $\mathcal{N}$ un sous ensemble de $\mathbb{R}_+ \times \mathcal{P}_\infty \times \mathbb{N}$ tel que, pour tout temps $t > 0$ et tout entier $n \geq 1$,

$$\#\{(s, \pi, k) \in \mathcal{N},\ s \in [0, t],\ \pi \neq \mathbf{1}_{[n]}, k \leq n\} < \infty,$$

où $\mathbf{1}_{[n]} := \{[n]\}$ désigne la partition contenant un seul bloc. On suppose aussi que $\mathcal{N}$ n'a au plus qu'un élément par fibre de temps, c'est-à-dire que pour tout $t \geq 0$, on a

$$\#\left(\mathcal{N} \cap (\{t\} \times \mathcal{P}_\infty \times \mathbb{N})\right) \in \{0, 1\}.$$

Un processus Π à valeurs dans $\mathcal{P}_\infty$ peut être défini à partir d'un tel ensemble $\mathcal{N}$, de la manière suivante. Pour $n \geq 1$, soient $0 < t_1 < t_2 < \ldots$ l'ensemble des temps auxquels sont associés les éléments (t_i, π_i, k_i) de $\mathcal{N}$ tels que $\pi_i \neq \mathbf{1}_{[n]}$ et $k_i \leq n$. On définit alors le processus Π^n à valeurs dans $\mathcal{P}_n$ comme constant sur les intervalles $[t_i, t_{i+1}[$, avec $\Pi^n(0) := \mathbf{1}_{[n]}$, et

$$\Pi^n(t_i) = \mathrm{Frag}_{k_i}(\Pi^n(t_{i-1}), \pi_i),$$

où par abus de langage, $\mathrm{Frag}_k(\pi, \pi')$ désigne la fragmentation de k-ième bloc de π par π'. Autrement dit, avec notre définition précédente, c'est $\mathrm{Frag}(\pi, (\mathbf{1}_{[n]}, \mathbf{1}_{[n]}, \ldots, \pi', \ldots))$, où π' est en k-ième position.

Alors on peut vérifier que les processus $(\Pi^n)_{n\geq 1}$ sont compatibles entre eux, c'est-à-dire que pour $1 \leq n < m$, on a

$$\Pi^m_{|[n]} = \Pi^n.$$

Ceci permet donc de définir le processus Π à valeurs dans $\mathcal{P}_\infty$ comme limite projective, i.e. comme l'unique processus tel que pour tout $n \geq 1$, $\Pi_{|[n]} = \Pi^n$. Le fait que tous les processus Π^n soient constants par morceaux (sur des intervalles de type $[t_i, t_{i+1}[$), donc càdlàg, implique de plus que le processus Π lui-même est càdlàg.

2.2.3 Représentation Poissonnienne des fragmentations

Une façon naturelle de construire une fragmentation aléatoire est de rendre l'ensemble $\mathcal{N} \subset \mathbb{R}_+ \times \mathcal{P}_\infty \times \mathbb{N}$ aléatoire, par exemple en lui donnant la loi d'un processus ponctuel de Poisson d'intensité $\mathrm{d}t \otimes \mu \otimes \#$, où $\mathrm{d}t$ est la mesure de Lebesgue, $\#$ est la mesure de comptage, et μ est une mesure (éventuellement de masse infinie) échangeable telle que

$$\mu(\{\mathbf{1}_{[\infty]}\}) = 0, \quad \text{et} \quad \mu\left(\pi_{|[n]} \neq \mathbf{1}_{[n]}\right) < \infty \text{ pour tout } n \geq 1. \tag{1}$$

Cette hypothèse implique que $\mathcal{N}$ a presque sûrement une forme qui permet de construire Π comme il vient d'être construit de manière déterministe. On peut en fait construire toute fragmentation homogène de cette manière.

Théorème 2.2. *(i) Le processus* Π *précédemment construit à partir de la mesure* μ *est une fragmentation homogène.*

(ii) Toute fragmentation homogène a la loi d'une telle construction pour une unique mesure μ *sur* $\mathcal{P}_\infty$ *qui vérifie* (1)*. Cette mesure est échangeable et est appelée la* ***mesure caractéristique*** *de la fragmentation.*

Une idée de preuve pour le deuxième point est d'observer les taux de transition des chaînes de Markov à espace discret $\Pi_{|[n]}$. Ceux-ci sont compatibles entre eux au sens où, si q^n_π désigne le taux de transition de $\mathbf{1}_{[n]}$ vers π, alors, pour $n < m$ et $\pi \in \mathcal{P}_n$,

$$q^n_\pi = \sum_{\pi' \in \mathcal{P}_m,\, \pi'_{|[n]} = \pi} q^m_{\pi'}.$$

On en déduit l'existence de la mesure μ, qui peut être identifiée comme la mesure de la construction Poissonnienne (cf. Bertoin [3] pour la preuve).

On peut décomposer plus encore la mesure caractéristique μ en deux termes, un terme d'*érosion* et un terme de *dislocation*. Pour $n \geq 1$, on définit la partition $\mathfrak{e}_n \in \mathcal{P}_\infty$ comme la partition qui isole seulement n :

$$\mathfrak{e}_n := \{\{n\}, \mathbb{N} \setminus \{n\}\}.$$

On peut donc définir la mesure d'érosion $\mathfrak{e}$, échangeable, de masse infinie, mais qui vérifie bien (1) :

$$\mathfrak{e} := \sum_{n \geq 1} \delta_{\mathfrak{e}_n}.$$

Pour construire le terme de dislocation, on rappelle le procédé de boîtes de peinture grâce aux mesures $\rho_{\mathbf{s}}$, pour $\mathbf{s} \in \mathcal{P}_{\mathtt{m}}$. Soit ν une mesure sur $\mathcal{P}_{\mathtt{m}}$ qui vérifie

$$\nu(\{(1, 0, \ldots)\}) = 0, \quad \text{et} \quad \int_{\mathcal{P}_{\mathtt{m}}} (1 - s_1)\, \nu(\mathrm{d}\mathbf{s}) < \infty. \tag{2}$$

Alors la mesure sur $\mathcal{P}_\infty$ définie par

$$\rho_\nu(\,\cdot\,) := \int_{\mathcal{P}_{\mathtt{m}}} \rho_{\mathbf{s}}\ (\,\cdot\,), \nu(\mathrm{d}\mathbf{s})$$

est échangeable et vérifie (1), car

$$\begin{aligned}
\rho_\nu\left(\pi_{|[n]} \neq \mathbf{1}_{[n]}\right) &\leq \sum_{k=2}^{n} \rho_\nu(1 \nsim k) \\
&= (n-1)\rho_\nu(1 \nsim 2) \\
&= (n-1) \int_{\mathcal{P}_{\mathtt{m}}} \Big(s_0 + \sum_{i \geq 1} s_i(1 - s_i)\Big) \nu(\mathrm{d}\mathbf{s}) \\
&\leq 2(n-1) \int_{\mathcal{P}_{\mathtt{m}}} (1 - s_1) \nu(\mathrm{d}\mathbf{s})
\end{aligned}$$

On peut donc considérer ρ_ν en tant que mesure caractéristique de fragmentation homogène, et la réciproque est donnée par le théorème suivant.

Théorème 2.3. *Soit μ une mesure échangeable sur $\mathcal{P}_\infty$ qui vérifie (1). Alors il existe un unique réel $c \geq 0$ et une unique mesure ν sur $\mathcal{P}_\mathbb{m}$ qui vérifie (2) tels que*

$$\mu = c\mathfrak{e} + \rho_\nu.$$

*Ainsi, un processus homogène de fragmentation Π est caractérisé de manière unique par son **taux d'érosion** c et sa **mesure de dislocation** ν.*

L'érosion correspond à la séparation de singletons des blocs, ce qui cause une diminution continue de la taille des blocs. On peut comprendre ce phénomène de la manière suivante : à chaque entier n indépendamment est associé un temps exponentiel de paramètre c le taux d'érosion, au bout duquel le singleton $\{n\}$ se détache du bloc auquel il appartenait. En conséquence, un processus de fragmentation avec un taux d'érosion $c \geq 0$ et une mesure de dislocation nulle, issu de la partition $\mathbf{1}_{[\infty]}$, est constitué d'un bloc infini unique de fréquence asymptotique $|\Pi(t)| = \mathrm{e}^{-ct}$. La dislocation en revanche, est le terme qui correspond à la fragmentation spontanée d'un bloc en plusieurs, ce qui cause les sauts du processus des fréquences asymptotiques.

2.2.4 Arbres de fragmentation

On définit ici une suite d'arbres discrets compatibles entre eux associés à une fragmentation. Pour un processus standard $\Pi = (\Pi(t))_{t\geq 0}$ de fragmentation on sait que, pour tout $n \geq 1$, le processus $\Pi^n := \Pi_{|[n]}$ est absorbé en $\mathbf{0}_{[n]} := \{\{1\}, \ldots, \{n\}\}$ au bout d'un temps assez grand. On peut alors définir l'ensemble de sous-ensembles de $[n]$:

$$T_n = \bigcup_{t\geq 0} \Pi^n(t).$$

Cet ensemble contient $[n]$ car $\Pi^n(0) = \{[n]\}$, et vu la remarque qui vient d'être faite, T_n contient aussi chaque singleton $\{i\}$, pour $i \in [n]$. Partiellement ordonné par la relation $\subset$, l'ensemble T_n peut-être représenté de manière unique par un arbre discret, de racine correspondant à l'élément maximal unique $[n]$ et aux feuilles correspondant aux éléments minimaux $\{1\}, \ldots, \{n\}$ (cf. figure 1 pour un exemple).

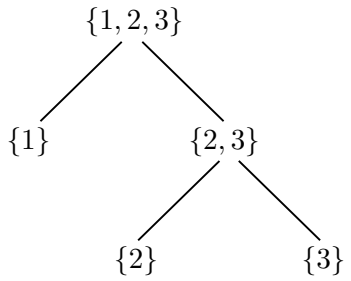

FIGURE 1 – Arbre correspondant à $T_3 = \{\{1\}, \{2\}, \{3\}, \{2, 3\}, \{1, 2, 3\}\}$

On considèrera donc naturellement T_n comme un arbre discret à n feuilles étiquetées par $[n]$, et on appellera la suite $(T_n)_{n\geq 1}$ la suite des **arbres de fragmentation** associée à Π.

Il est intéressant de constater que ces arbres sont fortement compatibles par échantillonnage, c'est-à-dire que T_n est l'arbre obtenu si l'on enlève à T_{n+1} la feuille étiquetée par l'entier $n + 1$.

3 Arbres branchants Markoviens

Dans cette section, on présente le modèle des arbres branchants Markoviens, introduits par Aldous [2] sous une forme particulière dans une tentative d'axiomatisation de modèle d'arbres phylogénétiques. Pour les arbres binaires à n feuilles, Aldous développe les modèles dit de *β-splitting*, famille de modèles à un paramètre continu qui permet de retrouver la loi d'un arbre de Yule ou encore d'un arbre uniforme. Les axiomes retenus permettent en fait de construire des familles d'arbres plus généraux, dont les lois sont liés aux lois des processus de fragmentation homogène.

3.1 Définitions

Pour $n \in \mathbb{N}$ fixé, on considère l'ensemble $\mathcal{T}_n$ des arbres binaires à n feuilles, numérotées par l'ensemble $[n] = \{1, \dots, n\}$.

Définition 4. Un modèle d'**arbre branchant Markovien** (MBT pour *Markov branching tree*) est une famille $(P_n)_{n\geq 1}$, où P_n est une mesure de probabilité sur $\mathcal{T}_n$, qui peut être décrite de la manière suivante.

Il existe une famille $(q_n)_{n\geq 2}$ de mesures de probabilités sur $\mathbb{N}$ telle que :
— $q_n([n-1]) = 1$,
— Pour tout $i \in [n-1]$, on a $q_n(i) = q_n(n-i)$.

Un arbre aléatoire T qui suit la loi P_n peut être construit ainsi :

1. On se donne une variable aléatoire K selon la loi q_n,
2. Deux arbres aléatoires T' et T'', de lois respectives P_K et P_{n-K} sont tirés,
3. L'arbre tel que T' et T'' soient les deux sous-arbres issus de la racine, dont les feuilles sont ré-étiquetées uniformément au hasard par $[n]$, a la loi de T.

Un MBT est donc déterminé par la famille $(q_n)_{n\geq 1}$ qui donne la loi du nombre de feuilles dans le sous-arbre à gauche de la racine (si l'on se donne une représentation planaire de l'arbre choisie uniformément). Il découle de la définition que les MBT sont nécessairement échangeables, c'est-à-dire invariant par les permutations des étiquettes des arbres de $\mathcal{T}_n$. Cependant, pour pouvoir former des modèles cohérents d'arbres phylogénétiques, on demande une propriété de plus : la compatibilité par échantillonnage (*sampling consistency*). On définit pour $n \geq 1$ l'application $r_n : \mathcal{T}_{n+1} \to \mathcal{T}_n$, tel que $r_n(T)$ soit l'arbre T auquel on enlève la feuille numérotée par $n+1$.

Définition 5. Un modèle d'arbre branchant Markovien $(P_n)_{n\geq 1}$ est dit **compatible par échantillonnage** si la mesure image de P_{n+1} par r_n est P_n, pour tout n :

$$P_{n+1} \circ r_n^{-1} = P_n,$$

c'est-à-dire que si T_{n+1} est tiré selon la loi P_{n+1}, alors l'arbre $r_n(T_{n+1})$ a la loi P_n.

3.2 Exemple : le modèle de beta-splitting

Aldous [2] introduit une famille de MBT compatibles par échantillonnage, indexée par un paramètre $\beta > -2$. Pour $\beta > -1$, on considère la fonction

$$f :]0,1[\to \mathbb{R}_+,\ x \mapsto x^\beta (1-x)^\beta.$$

Comme $C = \int_0^1 f(x)\mathrm{d}x = \frac{\Gamma(\beta+1)^2}{\Gamma(2\beta+2)}$ est finie, on peut tirer un nombre X aléatoire selon la loi de densité f renormalisée :

$$X \sim \frac{f(x)}{C}\mathrm{d}x.$$

Pour $n \geq 1$ et $(U_1, \ldots, U_n)$ des variables uniformes sur $[0,1]$ indépendantes, on définit q_n comme la loi du nombre de ces variables à gauche de X, conditionné à être compris entre 1 et $n-1$:

$$q_n(k) = \frac{1}{\alpha_n(\beta)}\binom{n}{k}\int_0^1 x^\beta(1-x)^\beta\, x^k(1-x)^{n-k}\,\mathrm{d}x,$$

avec $\alpha_n(\beta) = \int_0^1 x^\beta(1-x)^\beta\,(1-x^n-(1-x)^n)\mathrm{d}x$. On peut remarquer que ces formules restent bien définies pour $\beta > -2$, ce qui permet de définir un modèle de β-splitting pour toutes ces valeurs. De plus, on peut voir, par le calcul ou par la construction avec des variables aléatoires uniformes sur $[0,1]$, que ces modèles sont compatibles par échantillonnage.

Une interpolation entre différents modèles La particularité de cette famille de modèle de β-splitting est qu'elle permet de retrouver des lois classique pour différentes valeurs de β. On peut montrer en particulier que :

- pour $\beta = -3/2$, la loi P_n obtenue est la loi uniforme sur $\mathcal{T}_n$;
- pour $\beta = 0$, P_n est la loi de la structure d'un arbre de Kingman ("n coalescent" dans [9]). C'est aussi la loi de la généalogie des individus vivants dans n'importe quel processus de naissance et de mort stoppé à un temps T fixe, conditionné à avoir n individus vivants.

3.3 MBT compatibles par échantillonnage

Il est possible de caractériser tous les arbres branchants Markoviens qui sont compatibles par échantillonnage. Dans un article de Haas et al. [8], les auteurs caractérisent ainsi toutes les mécaniques d'arbres branchants Markoviens (c'est-à-dire avec des branchements généraux, pas seulement binaires). Le résultat est que l'on peut comprendre ces modèles d'arbres aléatoires comme la structure discrète associée à un processus de fragmentation homogène. On présente ici la version "branchements binaires" du théorème.

Théorème 3.1. *Soient $(q_n)_{n\geq 2}$ les lois de branchement d'un arbre branchant Markovien compatible par échantillonnage. Alors il existe une mesure μ sur $]0,1[$, symétrique (c'est-à-dire invariante par l'application $x \mapsto (1-x)$), qui vérifie*

$$\int_{[0,1]} x(1-x)\mu(\mathrm{d}x) < \infty,$$

et telle que l'on ait, pour $n \geq 2$ et $k \leq n$:

$$q_n(k) = \frac{1}{\alpha_n}\binom{n}{k}\left(\int_{]0,1[} x^k(1-x)^{n-k}\mu(\mathrm{d}x) + n\mu(\{0\})\mathbb{1}_{k=1} + n\mu(\{1\})\mathbb{1}_{k=n-1}\right), \quad (3)$$

avec

$$\alpha_n = \int_{]0,1[} 1 - x^n - (1-x)^n\mu(\mathrm{d}x) + n\mu(\{0,1\}).$$

Pour rendre le lien avec la théorie des fragmentations plus clair, étant donnée une mesure μ vérifiant les conditions du théorème, on peut définir

$$c := \mu(\{0,1\}) \quad \text{et} \quad \nu := \mu_{|]0,1[} \circ \phi^{-1},$$

avec $\phi :]0,1[\to \mathcal{P}_{\mathtt{m}}$ l'application définie par

$$\phi(x) = \begin{cases} (x, 1-x, 0, \ldots) & \text{si } x \geq 1/2, \\ (1-x, x, 0, \ldots) & \text{si } x < 1/2. \end{cases}$$

Alors si Π est une fragmentation homogène de taux d'érosion c et de mesure de dislocation ν, la suite $(T_n)_{n\geq 1}$ des arbres de fragmentation associée à Π est un couplage fortement compatible par échantillonnage des lois des MBT définis par (3). C'est-à-dire que pour tout n, T_n a loi P_n, et l'on a aussi $T_n = r_n(T_{n+1})$.

4 Perspective : processus de fragmentation imbriqué

Les processus de fragmentation, à valeurs dans les partitions des entiers, permettent donc de modéliser des phylogénies. On peut toujours se demander si l'on peut étendre ces processus de fragmentations à des espaces plus généraux. Alors il faudrait peut-être redéfinir le concept d'échangeabilité, et imposer des hypothèses de branchement particulières. Pour citer un exemple de généralisation, Crane [6] s'intéresse au problème arbre des espèces – arbre des gènes, et crée un modèle d'arbre branchants généralisé, où le branchement de l'arbre des gènes est défini conditionnellement à une statistique de l'arbre des espèces.

4.1 Problème arbres des espèces – arbre des gènes

Le problème évoqué ci-dessus désigne la difficulté à reconstruire le véritable arbre de spéciation à partir des différents arbres phylogénétiques obtenus en observant les données génétiques de plusieurs espèces. Plusieurs facteurs biologiques empêchent une reconstruction "facile" de la vraie phylogénie, du fait que les arbres de gènes construits par vraisemblance ne sont pas forcément compatibles entre eux. Pour modéliser ce phénomène, on imagine une évolution à deux niveaux : les espèces se distinguent au fil du temps, et au sein des espèces, l'information génétique se différencie également. Par exemple, sur la figure 2 dans l'arbre de gauche, chaque nœud noir représente une information génétique distincte, et chaque enveloppe (en pointillés) représente une espèce distincte. Chaque branchement représente un point du temps où s'effectue une différenciation à un niveau et/ou à l'autre. Alors on peut reconstruire l'arbre des espèces en identifiant les points noirs à l'intérieur d'une même enveloppe, et au contraire reconstruire l'arbre des gènes en ignorant les enveloppes. Sur cet exemple simple, on voit que les deux arbres reconstruits sont différents, ce qui laisse penser que considérer une telle évolution à deux niveaux permet de modéliser le problème.

4.2 Représentation des arbres à deux niveaux

Pour pouvoir représenter de tels arbres, on peut considérer une suite de couples de partitions imbriquées. C'est-à-dire que si $\pi \prec \pi'$ signifie que chaque bloc de π' est

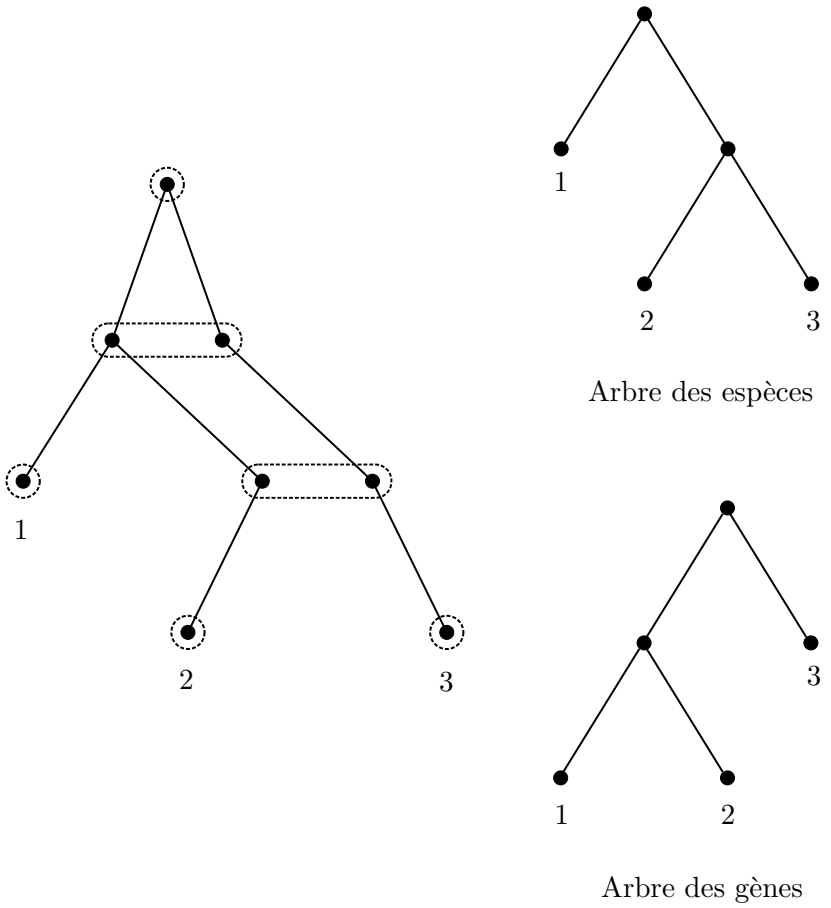

FIGURE 2 – Arbre double, induisant deux arbres des espèces et des gènes distincts

une union de blocs de π, on peut représenter un arbre double à n feuilles par une suite $(\pi_i, \pi'_i)_{1\leq i\leq N}$ de partitions de $[n]$ telle que :

$$\forall i \geq 1,\ \pi_{i+1} \prec \pi_i \prec \pi'_i \quad \text{et} \quad \pi'_{i+1} \prec \pi'_i,$$

$$\pi_1 = \pi'_1 = \mathbf{1}_{[n]} \quad \text{et} \quad \pi_N = \pi'_N = \{\{1\}, \{2\}, \ldots, \{n\}\}.$$

Alors la structure de l'arbre double serait donnée par l'arbre discret des gènes construit par la suite $(\pi_i)_{1\leq i\leq N}$, dont les sommets sont regroupés à chaque étape i selon la partition plus grossière π'_i. Pour donner un exemple simple, l'arbre de la figure 2 est codé par la suite de partitions :

$$\begin{aligned}
\pi_1 &= \{\{1,2,3\}\}, & \pi'_1 &= \{\{1,2,3\}\},\\
\pi_2 &= \{\{1,2\},\{3\}\}, & \pi'_2 &= \{\{1,2,3\}\},\\
\pi_3 &= \{\{1\},\{2\},\{3\}\}, & \pi'_3 &= \{\{1\},\{2,3\}\},\\
\pi_4 &= \{\{1\},\{2\},\{3\}\}, & \pi'_4 &= \{\{1\},\{2\},\{3\}\}.
\end{aligned}$$

De même que pour les MBT, on aimerait définir des axiomes de branchement et de compatibilité pour les arbres à deux niveaux qui permettrait de faire le lien avec une fragmentation (à définir) à valeurs dans l'espace des partitions imbriqués $\{(\pi, \pi') \in \mathcal{P}^2_\infty,\ \pi \prec \pi'\}$.

Références

[1] R. Abraham and L. Serlet. Poisson snake and fragmentation. *Electron. J. Probab.*, 7 :1–15, 2002.

[2] D. Aldous. Probability distributions on cladograms. In *Random discrete structures*, pages 1–18. Springer, 1996.

[3] J. Bertoin. A fragmentation process connected to Brownian motion. *Probability Theory and Related Fields*, 117(2) :289–301, 2000.

[4] J. Bertoin. *Random fragmentation and coagulation processes*, volume 102. Cambridge University Press, 2006.

[5] B. Chen, D. J. Ford, and M. Winkel. A new family of Markov branching trees : the alpha-gamma model. *Electron. J. Probab*, 14(15) :400–430, 2009.

[6] H. Crane. Generalized Markov branching trees. *Advances in Applied Probability*, to be published.

[7] D. J. Ford. Probabilities on cladograms : introduction to the alpha model. *arXiv preprint*, 2005.

[8] B. Haas, G. Miermont, J. Pitman, and M. Winkel. Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models. *The Annals of Probability*, 36(5) :1790–1837, 2008.

[9] J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13 :235–248, 1982.

[10] P. Marchal. Nested regenerative sets and their associated fragmentation process. In *Mathematics and Computer Science III : Algorithms, Trees, Combinatorics and Probabilities*, pages 461–470. Birkhäuser Basel, 2004.

[11] P. Marchal. A note on the fragmentation of a stable tree. In *Fifth Colloquium on Mathematics and Computer Science*, pages 489–500. Discrete Mathematics and Theoretical Computer Science, 2008.