# Arbres : combinatoire et modèles

Gilles Didier, Stéphane Guindon

# Arbres : Combinatoire et Modèles

Gilles Didier[1] et Stéphane Guindon[2]

[1] IMAG, Univ Montpellier, CNRS, Montpellier, France

[2] LIRMM, Univ Montpellier, CNRS, Montpellier, France

## 1 Introduction

Les arbres décrivent les processus évolutifs à différentes échelles temporelles. Ainsi, pour des temps longs, chaque branche dans l'arbre, que l'on désigne ici par le terme «phylogénie», représente généralement l'évolution d'un ensemble d'organismes échangeant du matériel génétique, c'est à dire une *espèce*. Les nœuds dits «internes» (l'ensemble des nœuds distincts des feuilles) correspondent alors à la formation de nouveaux clades, ou «*cladogénèse*». La formation d'un nouveau clade est provoquée par (ou entraîne) une diminution voire une interruption dans l'échange de matériel génétique et définit une partition de l'ensemble des organismes constituant l'espèce en question. Dans certaines conditions, l'apparition d'une telle partition est synonyme de spéciation : l'espèce ancestrale donne naissance à une ou plusieurs nouvelle(s) espèce(s), générant ainsi une bi- ou multi-furcation dans l'arbre phylogénétique. Enfin, une feuille d'un arbre phylogénétique apparaît lorsqu'une espèce disparait. Une feuille peut également traduire le fait que l'espèce correspondante a été échantillonnée afin d'être incorporée au sein d'un corpus de données à analyser.

À une échelle de temps plus courte, un arbre peut décrire les relations de parentés évolutives entre organismes de la même espèce ou de la même population. Une branche d'un tel arbre — arbre que l'on désigne généralement par le terme «*généalogie*» — décrit alors une succession d'individus liés par des relations de parentés directes. Un nœud interne de la généalogie correspond à un évènement de reproduction où un parent donne naissance à une ou plusieurs nouvelles lignées, une lignée étant définie comme l'ascendance d'un individu. Enfin, une feuille est associée à la disparition ou l'échantilonnage d'une lignée.

Quelle que soit l'échelle temporelle considérée, les phénomènes de branchement évoqués ci-dessous se prêtent à la modélisation probabiliste. Ces processus ont d'ailleurs fait l'objet d'un nombre considérable de travaux mathématiques. Ainsi, dès 1845, le probabiliste et statisticien français Irénée-Jules Bienaymé propose un modèle stochastique de branchements pour étudier la disparition des patronymes. Le modèle proposé deviendra celui de Galton-Watson et sera à l'origine d'une nouvelle branche en mathématiques, combinant aspects combinatoires et probabilistes. Les processus de coalescence, qui, en simplifiant, équivalent à des processus de branchement lorsque la flèche du temps est inversée, sont également à l'origine d'un foisonnement d'études mathématiques (voir Aldous *et al.*, 1999; Möhle, 2000;

Berestycki *et al.*, 2009, pour des revues sur ce thème). Ces derniers ont permis d'éclairer sous un jour nouveau des questions importantes en génétique des populations, avec, en particulier, l'établissement de relations d'équivalence entre modèles de biologie des populations reposant sur des hypothèses très distinctes.

Ce chapitre se focalise sur les deux principaux modèles probabilistes générateurs d'arbres, à savoir le processus de naissance et mort et le coalescent de Kingman. Nous donnerons dans un premier temps les éléments de combinatoire nécéssaires au comptage de ces arbres. Les probabilités des arbres correspondants, en omettant les âges des nœuds, seront données ensuite. Les densités de probabilité des arbres munis d'âges aux nœuds seront enfin détaillées. Ces dérivations nous permettront notamment d'exposer les hypothèses sous-jacentes aux deux modèles d'arbres et de discuter des liens avec d'autres modèles populaires en biologie des populations.

# 2 Définitions préliminaires

Si les arbres sont utilisés pour modéliser le support de l'évolution biologique, ils sont également des objets formels étudiés en tant que tels en informatique et en mathématiques. Dans ces deux disciplines, un arbre se définit comme un graphe (non-orienté) connexe et sans cycle. Autrement dit, un arbre est formé d'un ensemble d'éléments appelés *nœuds* ou *sommets* dont certains sont reliés entre eux par des *arêtes* non-orientées de telle sorte qu'il existe un et un seul chemin reliant deux sommets quelconques. On dit que deux sommets directement reliés par une arête dans un arbre ou plus généralement dans un graphe sont *voisins* et on appelle *chemin* de l'arbre une succession de sommets $(a_1, a_2, \ldots, a_n)$ telle que $a_i$ et $a_{i+1}$ sont reliés entre eux pour tout $1 \leq i < n$. On distingue les sommets d'un arbre qui ne sont reliés qu'à un seul autre sommet et qu'on appelle *feuilles* de l'arbre. Les autres sommets de l'arbre sont appelés *nœuds internes*.

On rencontrera plus loin deux types d'arbres : les arbres non-enracinés dont la définition est exactement celle ci-dessus et les arbres enracinés qui sont des arbres dans lesquels l'un des sommets est distingué et appelé *racine* de l'arbre. Nous nous intéresserons ici plus particulièrement aux arbres enracinés car ceux-ci ont une interprétation plus directe en terme d'évolution sur laquelle nous reviendrons plus loin. Dans un cadre phylogénétique, les arbres non-enracinés apparaissent essentiellement pour représenter de manière visuelle des relations de proximités entre espèces. Dans un contexte évolutif, les jeux de données sont d'ailleurs souvent choisis de telle sorte qu'il soit ensuite plus ou moins facile de déterminer la racine de l'arbre.

Revenons aux arbres enracinés. De manière quelque peu contre-intuitive, ils sont en général représentés avec leur racine en haut. Un point important est que la racine d'un arbre permet de l'orienter car emprunter une arête (ou plus généralement un chemin) de l'arbre soit éloigne, soit rapproche de la racine. En effet, comme pour tout sommet $a$ de l'arbre différent de sa racine, il existe toujours un unique chemin de la racine vers $a$ (car l'arbre est connexe) et que celui-ci passe par un et un seul voisin de $a$ (sinon il y aurait un cycle), qui est nécessairement le seul voisin de $a$ plus proche de la racine et qu'on appelle

l'*ancêtre direct* ou le *père* de $a$. Ceci permet de définir une relation d'ordre partiel de type «est ancêtre de» (dans un sens plus large que «est ancêtre direct de») ou symétriquement «est descendant de» sur les sommets de l'arbre, un sommet $b$ étant ancêtre du sommet $a$ si l'unique chemin menant de la racine à $a$ passe par $b$. Parmi les arbres enracinés, on distinguera les *arbres binaires* dans lesquels tout nœud qui n'est pas une feuille possède exactement deux descendants directs. Une propriété importante des arbres enracinés est que si l'on supprime le nœud racine et toutes les arêtes qui en partent, on obtient autant d'arbres enracinés que la racine avait de descendants directs, ceux-ci ayant pour racine ces descendants directs. Cette propriété est très largement utilisée à la fois pour effectuer des calculs de quantités caractéristiques sur les arbres et pour en démontrer des propriétés par induction. Le principe général est le suivant : dans la situation (assez fréquente) où la propriété à démontrer est trivialement vraie pour l'arbre constitué d'un seul sommet et où l'on peut déduire que la propriété est vérifiée pour un arbre donné si celle-ci l'est pour tous les sous-arbres partant de sa racine, le principe d'induction permet de vérifier que la propriété est ainsi valide pour tout arbre. De même, si l'on sait calculer une certaine quantité caractéristique pour l'arbre constitué d'un seul sommet et combiner les quantités des sous-arbres de la racine d'un arbre donné pour déterminer sa propre quantité caractéristique, il est possible d'en déduire un algorithme de calcul récursif applicable à tout arbre et dont la complexité est linéaire avec la taille de l'arbre considéré. Par exemple, pour calculer le nombre de feuilles d'un arbre, il suffit de remarquer que celui-ci vaut un dans le cas de l'arbre formé d'un seul sommet puis de noter que le nombre de feuilles d'un arbre quelconque est égal à la somme des nombres de feuilles de tous les sous-arbres partant de sa racine. Pour montrer que dans un arbre binaire, le nombre de feuilles est égal au nombre de nœuds internes plus un, on remarque que la propriété est vraie dans le cas de l'arbre à un seul sommet, puis de voir que la propriété est vérifiée pour tout arbre avec $n+1$ sommets si celle-ci est vraie pour tout arbre avec moins de $n$ sommets. En effet, le nombre total de nœuds internes de l'arbre est égal à la somme de ceux des deux sous-arbres partant de la racine plus un (la racine) et le nombre total de feuilles est égal à la somme de celles de ces mêmes sous-arbres, qui, ayant nécessairement moins de $n$ sommets, vérifient tous deux la propriété.

# 3 Compter les arbres

Un problème mathématique naturel dans le contexte qui nous intéresse ici est de compter le nombre d'arbres différents. C'est une question assez ancienne puis qu'elle remonte au moins au XIX$^{\text{ème}}$ siècle et certains des travaux d'Arthur Cayley. Notons que le nombre d'arbres différents dépend à la fois du type d'arbres que l'on considère (e.g. enraciné ou non) mais aussi de ce qui les distingue entre eux, en particulier selon comment on étiquette ou pas leur nœuds. Deux arbres où aucun nœud n'est étiqueté sont identiques s'il existe une bijection entre les sommets de l'un et de l'autre qui conserve les relations de voisinage (c'est à dire que tout voisin de l'image d'un nœud est l'image d'un voisin de ce dernier et réciproquement). Intuitivement, cela signifie que les graphes ont alors la même forme. Si l'on étiquette, même partiellement, les nœuds des arbres, deux arbres seront considérés identiques seulement s'il

existe une bijection entre les sommets de l'un et de l'autre qui conserve à la fois les relations de voisinage et les étiquettes. Naturellement, deux arbres étiquetés identiques resteront identiques si l'on supprime leurs étiquettes. Dans ce qui suit, on suppose implicitement que deux nœuds étiquetés d'un même arbre ont des étiquettes différentes.

## 3.1 Arbres non enracinés entièrement étiquetés

Le nombre d'arbres non enracinés où tous les nœuds sont étiquetés (chacun avec une étiquette unique) avec $n$ nœuds est $n^{n-2}$. Ce résultat est du à Cayley. Prüfer (1918) en a proposé une preuve en établissant une bijection entre les arbres étiquetés de taille $n$ et les séquences formées de $n-2$ nombres pris de 1 à $n$.

Cette bijection associe à tout arbre de taille $n$ où les nœuds sont étiquetés (ici plutôt numérotés) de 1 à $n$ la séquence $s$ de longueur $n-2$ dont les termes sont pris entre 1 et $n$ construite de la façon suivante. On initialise $s$ à la séquence vide. Tant que l'arbre contient plus de deux nœuds, on concatène à la fin de $s$ le nombre étiquetant le nœud voisin de la feuille étiquetée avec le plus petit nombre. On met ensuite à jour l'arbre en supprimant la feuille étiquetée avec le plus petit nombre et on itère le même processus.

Réciproquement, on associe à toute séquence $s = s_1 s_2 \dots s_{n-2}$ de nombres pris de 1 à $n$ l'arbre construit de la façon suivante. On considère l'ensemble $I$ qu'on initialise comme l'ensemble des nombres entre 1 et $n$ n'apparaissant pas dans $s$. On construit ensuite le graphe $\mathcal{G}$, initialement vide, en itérant le processus suivant. Soit $m$ le plus petit nombre de $I$. On ajoute dans le graphe $\mathcal{G}$ un nœud étiqueté par $m$ et un nœud étiqueté par le premier terme de $s$ s'ils n'y étaient pas et on ajoute une arête entre ces deux nœuds. On met ensuite à jour l'ensemble $I$ en y supprimant $m$ et en y ajoutant le premier terme de $s$ si et seulement si c'était la dernière occurrence de ce nombre dans $s$. On met également à jour la séquence $s$ en supprimant son premier terme. On répète ces opérations jusqu'à arriver à une séquence vide. Il y a alors exactement deux nombres entre 1 et $n$ dans $I$ (à ce stade, on a retiré $n-2$ nombres entre 1 et $n$ de $I$ et tout nombre entre 1 et $n$ est à un moment ou un autre élément de $I$). On ajoute deux nœuds étiquetés par ces nombres dans $\mathcal{G}$ s'ils ne s'y trouvaient pas et on les connecte par une arête (Fig. 1). Par construction, on en déduit que

- le graphe $\mathcal{G}$ contient bien exactement $n$ nœuds étiquetés de 1 à $n$ après la dernière itération,
- le graphe $\mathcal{G}$ ne contient aucun cycle puisque à chaque fois que l'on ajoute une arête, l'un des deux nœuds qui la constitue ne sera plus connecté à aucun autre nœud par la suite,
- après chaque itération, tout nœud du graphe $\mathcal{G}$ est connecté soit à un nœud étiqueté par un élément de $I$, soit à un nœud étiqueté par un terme de la séquence $s$ (mise à jour). À la dernière itération, tout nœud de $\mathcal{G}$ est donc connecté à l'un ou l'autre des nœuds étiquetés par les deux nombres restant dans $I$ et cette dernière itération connecte l'ensemble du graphe.

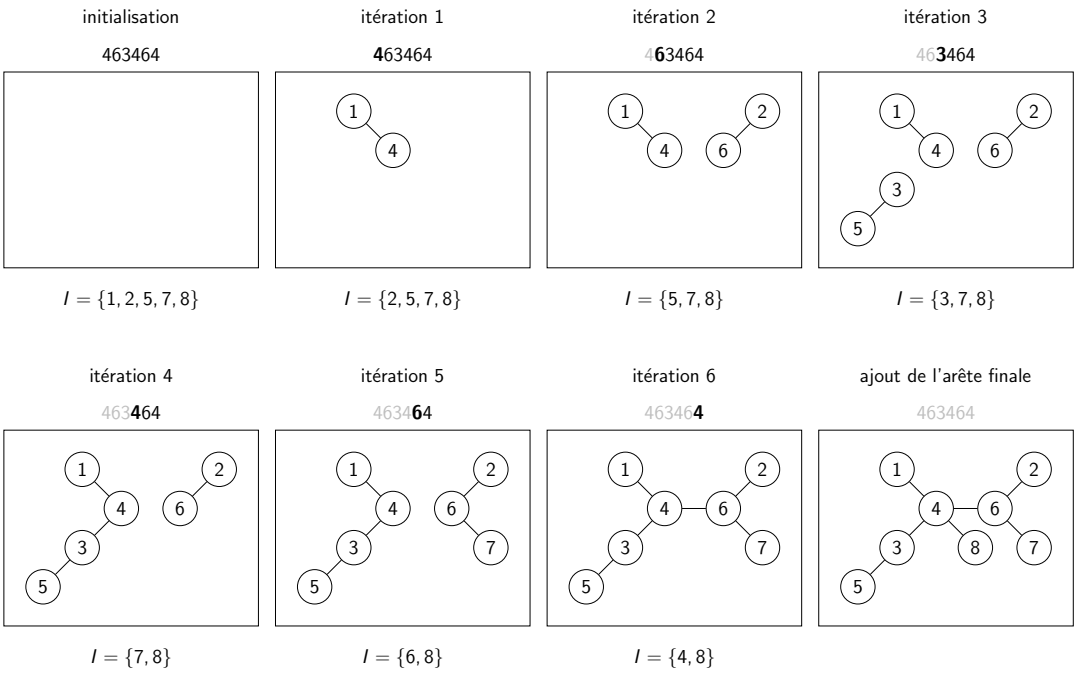


Figure 1: Calcul de l'arbre associé à la séquence 463464. Sous chaque itération, le premier terme de la séquence courante est en gras et on présente les états du graphe et de l'ensemble $I$ juste après cette itération.

En résumé, le graphe $\mathcal{G}$ obtenu à partir de n'importe quelle séquence de $n-2$ nombres pris entre 1 et $n$ est bien un graphe connexe et sans cycle, autrement dit un arbre, de $n$ sommets étiquetés de 1 à $n$.

On vérifie que l'opération transformant un arbre en séquence, qu'on appellera $f$, est bien réciproque de celle transformant une séquence en arbre, qu'on appellera $g$. En effet, soit $s$ une séquence quelconque de $n-2$ nombres pris entre 1 et $n$. On note $\mathcal{T}$ l'arbre obtenu en appliquant la transformation $g$ à $s$ et $s'$ la séquence obtenue en appliquant la transformation $f$ à $\mathcal{T}$. Au début de l'itération 1, de la transformation $g$ appliquée à $s$, l'ensemble $I$ contient les nombres étiquetant les feuilles de $\mathcal{T}$ qui est l'arbre courant au début de l'itération 1 de la transformation $f$ appliquée à $\mathcal{T}$. Par construction, l'étiquette du voisin de la feuille avec la plus petite étiquette de $\mathcal{T}$ est $s_1$ et l'on a donc bien $s'_1 = s_1$. Comme l'arbre mis à jour à la fin de l'itération 1 de $f$ est l'arbre $\mathcal{T}$ privé de sa feuille avec la plus petite étiquette qui est également l'image par $g$ de la séquence $s_2 \dots s_{n-2}$ (en renumérotant les nœuds et la séquence), on a de la même façon que $s'_2 = s_2$ etc. On en conclut que $f(g(s)) = s$ pour toute séquence $s$ et l'on a bien établi une bijection entre les arbres étiquetés non enracinés de taille $n$ et les séquences de longueur $n-2$ de nombres pris entre 1 et $n$.

## 3.2 Arbres binaires où seules les feuilles sont étiquetées

Intéressons nous maintenant au nombre d'arbres binaires (donc enracinés) dont (toutes et seulement) les feuilles sont étiquetées, disons de 1 à $n$, qui est la situation la plus couramment rencontrée en phylogénie. On note $\mathcal{B}(n)$ le nombre d'arbres binaires de $n$ feuilles étiquetées. Attention, $n$ désigne ici le nombre de feuilles et pas le nombre total de nœuds de l'arbre comme précédemment. On a vu ci-dessus que le nombre de nœuds internes est alors $n-1$ (le nombre total de nœuds de l'arbre est donc $2n-1$). Considérons un arbre binaire avec $n$ feuilles étiquetées de 1 à $n$. L'ancêtre direct de la feuille étiquetée par $n$ est soit la racine de l'arbre, soit un autre nœud interne. Dans le premier cas, si l'on supprime la feuille $n$ et la racine (son ancêtre direct) ainsi que les deux arêtes où ils interviennent, on obtient un arbre binaire de $n-1$ feuilles étiquetées de 1 à $n-1$. Dans le second cas, l'ancêtre direct de la feuille $n$ est un nœud interne qui a lui-même un ancêtre direct. Comme l'arbre est binaire, l'ancêtre direct de $n$ a également un descendant direct autre que $n$. Si l'on supprime la feuille $n$ et son ancêtre direct ainsi que les trois arêtes où l'un ou l'autre intervient et que l'on ajoute une arête entre l'ancêtre direct de l'ancêtre direct de $n$ et le nœud frère de $n$, on obtient à nouveau un arbre binaire de $n-1$ feuilles étiquetées de 1 à $n-1$. On en déduit que tout arbre binaire de $n-1$ feuilles étiquetées de 1 à $n-1$ peut être obtenu en supprimant la feuille $n$ de $2n-3$ arbres binaires de $n$ feuilles étiquetées de 1 à $n$. En effet, un arbre binaire de $n-1$ feuilles a $2n-3$ nœuds (feuilles ou internes, racine inclue). L'ancêtre direct de la feuille $n$ peut être ajouté comme ancêtre direct de n'importe lequel d'entre eux de telle sorte qu'on obtienne le même arbre binaire en le supprimant comme ci-dessus. De plus, si les arbres binaires obtenus en supprimant la feuille $n$ de deux arbres binaires avec $n$ feuilles étiquetées sont différents, les deux arbres binaires avec $n$ feuilles étiquetées le sont également. On en conclut que le nombre d'arbres avec $n$ feuilles étiquetées est égal au nombre d'arbres binaires de $n-1$ feuilles étiquetées multiplié par $2n-3$, i.e.,

$\mathcal{B}(n) = (2n-3)\mathcal{B}(n-1)$. Comme il n'y a qu'un seul arbre binaire avec deux feuilles étiquetées, on en déduit $\mathcal{B}(n) = (2n-3)(2n-5)\dots 1 = \prod_{i=2}^{n}(2i-3) = \frac{(2n-3)!}{2^{n-2}(n-2)!}$, une quantité que l'on note $(2n-3)!!$.

## 3.3 Arbres binaires à feuilles étiquetées et à nœuds internes ordonnés

Dans un contexte évolutif, on considère parfois des arbres binaires dont les feuilles sont étiquetées et dont les nœuds internes sont ordonnés les uns par rapport aux autres selon un ordre compatible avec l'arbre. Un ordre sur les nœuds est *compatible* avec leur arbre s'il est tel que le rang de tout nœud est supérieur à celui de n'importe lequel de ses ancêtres (et inférieur à celui de n'importe lequel de ses descendants). En particulier, la racine a toujours le rang 1 dans un ordre compatible. Dans un cadre phylogénétique, un ordre compatible s'interprète comme représentant les relations temporelles des évènements de divergences évolutives ayant mené à cet arbre (cf la section suivante).

Compter le nombre d'arbres binaires ordonnés et étiquetés peut être réalisé de la même façon que dans le cas non ordonné. Soit un arbre binaire ordonné avec $n$ feuilles étiquetées de 1 à $n$. Si l'on enlève la feuille $n$ et son ancêtre direct de la même façon que précédemment, on obtient un arbre binaire ordonné de $n-1$ feuilles étiquetées de 1 à $n-1$ dans lequel les rangs des nœuds d'ordre supérieur à celui de l'ancêtre direct de $n$ ont été diminués de 1.

Le nombre de positions possibles de l'ancêtre direct de la feuille $n$ dépend de son rang parmi les nœuds internes. On montre par récurrence que si son rang est $r$ alors l'ancêtre direct de la feuille $n$ a $r$ positions possibles. C'est trivialement vrai pour $r = 1$, car l'ancêtre direct est alors nécessairement la racine. Supposons que la propriété est vraie pour un certain $1 < r < n-1$, autrement dit, qu'il y a exactement $r$ branches (de l'arbre à $n-1$ feuilles) qui partent d'un nœud de rang strictement inférieur à $r$ et arrivent à un nœud de rang supérieur ou égal à $r$ ou une feuille. Parmi celles-ci, une branche arrive au nœud de rang $r+1$ et les $r-1$ autres arrivent à des nœuds internes de rangs strictement supérieurs à $r+1$ ou des feuilles. L'arbre étant binaire, il part deux branches du nœud de rang $r+1$. On en déduit qu'il y a $r+1$ branches dont l'ancêtre a un rang strictement inférieur à $r+1$ et le descendant a un rang supérieur ou égal à $r+1$ ou une feuille et la propriété est donc vraie pour tout rang $r$. Enfin, si son rang est $r$, il y autant de positions possibles de l'ancêtre direct de la feuille $n$ que de branches partant d'un d'un nœud de rang strictement inférieur à $r$ et arrivant à un nœud de rang supérieur ou égal à $r$ ou une feuille, c'est à dire $r$. Comme chaque arbre binaire ordonné avec $n-1$ feuilles étiquetées, chaque rang et position possible pour l'ancêtre direct de la feuille donnant un arbre ordonné différent et en notant $\mathcal{O}(n)$ le nombre d'arbres binaires ordonnés de taille $n$, on a $\mathcal{O}(n) = \mathcal{O}(n-1)\sum_{r=1}^{n-1} r = \mathcal{O}(n-1)\frac{n(n-1)}{2}$. Comme il n'y a qu'un seul arbre ordonné avec deux feuilles étiquetées, on en conclut que $\mathcal{O}(n) = \frac{n!(n-1)!}{2^{n-1}}$.

## 3.4 Nombre d'ordres de nœuds internes d'un arbre donné

L'ordre des nœuds internes constituant un information supplémentaire, on compte plus d'arbres ordonnés et étiquetés que d'arbres seulement étiquetés. Le résultat suivant montre comment calculer le nombre d'arbres ordonnés correspondants à un arbre (non-ordonné) donné, autrement dit le nombre d'ordres compatibles avec un arbre donné. Le nombre d'arbres ordonnés déterminé dans la section précédente peut être obtenu en sommant ce nombre d'ordres compatibles sur l'ensemble des arbres binaires avec feuilles étiquetées.

**Théorème 1.** *Le nombre $\mathcal{R}(\mathcal{T})$ d'ordres de nœuds internes compatibles avec un arbre binaire, enraciné $\mathcal{T}$ dont les feuilles sont étiquetées est égal à*

$$\mathcal{R}(\mathcal{T}) = \mathcal{R}(\mathcal{T}_a)\mathcal{R}(\mathcal{T}_b)\binom{\mathbf{L}_{\mathcal{T}_a} + \mathbf{L}_{\mathcal{T}_b} - 2}{\mathbf{L}_{\mathcal{T}_a} - 1}$$

*où $\mathcal{T}_a$, $\mathcal{T}_b$ , $\mathbf{L}_{\mathcal{T}_a}$ et $\mathbf{L}_{\mathcal{T}_b}$ sont respectivement les deux sous-arbres partant de la racine de $\mathcal{T}$ et leurs nombres de feuilles.*

*Preuve.* On remarque que la restriction de tout ordre de nœuds internes compatible avec $\mathcal{T}$ aux nœuds internes de $\mathcal{T}_a$ (resp. de $\mathcal{T}_b$) est un ordre compatible avec $\mathcal{T}_a$ (resp. avec $\mathcal{T}_b$). Réciproquement, tout ordre obtenu en fusionnant un ordre compatible avec $\mathcal{T}_a$ avec un ordre compatible avec $\mathcal{T}_b$ est un ordre compatible avec $\mathcal{T}$. On en déduit qu'il y a autant d'ordres compatibles avec $\mathcal{S}$ que de façons de fusionner un ordre de $\mathcal{T}_a$ avec un ordre de $\mathcal{T}_b$. Une fusion étant parfaitement déterminée par l'ensemble des positions des nœuds internes de $\mathcal{T}_a$ dans l'ordre sur $\mathcal{T}$. En effet, les positions de ceux de $\mathcal{T}_b$ s'en déduisent et l'ordre relatif des nœuds internes de $\mathcal{T}_a$ (resp. de $\mathcal{T}_b$) à l'intérieur de ses positions est donné par l'ordre sur ce sous-arbre. En résumé, on a

- $\mathcal{O}(\mathcal{T}_a)$ ordres compatibles avec $\mathcal{T}_a$,
- $\mathcal{O}(\mathcal{T}_b)$ ordres compatibles avec $\mathcal{T}_b$,
- $\binom{\mathbf{L}_{\mathcal{T}_a}-1}{\mathbf{L}_{\mathcal{T}_a}+\mathbf{L}_{\mathcal{T}_b}-2}$ façon de fusionner un ordre de nœuds internes de $\mathcal{T}_a$ avec un ordre de nœuds internes de $\mathcal{T}_b$.

Comme toutes ces possibilités peuvent être combinées indépendamment pour donner un ordre compatible avec $\mathcal{T}$, on obtient bien que $\mathcal{O}(\mathcal{T}) = \mathcal{O}(\mathcal{T}_a)\mathcal{O}(\mathcal{T}_b)\binom{\mathbf{L}_{\mathcal{T}_a}+\mathbf{L}_{\mathcal{T}_b}-2}{\mathbf{L}_{\mathcal{T}_a}-1}$. □

Remarquons que le nombre d'ordres compatibles dépend de la forme de l'arbre considéré. En particulier, il est linéaire avec le nombre de feuilles pour l'arbre "peigne" mais exponentiel dans le cas d'un arbre équilibré.

## 3.5 Arbres binaires orientés

Un arbre binaire *orienté* (à ne surtout pas confondre avec la notion usuelle d'orientation des graphes) est un arbre binaire dans lequel les deux descendants directs de tout nœud interne

sont distingués. On peut par exemple distinguer l'aîné et le cadet mais la nature de cette distinction n'importe pas ici et on parlera de "fils gauche" et de "fils droit". Comme l'ordre des nœuds internes, l'orientation des descendants directs est une information supplémentaire. On compte donc plus d'arbres binaires orientés que d'arbres binaires non-orientés. Pour orienter un arbre binaire dont les feuilles sont étiquetées, il est nécessaire et suffisant de faire autant de choix binaires indépendants ("quel est son fils droit?") qu'il y a de nœuds internes dans l'arbre, c'est à dire $n-1$ si l'arbre a $n$ feuilles. On a donc

- $2^{n-1}\mathcal{B}(n) = 2^{n-1}(2n-3)!!$ arbres binaires orientés et non-ordonnés avec $n$ feuilles étiquetées et
- $2^{n-1}\mathcal{O}(n) = n!(n-1)!$ arbres binaires orientés et ordonnés avec $n$ feuilles étiquetées.

De la même façon, il y a $2^{n-1}\mathcal{R}(\mathcal{T})$ arbres binaires ordonnés et orientés correspondants à un arbre binaire $\mathcal{T}$ avec $n$ feuilles étiquetées.

# 4 Probabilités d'arbres résultant de processus de branchement

Les arbres binaires (orientés, ordonnés ou non) dont les feuilles sont étiquetées apparaissent naturellement comme résultant de processus de branchement. Afin de rester le plus général possible, on considère ici des processus qui portent sur des éléments que nous appellerons *lignages* et dans lesquels le seul type d'évènement possible à un temps quelconque est qu'un lignage présent à ce temps donne naissance à un nouveau lignage. Les seules autres hypothèses faites sont que les processus considérés démarrent avec un seul lignage et qu'ils sont *homogènes en lignage*, c'est à dire qu'à n'importe quel temps $t$, aucun lignage vivant au temps $t$ n'a plus de chance qu'un autre de donner naissance à un nouveau lignage.

L'évolution d'un tel processus se représente naturellement sous la forme d'un arbre. Plus formellement, on associe à toute réalisation d'un processus de ce type, l'arbre dans lequel :

1. les nœuds internes et les feuilles de l'arbre sont en bijection respectivement avec les évènements de naissances et avec les lignages de la réalisation ;
2. pour toute feuille $x$, le nœud ancêtre direct de la feuille associée au lignage $x$ est celui correspondant au dernier évènement de naissance impliquant $x$, qui peut être soit sa propre naissance, soit la dernière fois qu'il a donné naissance à un nouveau lignage ;
3. pour toute feuille $x$, le nœud ancêtre direct du nœud interne associé à la naissance du lignage $x$ est celui correspondant au dernier évènement avant la naissance de $x$ impliquant le lignage parent de $x$.

Par construction, un tel arbre est binaire et ses feuilles sont étiquetées mais il n'est pas orienté et ses nœuds internes ne sont pas ordonnés. L'arbre défini en ajoutant aux points ci-dessus:

4. les nœuds internes sont ordonnés selon l'ordre temporel des évènements auquels ils sont associés ;

5. le fils gauche (resp. droit) du nœud interne associé à l'évènement «$x$ donne naissance à $y$» est la racine du sous-arbre contenant la feuille $x$ (resp. $y$);

est binaire et orienté avec des nœuds internes ordonnés et des feuilles étiquetées. Si l'on ne considère que le point 4 (resp. 5), l'arbre ainsi défini est seulement ordonné (resp. orienté).

En résumé, on peut associer à toute réalisation d'un processus de branchement, un arbre binaire avec feuilles étiquetées que l'on peut munir d'une orientation et/ou d'un ordre sur les nœuds internes.

L'ordre de naissance des lignages (à ne pas confondre avec l'ordre des nœuds internes) permet à la fois d'orienter et d'ordonner les nœuds internes de l'arbre (non-ordonné et non orienté) associé à une réalisation de processus. Commençons par l'orientation, l'ordre de naissance des lignages permet d'associer récursivement un lignage/feuille à tout nœud de l'arbre de la façon suivante. Chaque feuille est associée à elle-même et chaque nœud est associé à la feuille/lignage la plus ancienne parmi celles associées à ses deux descendants directs (l'un de ses descendants est donc toujours associé à la même feuille que lui). L'orientation est ensuite directe : le fils gauche d'un nœud interne est le descendant direct qui est associé à la même feuille que lui et son fils droit est son autre descendant direct. Par construction, le nœud interne correspondant à la naissance du lignage $x$ est alors l'ancêtre direct du nœud le plus profond associé à $x$ si celui-ci n'est pas la racine (sinon $x$ est le lignage initial). L'ordre de naissance des lignages détermine alors parfaitement l'ordre des nœuds internes de l'arbre défini par le point 4 ci-dessus.

Réciproquement, l'orientation et l'ordre des nœuds internes d'un arbre permet de déterminer sans ambiguïté l'ordre de naissance de ses lignages. En effet, l'orientation permet d'associer récursivement un lignage/feuille à tout nœud de l'arbre. On procède là-encore récursivement en associant à tout nœud interne la feuille associée à son fils gauche (une feuille étant associée à elle-même). Par construction, le nœud interne associé à la naissance du lignage $x$ est alors l'ancêtre direct du nœud le plus profond associé à $x$ si celui-ci existe (sinon $x$ est le lignage initial). L'ordre des nœuds détermine alors parfaitement l'ordre de ces évènements.

Comme la façon dont sont étiquetées les feuilles, donc les lignages, est arbitraire (dans le sens où elle ne dépend ni de l'arbre, ni de leur ordre de naissance), un argument de symétrie amène à la remarque suivante.

**Remarque 1.** *Si les lignages sont étiquetés arbitrairement, tous leurs ordres de naissance sont équiprobables.*

Rappelons qu'un processus est dit *homogène en lignage* si une naissance ayant lieu à un temps donné est rattachée à n'importe quel lignage présent à ce temps avec une probabilité uniforme.

**Lemme 1.** *Étant donnés $n$ lignages et leur ordre de naissance résultant d'une réalisation d'un processus de naissance homogène (sans connaître leurs relations de parenté), tous les arbres ont probabilité $\frac{1}{(n-1)!}$.*

*Preuve.* Comme le processus est homogène en lignage, le parent du $i^{\text{ème}}$ lignage est pris uniformément parmi les $(i-1)$ lignages vivant au moment de sa naissance, i.e., avec probabilité $\frac{1}{i-1}$ pour chacun d'eux. On en déduit qu'étant donné l'ordre de naissance des lignages, la probabilité jointe d'un ensemble de relations parent-enfant donné, donc un arbre, est $\frac{1}{(n-1)!}$. □

**Théorème 2.** *Soient $\mathcal{T}^{o,r}$, $\mathcal{T}^{o}$, $\mathcal{T}^{r}$ et $\mathcal{T}$ les arbres binaires avec feuilles étiquetées, respectivement à la fois orienté et ordonné, orienté mais pas ordonné, ordonné mais pas orienté et ni orienté, ni ordonné, résultant d'un processus de naissance homogène en lignages. Conditionnellement au nombre $n$ de lignages finaux, les probabilités de ces arbres sont*

$$\begin{aligned}
\mathbf{P}(\mathcal{T}^{o,r} \mid \mathbf{L}_{\mathcal{T}^{o,r}} = n) &= \frac{1}{(n-1)!n!}, \\
\mathbf{P}(\mathcal{T}^{o} \mid \mathbf{L}_{\mathcal{T}^{o}} = n) &= \frac{2^{n-1}}{(n-1)!n!}, \\
\mathbf{P}(\mathcal{T}^{r} \mid \mathbf{L}_{\mathcal{T}^{r}} = n) &= \frac{\mathcal{R}(\mathcal{T})}{(n-1)!n!} \ \textit{et} \\
\mathbf{P}(\mathcal{T} \mid \mathbf{L}_{\mathcal{T}} = n) &= \frac{2^{n-1}\mathcal{R}(\mathcal{T})}{(n-1)!n!}.
\end{aligned}$$

*Preuve.* D'après la Remarque 1, la probabilité d'un ordre de naissance de $n$ lignages donnés est $\frac{1}{n!}$. Si l'on y ajoute le Lemme 1, la probabilité jointe d'un couple arbre/ordre de naissance de $n$ lignages donnés est $\frac{1}{(n-1)!n!}$. Comme donner l'ordre des naissances équivaut à orienter l'arbre et à ordonner ses nœuds, on a bien que $\mathbf{P}(\mathcal{T}^{o,r} \mid \mathbf{L}_{\mathcal{T}^{o,r}} = n) = \frac{1}{(n-1)!n!}$.

Étant donné un arbre binaire avec feuilles étiquetées, toute paire "orientation de cet arbre et ordre compatible de ses nœud internes" correspond à un unique ordre de naissance de ses lignages et réciproquement. La probabilité d'un arbre orienté mais pas ordonné, (resp. ordonné mais pas orienté et ni orienté, ni ordonné) s'obtient en sommant la probabilité jointe d'un couple arbre/ordre de naissance sur toutes les orientations possibles (resp. tous les ordres de nœuds internes possibles et toutes les orientations et ordres possibles). Les Sections 3.5 et 3.4 nous donnent alors les probabilités $\mathbf{P}(\mathcal{T}^{o} \mid \mathbf{L}_{\mathcal{T}^{o}} = n)$, $\mathbf{P}(\mathcal{T}^{r} \mid \mathbf{L}_{\mathcal{T}^{r}} = n)$ et $\mathbf{P}(\mathcal{T} \mid \mathbf{L}_{\mathcal{T}} = n)$. □

La distribution de la forme d'un arbre binaire conditionnée à son nombre de feuilles ci-dessus (i.e., $\mathbf{P}(\mathcal{T} \mid \mathbf{L}_{\mathcal{T}} = n)$) est appelée la *distribution de Yule-Harding* et a été établie par Harding (1971).

# 5 Processus de naissance-mort

Un processus de naissance-mort suppose qu'à un instant quelconque $t$ tout individu vivant se reproduit à un taux $\lambda$ pour donner un (seul) nouvel individu et disparaît (meurt) selon un taux $\mu$ Kendall (1948). Autrement dit, en notant $N_t$ pour la variable aléatoire comptant le nombre d'individus vivants à l'instant $t$, le modèle de naissance-mort nous donne les

probabilités d'évolution du nombre d'individus durant un intervalle de temps infinitésimal $\mathrm{d}t$. Pour tout entier $n$ strictement positif, on a

$$\begin{aligned}
\mathbf{P}(N_{t+\mathrm{d}t} = n+k \mid N_t = n) &= o(\mathrm{d}t) \text{ pour tout } k \geq 2,\\
\mathbf{P}(N_{t+\mathrm{d}t} = n+1 \mid N_t = n) &= n\lambda \mathrm{d}t + o(\mathrm{d}t),\\
\mathbf{P}(N_{t+\mathrm{d}t} = n \mid N_t = n) &= 1 - n(\lambda+\mu)\mathrm{d}t + o(\mathrm{d}t),\\
\mathbf{P}(N_{t+\mathrm{d}t} = n-1 \mid N_t = n) &= n\mu \mathrm{d}t + o(\mathrm{d}t),\\
\mathbf{P}(N_{t+\mathrm{d}t} = n-k \mid N_t = n) &= o(\mathrm{d}t) \text{ pour tout } k \geq 2.
\end{aligned}$$

Si à un temps quelconque $t$, on a $N_t = 0$, alors plus aucun individu ne peut naître ou mourir et l'on a $N_{t'} = 0$ pour tout temps $t' \geq t$. Soit $I$ le nombre d'individus présents au temps 0 qui est le départ du processus (dans le cas le plus usuel $I = 1$). Posons $p_n(t) = \mathbf{P}(N_t = n)$. On a en particulier que $p_I(0) = 1$ et $p_n(0) = 0$ pour tout $n \neq I$. Pour un instant $t \geq 0$, on a que

$$p_n(t+\mathrm{d}t) = p_{n+1}(t)(n+1)\mu \mathrm{d}t + p_{n-1}(t)(n-1)\lambda \mathrm{d}t + (1 - n(\lambda+\mu)\mathrm{d}t)p_n(t) + o(\mathrm{d}t)$$

$$\text{pour tout } n \geq 1 \text{ et } p_0(t+\mathrm{d}t) = p_1(t)\mu \mathrm{d}t + p_0(t) + o(\mathrm{d}t)$$

D'où l'on tire que

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = (n+1)\mu p_{n+1}(t) + (n-1)\lambda p_{n-1}(t) - n(\lambda+\mu)p_n(t) \text{ pour tout } n \geq 1 \text{ et}$$

$$\frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = \mu p_1(t).$$

Considérant ensuite la fonction génératrice $\phi$ définie comme

$$\phi(z,t) = \sum_{n=-\infty}^{\infty} p_n(t) z^n,$$

on obtient que

$$\frac{\phi(z,t)}{\partial t} = (z-1)(\lambda z - \mu)\frac{\phi(z,t)}{\partial z}.$$

L'équation différentielle ci-dessus peut être résolue, par exemple par la méthode des carac-

téristiques, pour obtenir:

$$\begin{aligned}\phi(z,t) &= \left(\frac{\mu(1-e^{-(\lambda-\mu)t}) - z(\mu-\lambda e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t} - z\lambda(1-e^{-(\lambda-\mu)t})}\right)^I \\ &= \left(\frac{\mu(1-e^{-(\lambda-\mu)t}) - z(\mu-\lambda e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t} - \lambda(1-e^{-(\lambda-\mu)t})}\right)^I \left(\frac{1-\frac{\lambda(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}}{1-z\frac{\lambda(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}}\right)^I \\ &= \left(\frac{\mu(1-e^{-(\lambda-\mu)t}) - z(\mu-\lambda e^{-(\lambda-\mu)t})}{(\lambda-\mu)e^{-(\lambda-\mu)t}}\right)^I \\ &\qquad \left(\frac{\frac{(\lambda-\mu)e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}}{1-z\left(1-\frac{(\lambda-\mu)e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)}\right)^I.\end{aligned}$$

La formule de la binomiale négative nous indique que

$$\left(\frac{p}{1-(1-p)z}\right)^n = \sum_{k=0}^{\infty}\binom{k+n-1}{n-1}p^n(1-p)^k z^k.$$

D'où l'on tire que

$$\left(\frac{\frac{(\lambda-\mu)e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}}{1-z\left(1-\frac{(\lambda-\mu)e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)}\right)^I = \sum_{k=0}^{\infty}\binom{k+I-1}{I-1}\left(\frac{(\lambda-\mu)e^{-(\lambda-\mu)t}}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)^I\left(\frac{\lambda(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}\right)^k z^k.$$

Comme

$$\left(\frac{\mu(1-e^{-(\lambda-\mu)t}) - z(\mu-\lambda e^{-(\lambda-\mu)t})}{(\lambda-\mu)e^{-(\lambda-\mu)t}}\right)^I = \frac{\sum_{j=0}^{I}\binom{I}{j}(\mu(1-e^{-(\lambda-\mu)t}))^{I-j}(-\mu+\lambda e^{-(\lambda-\mu)t})^j z^j}{((\lambda-\mu)e^{-(\lambda-\mu)t})^I},$$

on obtient finalement que le coefficient de $z^n$ dans $\phi(z,t)$ est pour tout $n \geq 0$

$$p_n(t) = \sum_{k=0}^{\min(I,n)}\binom{I}{k}\binom{I+n-k-1}{I-1} \frac{\left(\mu(1-e^{-(\lambda-\mu)t})\right)^{I-k}\left(\lambda(1-e^{-(\lambda-\mu)t})\right)^{n-k}\left(-\mu+\lambda e^{-(\lambda-\mu)t}\right)^k}{(\lambda-\mu e^{-(\lambda-\mu)t})^{I+n-k}}.$$

En particulier pour $I=1$, le cas usuel considéré en évolution, on a

$$p_n(t) = \frac{e^{-(\lambda-\mu)t}(\lambda-\mu)^2\left(\lambda(1-e^{-(\lambda-\mu)t})\right)^{n-1}}{(\lambda-\mu e^{-(\lambda-\mu)t})^{n+1}} \text{ pour tout } n \geq 1 \text{ et,}$$

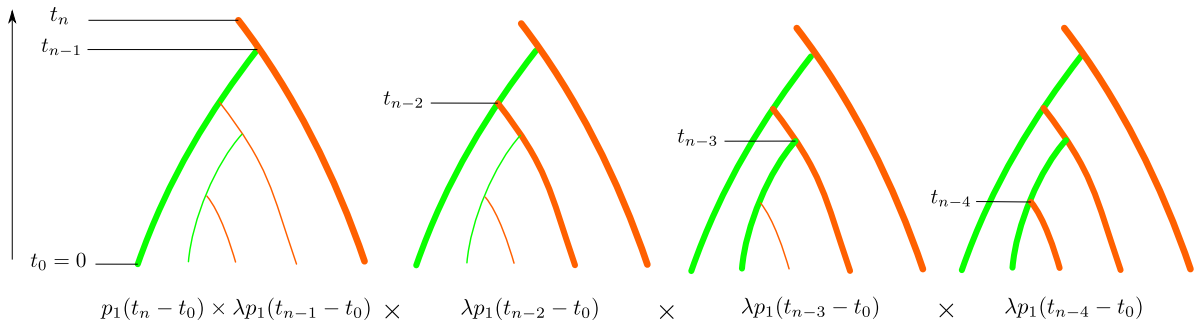


Figure 2: **Calcul de la densité de probabilité d'un arbre orienté généré selon un modèle de naissance et mort**. Chaque bifurcation dans l'arbre a pour probabilité $\lambda \mathrm{d}t$. $p_1(t)$ est la probabilité qu'une lignée née au temps $t$ donne un descendant au temps présent ($t_0 = 0$). Le terme $p_1(t_i)$ donné sous chaque arbre correspond à la densité de probabilité associée à la bifurcation se produisant au temps $t_i$. L'arbre généré ici est orienté (chaque branche a une couleur), et est noté $\tau_o$.

$$p_0(t) = \frac{\mu(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}}.$$

## 5.1 Densité de probabilité d'un arbre de naissance-mort

Nous nous intéressons ici à la densité de probabilité jointe d'un ensemble de variables aléatoires dont la combinaison définit un arbre phylogénétique enraciné avec des noeuds datés. Les variables aléatoires en question sont les suivantes : un arbre orienté sans label aux feuilles ($\tau_o$), les âges des nœuds associés ($t_1, \ldots, t_{n-1}$) et le nombre de feuilles ($n$). La densité jointe de ces trois variables est définie conditionnellement au temps auquel l'observation des $n$ lignées est réalisée ($t_0$). Ce temps est considéré comme étant unique dans le cas présent. Nous considérons également comme connues les valeurs des paramètres de naissance et mort, $\lambda$ et $\mu$, ainsi que le temps auquel la première lignée est née ($t_n$), ou temps d'origine du processus. Enfin, nous faisons l'hypothèse que le temps d'origine du processus est une variable aléatoire uniformément distribuée dans l'intervalle $[t_0, t_b]$. Nous considérons donc ici que les réalisations du processus de naissance et mort sont initiées après le temps $t_b$ et avant $t_0$. Pour des raisons pratiques, on renverse l'axe du temps par rapport à la course du processus qui va bien de $t_b$ à $t_0$ (i.e., on a $t_b \geq t_0$ même si du point de vue du processus $t_b$ est antérieur à $t_0$). La densité jointe de l'arbre est donnée ci-dessous (Fig. 2) :

$$p(\tau_o, t_1, \ldots, t_{n-1}, n | \lambda, \mu, t_n, t_0, t_b) = p_1(t_n - t_0) \prod_{i=1}^{n-1} \lambda p_1(t_i - t_0). \quad (1)$$

Le temps d'origine du processus $t_n$ n'est généralement pas connu. Il nous faut donc ici intégrer sur l'ensemble des valeurs que peut prendre cette variable aléatoire. Sous l'hypothèse

d'une distribution uniforme sur $[t_0, t_b]$, la densité d'intérêt devient alors :

$$\begin{aligned}
&p(\tau_o, t_1, \ldots, t_{n-1}, n|\lambda, \mu, t_0, t_b) \\
&= \int_{t_0}^{t_b} p(\tau_o, t_1, \ldots, t_{n-1}, n|\lambda, \mu, t_n, t_0, t_b) p(t_n|\lambda, \mu, t_0, t_b) \mathrm{d}t_n \\
&= \lambda^{n-1} \prod_{i=1}^{n-1} p_1(t_i - t_0) \left(\frac{1}{t_b - t_0}\right) \int_{t_0}^{t_b} p_1(t_n - t_0) \mathrm{d}t_n \\
&= \lambda^{n-1} \prod_{i=1}^{n-1} p_1(t_i - t_0) \left(\frac{1}{t_b - t_0}\right) \\
&\quad \frac{\lambda - \mu}{\mu} \left(\frac{1}{\lambda - \mu e^{-(\lambda-\mu)t_0}} - \frac{1}{\lambda - \mu e^{-(\lambda-\mu)t_b}}\right) \qquad (2)
\end{aligned}$$

Lorsque $t_0 = 0$ et $t_b \to \infty$, l'intégrale présente dans l'expression précédente est égale à $1/\lambda$. On obtient alors :

$$p(\tau_o, t_1, \ldots, t_{n-1}, n|\lambda, \mu, t_0 = 0, t_b \to \infty) \propto \lambda^{n-2} \prod_{i=1}^{n-1} p_1(t_i),$$

et le symbole «proportionnel à» ($\propto$) devient nécessaire car $1/(t_b - t_0) \to 0$ lorsque $t_b \to \infty$ et $t_0 = 0$.

Les arbres phylogénétiques ne sont généralement pas orientés et possèdent des labels aux feuilles, souvent associés aux noms des séquences génétiques analysées. Il existe $2^{n-1}$ arbres orientés qui diffèrent exclusivement par la couleur de leurs branches (voir Fig. 2). Chacune de ces combinaisons est équiprobable. Il existe également $n!$ permutations de labels aux feuilles de l'arbre. Chacune de celle-ci est, là-encore, équiprobable. La densité de probabilité jointe d'un arbre non-orienté, $\tau$, et présentant des labels aux feuilles est donc donnée par l'expression ci-dessous :

$$p(\tau, t_1, \ldots, t_{n-1}, n|\lambda, \mu, t_0 = 0, t_b \to \infty) \propto \frac{2^{n-1}}{n!} \lambda^{n-2} \prod_{i=1}^{n-1} p_1(t_i)$$

Le nombre de feuilles dans l'arbre est généralement déterminé par les contingences de l'échantillonnage plutôt que par le processus stochastique générant les données. Il est alors

préférable de donner la densité jointe de l'arbre conditionnellement à $n$. On a donc:

$$\begin{aligned}
&p(\tau, t_1, \ldots, t_{n-1} | n, \lambda, \mu, t_0 = 0, t_b \to \infty) \\
&= \frac{p(\tau, t_1, \ldots, t_{n-1}, n | \lambda, \mu, t_0 = 0, t_b \to \infty)}{\Pr(n | \lambda, \mu, t_0 = 0, t_b \to \infty)} \\
&= \frac{p(\tau, t_1, \ldots, t_{n-1}, n | \lambda, \mu, t_0 = 0, t_b \to \infty)}{\int_{t_0}^{t_b} p_n(t_n) p(t_n | \lambda, \mu, t_0 = 0, t_b \to \infty) \mathrm{d}t_n} \\
&= \frac{\frac{2^{n-1}}{n!} \lambda^{n-2} \prod_{i=1}^{n-1} p_1(t_i)}{\int_{t_0}^{t_b} p_n(t_n) \mathrm{d}t_n} \\
&= \frac{\frac{2^{n-1}}{n!} \lambda^{n-2} \prod_{i=1}^{n-1} p_1(t_i)}{1/\lambda n} \\
&= \frac{2^{n-1}}{(n-1)!} \lambda^{n-1} \prod_{i=1}^{n-1} p_1(t_i) \\
&= \prod_{i=1}^{n-1} \left( \frac{2\lambda}{i} \right) p_1(t_i) \qquad (3)
\end{aligned}$$

Il est également utile de donner la densité jointe de l'arbre conditionnellement à $t_{n-1}$, c'est-à-dire l'âge de la première divergence. Les analyses de datation moléculaire s'appuient souvent sur une distribution marginale de l'âge de ce nœud dérivé de l'analyse de fossiles. Le produit de la densité jointe de l'arbre sachant $t_{n-1}$ selon le processus de naissance et mort par la densité marginale derivée des fossiles permet ainsi de calibrer la datation moléculaire de manière adéquate. On a :

$$\begin{aligned}
&p(\tau, t_1, \ldots, t_{n-2} | t_{n-1}, n, \lambda, \mu, t_0 = 0, t_b \to \infty) \\
&= \frac{p(\tau, t_1, \ldots, t_{n-1} | n, \lambda, \mu, t_0 = 0, t_b \to \infty)}{p(t_{n-1} | n, \lambda, \mu, t_0 = 0, t_b \to \infty)} \\
&= \frac{p(\tau, t_1, \ldots, t_{n-1} | n, \lambda, \mu, t_0 = 0, t_b \to \infty)}{\sum_\tau \int_{0 \le t_1 \le \ldots \le t_{n-1}} \ldots \int p(\tau, t_1, \ldots, t_{n-1} | n, \lambda, \mu, t_0 = 0, t_b \to \infty) \mathrm{d}t_1 \ldots \mathrm{d}t_{n-2}} \\
&= \frac{\prod_{i=1}^{n-1} p_1(t_i)}{\sum_\tau \frac{1}{\mu^{n-2}(n-2)!} p_0(t_{n-1})^{n-2} p_1(t_{n-1})} \\
&= \frac{(n-2)!}{\sum_\tau 1} \prod_{i=1}^{n-2} \mu \frac{p_1(t_i)}{p_0(t_{n-1})}.
\end{aligned}$$

$\sum_\tau 1$ correspond au nombre d'arbres non-orientés avec labels aux feuilles. Celui-ci est égal à $\frac{(n-1)!n!}{2^{n-1}}$ (voir Section 3.3 de ce chapitre). La densité recherchée est donc donnée par

l'expression suivante :

$$p(\tau, t_1, \ldots, t_{n-2} | t_{n-1}, n, \lambda, \mu, t_0 = 0, t_b \to \infty) = \frac{2^{n-1}}{n!(n-1)} \prod_{i=1}^{n-2} \mu \frac{p_1(t_i)}{p_0(t_{n-1})} \quad (4)$$

L'expression ci-dessus est celle donnée par Yang and Rannala (1997, p. 718, Eq. 6). Notons ici que la densité de l'arbre est dérivée conditionnellement au nombre de lignées échantillonnées, $n$. Lorsque la valeur de $n$ est determinée par le processus de génération d'arbres, il est alors préférable d'utiliser l'Equation 3 comme support pour l'inférence des paramètres de naissance et mort. Ainsi, lors des phases précoces d'une épidémie, le nombre de lignées échantillonnées est parfois proche ou tout au moins fortement corrélé au nombre total d'agents pathogènes en circulation. La valeur de $n$ est alors porteuse d'information concernant le processus de naissance et mort et il est utile d'incorporer cette information dans le cadre de l'inférence. Lorsque $n$ est determiné par l'échantillonnage, et notamment par d'éventuelles contraintes financières limitant le nombre d'individus échantillonnés, il fait alors sens de se référer à l'Equation 4 ci-dessus pour l'inférence. Ainsi, cette dernière équation est plus appropriée lorsqu'il n'existe pas de raison de penser que $n$ est proche ou corrélé au nombre total de lignées vivantes au moment de l'échantillonnage, comme cela peut-être le cas lorsque l'abondance de l'agent pathogène dépasse largement les capacités de surveillance de l'épidémie.

# 6 Le coalescent

Deux particules coalescent lorsqu'à l'issue de leur collision l'une avec l'autre, elles n'en forment plus qu'une. Dans le cas où les particules en question correspondent à des lignées, la scission de l'une d'elle, suite à un événement de spéciation ou de duplication d'un gène ou d'un génome, correspond à une «coalescence» lorsque la flèche du temps est inversée, c.a.d . deux lignées coalescent pour n'en former plus qu'une. Il existe une grande variété de processus de coalescence. Certains de ces processus impliquent la fusion de plus de deux lignées par exemple. D'autres autorisent de multiples fusions au même instant.

Le coalescent proposé par Kingman est peut-être le plus simple parmi l'ensemble de ces processus et c'est celui que nous decrirons ci-dessous. Malgré sa relative simplicité, ce modèle a permis de grandes avancées dans le domaine de la biologie des populations. Ce succès s'explique par son étroite connection avec plusieurs modèles clasiques en génétique des populations, tels que celui de Wright-Fisher et celui de Moran, comme nous le verrons plus loin dans ce chapitre.

Soit $\mathcal{P}_n$ l'ensemble des partitions d'un ensemble de $n$ éléments qui correspondent ici à nos $n$ espèces ou gènes. Pour $n = 3$ par exemple, avec $[n] = \{a, b, c\}$ désignant les noms (ou labels) des trois espèces, alors cet ensemble de partitions est le suivant :

$$\mathcal{P}_3 = \{\{a|b|c\}, \{a, b|c\}, \{a, c|b\}, \{b, c|a\}, \{a, b, c\}\}.$$

La partition $\{a|b|c\}$ est constituée de trois «blocs», chaque bloc correspondant ici à une lignée, tandis que $\{a,b|c\}$ est composée de deux blocs. De manière générale, chaque bloc correspond à une particule. Le coalescent de Kingman est un processus stochastique noté $(\Pi_t^n, t \geq 0)$ avec $\mathcal{P}_n$ pour espace des états. Il est défini de la façon suivante:

- L'état initial $\Pi_0^n$ est celui correspondant à l'ensemble des singletons ($\{a|b|c\}$ dans l'exemple ci-dessus).
- $(\Pi_t^n, t > 0)$ est un processus Markovien dont les taux de transition de $x$ vers $y$, notés $q(x,y)$ sont nuls si $y$ n'est pas obtenu par fusion de deux blocs de $x$ et égaux à 1 sinon. Par exemple, $q(\{a,b|c\},\{a,c|b\}) = 0$ et $q(\{a|b|c\},\{a,c|b\}) = 1$.

En d'autres termes, l'état initial du processus stochastique qui nous intéresse ici correspond à un ensemble de $n$ espèces pris au temps $t = 0$. À chacune des $n(n-1)/2$ paires de blocs est associée une alarme qui sonne après une durée aléatoire. Le processus de fusion de paires blocs étant markovien, la distribution de cette dernière est une exponentielle de paramètre 1. La première alarme sonne à un temps correspondant au minimum des $n(n-1)/2$ durées aléatoires. Ce temps est donc distribué selon une exponentielle de paramètre $n(n-1)/2$. Le nombre de blocs passe ensuite de $n$ à $n-1$ et la fusion de deux blocs pris parmis $n-1$ a lieu après un temps aléatoire de distribution exponentielle de paramètre $(n-1)(n-2)/2$. Ce processus se poursuit jusqu'à l'obtention d'un unique bloc regroupant l'ensemble des $n$ lignées initiales.

## 6.1 Liens avec les modèles «classiques» en génétique des populations

Le coalescent de Kingman, ou $n$-coalescent, est remarquable de part ses connections fortes avec divers modèles mathématiques utilisés en génétique des populations, établissant ainsi un lien d'équivalence entre ces derniers. Ainsi, les généalogies obtenues sous le modèle de Moran sont identiques, d'un point de vue probabiliste, à celles dérivées du $n$-coalescent. Selon le modèle de Moran, chaque individu au sein d'une population de $N$ individus, meurt au bout d'un temps aléatoire et est remplacé par un individu pris au hasard uniformément parmi $N$. La distribution de la durée de vie de chaque individu est une exponentielle de paramètre 1. Sur la Figure 3, cette durée correspond au temps écoulé entre deux disques successifs le long d'une des $N = 5$ lignes verticales.

## 6.2 Modèle de Moran

Pour déterminer la distribution des temps de coalescence au sein d'une généalogie générée selon le modèle de Moran, on s'intéresse à un échantillon de $n$ individus pris parmi $N$ au temps $t = 0$ et on suit l'ensemble des $N$ lignées constituant la population en remontant du présent vers le passé. Le premier disque survient après une durée distribuée exponentiellement de paramètre $N$ (puisque le temps ecoulé entre deux disques successifs le long d'une des lignées est une exponentielle de paramètre 1). Ce disque correspond à une coalescence

entre les lignées de notre échantillon avec une probabilité égale à $\frac{n}{N} \times \frac{n-1}{N-1} \times (1 - \frac{1}{N}) = \frac{n(n-1)}{N^2}$. En effet, la probabilité que la lignée sur laquelle se trouve le premier disque soit issue de notre échantillon est de $n/N$. Aussi, il existe $n-1$ lignées avec lesquelles cette lignée peut fusionner pour donner une coalescence, parmi $N-1$ lignées au total. Enfin, la probabilité pour que ce même disque ne corresponde pas au remplacement d'une lignée par elle-même est de $(1 - \frac{1}{N})$. Ainsi, les lignées de notre échantillon coalescent à un taux égal à $n(n-1)/N^2 \times N = n(n-1)/N$ coalescences par unité de temps. Ce taux est distinct de celui du coalescent de Kingman, qui est de $n(n-1)/2$. Cependant, lorsque l'unité de temps du modèle de Moran est remplacée par une nouvelle unité, l'unité de temps du coalescent, telle qu'une unité de temps du coalescent vaut $N/2$ unités de temps de Moran, alors le taux de coalescence mesuré en espérance du nombre de coalescences par unité de temps du coalescent est de $n(n-1)/2$. Ainsi, la distribution des temps de coalescence selon le modèle de Moran est celle du coalescent de Kingman après une simple mise à l'échelle de l'unité de temps.

Notons que lorsque le temps écoulé entre deux disques successifs, pris sur l'ensemble des $N$ lignées, est fixe et égal à une unité de temps dit «de Moran» (plutôt qu'une quantité aléatoire distribuée suivant une exponentielle de paramètre $N$) alors la probabilité que deux lignées fusionnent est de $2/N^2$. Ainsi, la probabilité que deux lignées n'aient pas d'ancêtre commun avant $k$ unités de temps de Moran est alors égale à $(1-2/N^2)^k$. Lorsque l'unité de temps de Moran est mise à l'échelle de manière à ce que $k$ unités de ce temps équivalent à $(N^2/2)t$ unités du «nouveau» temps (cad le temps en nombre d'unités de Moran est donné par le temps en unités du coalescent multiplié par $N^2/2$), alors $\lim_{N\to\infty}(1-2/N^2)^{N^2t/2} = e^{-t}$. Le processus de Moran correspond donc encore, cette fois-ci asymptotiquement ($N \to \infty$), à un coalescent lorsqu'une unité du nouveau temps (le temps du coalescent) équivaut à $2/N^2$ unités de temps de Moran.

## 6.3 Modèle de Wright-Fisher

Le modèle de Wright-Fisher est central en génétique des populations car il permet de quantifier le phénomène de dérive génétique. Tout comme pour le modèle de Moran, nous considérons ici que la taille de population est fixe et égale à $N$. En revanche, les générations sont ici non-chevauchantes : les $N$ parents de la génération $t$ produisent $N$ descendants qui remplacent leurs parents, devenant ainsi les $N$ parents de la génération $t+1$. Les liens de parentés entre individus de deux générations successives sont définis en considérant la manière dont chaque descendant «choisit» un parent. Il s'agit ici d'un tirage aléatoire uniforme avec remise. Deux descendants choisissent donc le même parent avec une probabilité égale à $1/N$. La probabilité que deux individus n'aient pas d'ancêtre commun en remontant $g$ générations dans le passé est donc égale à $(1-1/N)^g$. Lorsque l'unité de temps, la génération, est mise à l'échelle de manière à ce que $g$ générations équivalent à $Nt$ unités du «nouveau» temps du coalescent, alors $\lim_{N\to\infty}(1-1/N)^{Nt} = \exp(-t)$. Le temps de coalescence est donc ici distribué, pour une paire d'individus, selon une exponentielle de paramètre 1. Plus généralement, tout comme pour le modèle de Moran, la distribution des temps de coalescence obtenue suivant le modèle de Wright-Fisher est celle du coalescent lorsque le

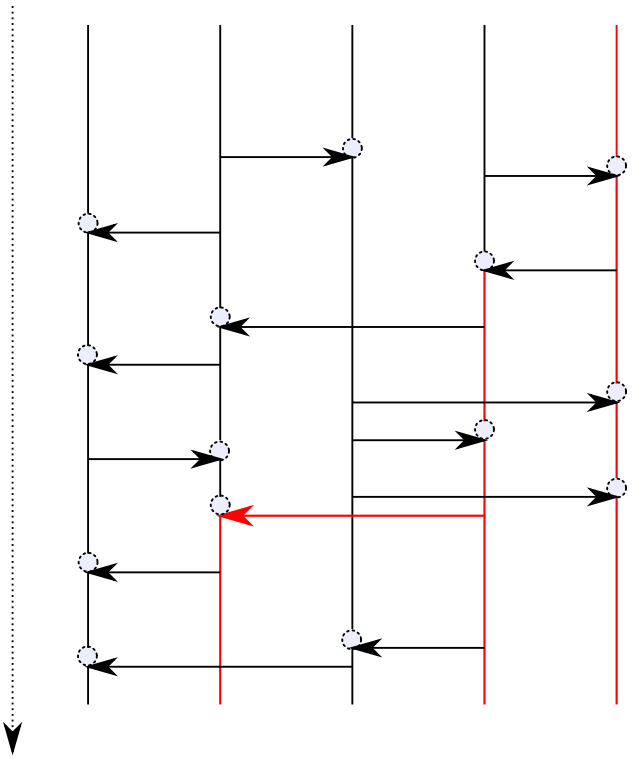

Figure 3: **Modèle de Moran.** Le temps s'écoule du haut vers le bas. Chaque disque correspond à la mort d'un individu. À l'issue d'un tel évenement, l'individu en question est immédiatement remplacé par un autre. Les lignées rouges constituent un échantillon de taille $n = 3$, pris parmi la population de taille $N = 5$.

temps est mis à l'échelle de manière adéquate.

## 6.4 Modèle générique

Les modèles de Moran et Wright-Fisher sont fortement liés au coalescent de Kingman dans le sens où la distribution des temps auxquels les lignées fusionnent au sein de la généalogie d'un échantillon est donnée (dans le cas du modèle de Moran avec une durée de vie de chaque individu distribuée selon une exponentielle) ou tend (dans celui de Wright-Fisher et de Moran avec un temps fixe entre deux événements de reproduction/extinction successifs) vers la distribution définie par le processus stochastique du coalescent de Kingman. Il est donc ici naturel de définir plus précisement les contours d'un modèle populationnel générique dont la distribution des temps de coalescence est donnée par le $n$-coalescent.

Soit $X_i$ le nombre de descendants directs du parent $i$. Le modèle générique est tel que $\sum_{i=1}^{N} X_i = N$, cad, le nombre d'individus dans la population est fixe. Les $X_i$, $i = 1, \ldots, N$, ont tous la même distribution. Ils ne sont cependant pas indépendents puisqu'ils sont contraints par la taille de la population. Le modèle générique est dit «échangeable» car l'ordre dans lequel les $N$ parents sont considérés dans la somme $\sum_{i=1}^{N} X_i$ n'a pas d'importance. Les $X_i$ étant tous identiquement distribués implique que le modèle générique est ignorant vis-à-vis de la sélection naturelle, et ceci à chaque génération. En faisant l'hypothèse que les générations sont non-chevauchantes, la probabilité $p$ que deux descendants aient le même parent est alors donnée par l'expression suivante:

$$\begin{aligned} p &= \mathbb{E}\Big(\sum_{i=1}^{N}\Big(\frac{X_i}{N}\Big)\Big(\frac{X_i-1}{N-1}\Big)\Big) \\ &= \frac{1}{N(N-1)}\sum_{i=1}^{N}\mathbb{E}\Big(X_i(X_i-1)\Big) \\ &= \frac{1}{N(N-1)}\sum_{i=1}^{N}\mathbb{V}\Big(X_i\Big) \\ &= \frac{\mathbb{V}\Big(X_1\Big)}{N-1}. \end{aligned}$$

Notons ici tout d'abord que $\sum_{i=1}^{N} X_i = N$ implique que $\sum_{i=1}^{N} \mathbb{E}(X_i) = N$ et donc $\mathbb{E}(X_i) = 1$ (et $[\mathbb{E}(X_i)]^2 = 1$) car tous les $X_i$ ont la même espérance. Ils ont également la même variance et donc $\sum_{i=1}^{N} \mathbb{V}\Big(X_i\Big) = N\mathbb{V}\Big(X_1\Big)$. Lorsque l'on considère que le temps s'écoule de manière discrète, la probabilité que deux lignées n'aient pas d'ancêtre commun en remontant $k$ unités de temps dans le passé est de $(1 - \mathbb{V}(X_1)/(N-1))^k$ et le coalescent de Kingman est donc obtenu ici asymptotiquement en changeant l'échelle du temps tel que $k$ unité de temps du modèle générique equivalent à $(N-1)t/\mathbb{V}(X_1)$ unité de temps du coalescent.

Ainsi, pour le modèle générique évoqué ci-dessus, la convergence vers le coalescent de Kingman s'effectue en accélérant le temps par un facteur proportionel à la taille de population

et inversement proportionnel à la variance du nombre de descendants produits par un parent. Ce facteur, communement appelé «taille efficace de population» et généralement noté $N_e$, est donc un scalaire permettant de mettre à la même échelle (celle du coalescent de Kingman) toute une variété de modèles populationnels.

La taille efficace de population est un concept central en génétique des populations car elle permet de quantifier la *dérive génétique*. Ainsi, lorsque l'on considère le modèle de Wright-Fisher en avant dans le temps (cad si $g > h$, alors $g$ désigne un point dans le temps situé dans le futur par rapport à $h$), pour une population constituée de deux types d'individus, la probabilité $p_g$ d'échantillonner deux individus de type distinct à la génération $g > 0$ est égale à $p_g = p_0(1 - 1/N)^g$. Cette probabilité, dite d'hétérozygotie, décroit donc au même taux, en avant dans le temps, que le taux de coalescence, mesuré cette fois-ci en inversant la flèche du temps.

## 6.5 Densité de probabilité d'un arbre généré par le coalescent

Le processus de fusion de lignées étant markovien, la dérivation de la densité de probabilité d'un arbre généré suivant le coalescent de Kingman est quasi-immédiate. Il est cependant nécessaire de s'attarder sur l'unité de temps utilisée. Nous considèrons ici que les ages des nœuds sont exprimés en unité de temps calendaire, comme c'était déjà le cas avec le modèle de naissance et mort. Sous le modèle de Wright-Fisher, le nombre de générations multiplié par la taille efficace de la population $N_e$ donne l'unité de temps correcte du coalescent de Kingman. Si l'on note $\rho$ la durée moyenne d'une génération, exprimée en temps calendaire, alors l'unité de temps du coalescent est donnée par le produit de $\rho N_e$ par le temps calendaire. On a ainsi :

$$\begin{aligned} & p(\tau, t_1, \ldots, t_{n-1} | n, N_e, \rho) \\ &= \prod_{i=2}^{n} \frac{i(i-1)}{2N_e\rho} \exp\left[\frac{i(i-1)}{2N_e\rho}\left(t_{n-i+1} - t_{n-i}\right)\right] \times \frac{2}{i(i-1)} \\ &= \left(\frac{1}{N_e\rho}\right)^{n-1} \exp\left[\sum_{i=2}^{n} \frac{i(i-1)}{2N_e\rho}\left(t_{n-i+1} - t_{n-i}\right)\right] \end{aligned}$$

Le produit ci-dessus porte sur l'ensemble des périodes de temps entre paires de nœuds "successifs", cad les paires dont les âges sont $(t_{i+1}, t_i)$ pour tout $i$ allant de 0 à $n-2$. Chaque terme dans ce produit correspond à la multiplication d'une densité exponentielle de paramètre $i(i-1)/2N_e$ par la probabilité que deux lignées (portant un label) parmi $i$ fusionnent. Notons ici que seul le produit $\rho N_e$ peut être estimé à partir d'un arbre. Ainsi, l'inférence de la taille efficace d'une population n'est possible qu'à la condition de connaître la durée moyenne d'une génération (et inversement).

# 7 Conclusions

Ce chapitre donne les principaux éléments de combinatoire concernant les arbres et détaille les dérivations des densités de probabilité d'arbres encracinés générés suivant les processus de naissance et mort et suivant le coalescent de Kingman. L'exposition des étapes de ces calculs permet de mettre en evidence les principales hypothèses sur lesquelles reposent ces deux modèles et d'en évaluer la pertinence biologique. Ainsi, ceux-ci font tous deux l'hypothèse que les lignées échantillonnées sont échangeables : la densité de probabilité des arbres reste identique quelle que soit la permutation des labels aux feuilles de l'arbre. Cette hypothèse simplifie grandement les calculs. Du point de vue biologique, elle implique cependant que les différentes lignées «se valent» toutes et ne permet donc pas à la sélection naturelle d'agir.

Les deux modèles de génération d'arbres présentés dans ce chapitre font par ailleurs l'hypothèse que leur paramètres, c'est à dire la taille efficace de population et les taux de naissance et mort, ne varient pas au cours de l'évolution. Par une mise à l'échelle de l'unitée de temps, où la taille de population est exprimée comme une fonction du temps (généralement une exponentielle), il est possible d'ajuster aux données un coalescent dont le paramètre varie au cours de l'évolution (tout en restant constant entre lignées prises au même instant). Ce type de modèle présente de multiples déclinaisons, les plus connues étant les suivantes : «classic skyline», «classic skyline», «generalized skyline», «Bayesian skyline», «Bayesian skyride» (voir Pybus *et al.* (2000) pour le premier modèle de type «skyline» et Ho and Shapiro (2011) pour une revue).

Les variations de tailles de populations détectées à partir de ce type de coalescent peuvent néanmoins se confondre avec des changements dans la structuration de la population au cours du temps. Ainsi, la transition d'une population panmictique vers une population scindée en dèmes distincts entraîne un accroissement des temps de coalescences, accroissement d'autant plus important que le taux de migration entre dèmes est faible. Or le même phénomène est également attendu pour une population panmictique dont la taille efficace augmente (voir Mazet *et al.* (2016) pour plus de détails). Des limitations similaires affectent également les modèles de types naissance et mort autorisant les taux des deux types d'événement à varier dans le temps. Ainsi, Louca and Pennell (2020) ont montré que la vraisemblance d'un arbre donné peut être la même pour une infinité de trajectoires distinctes de taux de naissance et mort (où une trajectoire décrit la variation des taux dans le temps, l'ensemble les lignées à un instant donné partagent toutes les mêmes taux). Ce résultat remet en question un certain nombre d'études en macroévolution. Ainsi, certaines des études se focalisant sur les corrélations éventuelles entre patterns de diversification et facteurs environnementaux doivent être re-interprétées avec prudence (voir Pagel (2020) et Morlon *et al.* (2020) pour une note plus optimiste sur cette question).

Par ailleurs, les deux classes de modèles introduites dans ce chapitre font des prédictions bien distrinctes concernant le nombre total de lignées asymptotiquement. Les modèles de Wright-Fisher et Moran, et donc pour le coalescent qui en découle, reposent sur une contrainte stricte de constance de taille de population. Pour le modèle de naissance et mort, la situation est plus complexe car, pour des temps infiniment longs, l'espérance du nombre total de lignées tend soit vers une quantité infinie soit vers zero. D'un point de vue bi-

ologique, il semble pertinent de se baser sur un modèle selon lequel la population ou l'espèce étudiée s'éteint obligatoirement à terme. Il est en revanche nécessaire d'accepter ici la possibilité que la population/l'espèce «explose» parfois pour atteindre une taille infinie, ce qui est évidemment plus difficile à justifier d'un point de vue biologique.

Des questions relatives à l'échantillonnage se posent également. Lorsque le nombre de lignées échantillonées est fixé par le plan experimental, comme c'est généralement le cas pour le coalescent et, dans certaines situations, pour le modèle de naissance et mort, il fait alors sens d'exprimer la distribution de probabilité des arbres conditionnelement à la taille de l'échantillon. Néanmoins, dans d'autres situations, le nombre de lignées échantillonnées doit être considéré comme une quantité aléatoire dont la valeur dérive du processus stochastique générant l'arbre. Du point de vue de l'inférence des paramètres démographiques, c'est alors la densité jointe de l'arbre et du nombre de lignées qui décrit au mieux les données. Bien que l'influence du type d'échantillonnage réalisé est importante du point de vue de l'inférence de paramètres évolutifs et des tests d'hypothèses biologiques, ces aspects n'ont été pris en compte que partiellement jusqu'à ce jour. Ainsi, dans le contexte de l'inférence en phylodynamie (voir Chapitre **??**), seules quelques études constatent et tentent de rectifier les biais dans l'inférence de tailles efficaces de populations du à l'échantillonnage temporel des séquences (Cappello and Palacios, 2020; Parag *et al.*, 2020; Karcher *et al.*, 2016; Volz and Frost, 2014). Stadler (2010) et Stadler *et al.* (2013) décrivent également une généralisation du modèle de naissance et mort prenant en compte l'intensité d'échantillonnage et sa variation dans le temps.

# References


Aldous, D. J. *et al.* (1999). Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists. *Bernoulli*, **5**(1), 3–48.

Berestycki, N., Etheridge, A., and Hutzenthaler, M. (2009). Survival, extinction and ergodicity in a spatially continuous population model. *Markov Processes and Related Fields*, **15**, 265–288.

Cappello, L. and Palacios, J. A. (2020). Adaptive preferential sampling in phylodynamics. *arXiv preprint arXiv:2009.02307*.

Harding, E. F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, **3**(1), 44–77.

Ho, S. Y. and Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular ecology resources*, **11**(3), 423–434.

Karcher, M. D., Palacios, J. A., Bedford, T., Suchard, M. A., and Minin, V. N. (2016). Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS computational biology*, **12**(3), e1004789.

Kendall, D. (1948). On the generalized "birth-and-death" process. *The annals of mathematical statistics*, **19**(1), 1–15.

Louca, S. and Pennell, M. W. (2020). Extant timetrees are consistent with a myriad of diversification histories. *Nature*, **580**(7804), 502–505.

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, **116**(4), 362–371.

Möhle, M. (2000). Ancestral processes in population genetics—the coalescent. *Journal of Theoretical Biology*, **204**(4), 629–638.

Morlon, H., Hartig, F., and Robin, S. (2020). Prior hypotheses or regularization allow inference of diversification histories from extant timetrees. *bioRxiv*.

Pagel, M. (2020). Evolutionary trees can't reveal speciation and extinction rates.

Parag, K. V., du Plessis, L., and Pybus, O. G. (2020). Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. *Molecular Biology and Evolution*.

Prüfer, H. (1918). Neuer beweis eines satzes über permutationen. *Arch. Math. Phys*, **27**(1918), 742–744.

Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, **155**(3), 1429–1437.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, **267**(3), 396–404.

Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, **110**(1), 228–233.

Volz, E. M. and Frost, S. D. (2014). Sampling through time and phylodynamic inference with coalescent and birth–death models. *Journal of The Royal Society Interface*, **11**(101), 20140945.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a markov chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.