UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LE *P*-COALESCENT : UN NOUVEAU MODÈLE PROBABILISTE INTÉGRANT LA GÉNÉTIQUE FAMILIALE AU PROCESSUS DE COALESCENCE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

RENAUD ALIE

JANVIER 2020

REMERCIEMENTS

Merci à mes codirecteurs Fabrice Larribe et Sorana Froda pour leur support, leurs contributions, leur rigueur et leur enthousiasme.

Merci à mes parents qui, heureusement, m'ont toujours encouragé et poussé à aller à l'université.

Merci à mes amis et collègues pour les discussions de mathématiques souvent éclairantes.

Merci aux professeurs du département de mathématiques de l'UQAM qui m'ont offert une éducation de qualité.

TABLE DES MATIÈRES

LISTE DES TABLEAUX v
LISTE DES FIGURES vi
RÉSUMÉ x
INTRODUCTION 1
CHAPITRE I GÉNÉTIQUE ET GRAPHES 4
1.1 Graphe dirigé acyclique 4
1.2 Arbres enracinés et génétique 8
CHAPITRE II MODÈLE DE WRIGHT-FISHER ET PROCESSUS DE COALESCENCE 13
2.1 Modèle de Wright-Fisher 13
2.1.1 Généalogie sous Wright-Fisher 16
2.1.2 Temps de coalescence et distance 17
2.1.3 Généalogies exprimées en termes de partitions 18
2.1.4 Processus ancestral 19
2.1.5 Propriétés des généalogies sous Wright-Fisher 22
2.2 Processus de coalescence 25
2.2.1 Propriétés du processus de coalescence 26
2.2.2 Simulation de généalogies 28
2.2.3 Dérivation du processus de coalescence à partir du modèle de Wright-Fisher 30
CHAPITRE III GÉNÉTIQUE FAMILIALE 36
3.1 Pedigree 36
3.1.1 Ploïdie 38
3.1.2 Indicateurs de méiose 40

3.2 Généalogies 40
3.2.1 Simulation de généalogies 43
3.2.2 État IBD 46
3.2.3 Probabilité de coalescence 50
3.3 Temps et distance sur la généalogie 54
3.3.1 Temps total 54
3.3.2 Distance 55
CHAPITRE IV *P*-COALESCENT 57
4.1 Généalogies 58
4.1.1 Construction de la généalogie 60
4.1.2 Simulation de généalogies 62
4.2 Temps sous le *p*-coalescent 63
4.2.1 Distance 63
4.2.2 Temps total 69
CHAPITRE V MODÈLE DE MUTATION 72
5.1 Mutation et généalogie 72
5.2 Sites polymorphes 77
5.2.1 Modèle des sites infiniment nombreux 79
5.2.2 Position des sites mutants 81
5.2.3 Nombre de sites polymorphes 85
5.3 Mesures du taux de mutation 88
5.3.1 Processus de coalescence 88
5.3.2 *P*-coalescent 90
CONCLUSION 93
APPENDICE A EXPÉRIENCES SUR LE CONCEPT DE GÉNÉALOGIE EN GÉNÉTIQUE FAMILIALE 95
BIBLIOGRAPHIE 104

LISTE DES TABLEAUX

Tableau Page

A.1 Résultats de l'expérience 1 98

A.2 Résultats de l'expérience 2 99

A.3 Résultats de l'expérience 3 100

A.4 Résultats de l'expérience 4 101

A.5 Résultats de l'expérience 5 102

A.6 Résultats de l'expérience 6 103

LISTE DES FIGURES

Figure		Page
1.1	Un graphe dirigé avec sommets $\mathbf{V} = \{1,2,3,4\}$ et arêtes $\mathbf{A} = \{[1,2],[2,3],[3,1],[3,4],[4,1)]\}$.	5
1.2	Un graphe dirigé acyclique avec les sommets $\mathbf{V} = \{1,2,\ldots,8\}$ dans un ordre topologique (résultat 1.1).	6
1.3	En rouge les parents du sommet $\mathbf{v}$; en bleu, les ancêtres.	6
1.4	En rouge les enfants du sommet $\mathbf{v}$; en bleu, les descendants. . . .	7
1.5	Un arbre enraciné en 1. Les feuilles sont les sommets $4,6,9,11,12$.	10
1.6	La structure de l'arbre de la figure 1.5 est préservée et les branches sont pondérées par des longueurs.	11
1.7	L'arbre minimal contenant les sommets 8 et 9 est en bleu. La taille de l'arbre ou la distance $d(8,9)$ est 5.	12
2.1	Exemple du processus de reproduction entre la génération τ et $\tau-1$ sous le modèle de Wright-Fisher pour une population de $N = 8$ individus.	14
2.2	Exemple d'une réalisation du modèle de Wright-Fisher sur 10 générations pour une population de $N = 8$ individus.	15
2.3	(a) La généalogie de g_1, g_2 et g_3 est superposée sur une réalisation du modèle de Wright-Fisher. (b) La généalogie (redressée) est un arbre enraciné.	16
2.4	(a) Une généalogie sous forme d'arbre réduit pondéré. (b) Les temps entre les évènements de coalescence résument les longueurs de branches.	17
2.5	Il existe $S(4,3)3! = \binom{4}{2}3! = 36$ fonctions surjectives de $\{1,2,3,4\}$ vers $\{1,2,3\}$. Chacune implique une seule coalescence d'exactement deux lignées.	21

2.6 Une réalisation moyenne du processus de coalescence avec $T_k = E[T_k]$ pour $k = 2, \ldots, 5$ ($n = 5$). 28

3.1 Pedigree sur trois générations illustrant la relation entre les frères et soeurs 11 et 12 et leur cousin 13. Les fondateurs sont les individus 1 à 6. 37

3.2 Un exemple de pedigree de hauteur $\Lambda(\mathcal{P}) = 3$. Les deux chemins de longueur maximale sont en bleu. 38

3.3 Pedigree montrant les relations familiales entre les gènes pour des individus diploïdes. Les liens en gris sont symboliques ; ils montrent l'appartenance de deux gènes à un même individu, mais ne font pas partie du graphe. Par exemple, les parents du gène 13 sont les gènes 1 et 2 qui appartiennent au même individu. 39

3.4 Une réalisation des indicateurs de méiose détermine le parent duquel (arête en noir) un gène est hérité (sauf pour les fondateurs). Sur la figure, la valeur 0 ou 1 correspond à choisir, respectivement, le parent de gauche ou de droite. 41

3.5 Pour un membre du pedigree, il existe un seul chemin reliant un fondateur à celui-ci lorsque les indicateurs de méiose **S** sont réalisés. 42

3.6 Un sous-graphe du pedigree est en rouge. Celui-ci correspond à une réalisation d'une généalogie. 43

3.7 Deux exemples de généalogies comme des sous-graphes d'un même pedigree et les arbres qui leurs correspondent. (a) La généalogie est composée de deux arbres disjoints correspondant à deux fondateurs distincts. (b) Toutes les lignées coalescent sur le pedigree et la généalogie est composée d'un seul arbre. 44

3.8 Généalogie d'un échantillon de sept gènes composé des trois gènes $\{g_2, g_4, g_6\}$ et de l'ensemble $\{g_1, g_3, g_5, g_7\}$ de leurs parents sur le pedigree. 45

3.9 Les trois gènes de l'échantillon coalescent dans le même ancêtre. . 45

3.10 Un exemple du déroulement de l'algorithme 3.1. Les sommets pleins représentent l'ensemble E à chacun des passage dans la boucle TANT QUE. 48

3.11 Les sommets e_1 et e_2 sont des entonnoirs par rapport à l'échantillon $\{g_1, g_2\}$; ils peuvent jouer le rôle de fondateur car si g_1 et g_2 sont IBD, ils le sont déjà à partir de e_1 ou e_2. La partie du pedigree en bleu peut être élaguée sans modifier la loi de probabilité de l'état IBD $\mathbf{J}(\mathbf{S})$. 49

3.12 Les arbres minimaux contenant respectivement $\{g_1, g_2\}$ et $\{g_3, g_6\}$ sont en bleu. Les distances $d_{\mathcal{P}}(g_1, g_2)$ et $d_{\mathcal{P}}(g_3, g_6)$ valent respectivement 1 et 3. 56

4.1 Les gènes $g_1, g_2, \ldots, g_5$ proviennent d'un pedigree tandis que les fondateurs appartiennent à une population se reproduisant selon le modèle de Wright-Fisher. 59

4.2 Une généalogie sur un pedigree composé de deux sous graphes disjoints. 60

4.3 Un arbre (binaire) de coalescence complète la généalogie provenant du pedigree à partir de l'état IBD $\{\{g_1, g_2\}, \{g_3\}, \{g_4, g_5\}\}$. 61

4.4 Un arbre représentant la généalogie sous le p-coalescent. En bleu, les arbres minimaux contenant respectivement $\{g_1, g_2\}$ et $\{g_3, g_5\}$. 64

4.5 (1) La coalescence se produit sur le pedigree : $d(a, b) = 2$. (2) La coalescence excède le pedigree : $d(a, b) = 2.4N + 4$. 65

4.6 Une généalogie sous le p-coalescent à l'échelle de N évènements de reproduction. Les distances $d(g_1, g_2)$ et $d(g_3, g_5)$ valent respectivement zéro et N. 69

5.1 (a) Aucun évènement de mutation : les gènes sont dupliqués dans une forme identique. (b) Une mutation sur l'arête $[\mathbf{u}_1, \mathbf{u}_2]$ a pour conséquence d'observer $\mathbf{u}_2$ et son descendant $\mathbf{u}_3$ dans une forme distincte de $\mathbf{u}_1$. 73

5.2 Une mutation entre les générations $\mathcal{G}_6$ et $\mathcal{G}_5$ explique le polymorphisme des gènes de la génération G_0. 75

5.3 Exemple d'une mutation causée par un changement d'état de $\mathcal{M}_g =$ 2 sites parmi $\kappa = 7$. 78

5.4 Les mutations apparaissent en une reproduction selon un processus de Poisson sur l'intervalle $[0, 1]$. 84

5.5 Le nombre de sites polymorphes entre deux gènes séparés par $j = 4$ reproductions est la somme de 4 variables aléatoires de Poisson indépendantes. 85

5.6 Les évènements de mutation sur la généalogie expliquent le nombre $K = 3$ de sites polymorphes observés dans un échantillon de $n = 4$ gènes. 87

A.1 (a) Un pedigree composé de $|\mathcal{I}| = 20$ gènes : $|\mathcal{F}| = 8$ fondateurs (en bleu) et $|\mathcal{D}| = 12$ non-fondateurs. (b) Une réalisation possible (en rouge) d'une généalogie pour un échantillon de $n = 4$ gènes. L'état IBD sur cet exemple est la partition $\mathbf{J(S)} = \{\{g_1\}, \{g_2, g_3\}, \{g_4\}\}$; $|\mathbf{J(S)}| = 3$. 96

RÉSUMÉ

L'objectif central de ce mémoire est de présenter une approche qui permet d'inclure les contraintes relevant de la génétique familiale et des pedigrees dans un modèle de génétique des populations inspiré du processus de coalescence. Ce modèle est formulé en termes de graphes dirigés acycliques et, plus précisément, d'arbres appelés généalogies. Quelques notions fondamentales concernant le modèle de Wright-Fisher et la génétique familiale sont d'abord formulées. Le modèle proposé, le p-coalescent, est ensuite présenté avec quelques résultats concernant le comportement des distances sur le graphe et la relation qui existe entre la taille de la généalogie et le nombre de mutations observées.

MOTS-CLÉS : génétique, graphes, généalogie, processus de coalescence, pedigree.

INTRODUCTION

L'hérédité de certains caractères chez les espèces vivantes est un phénomène connu depuis longtemps. Par exemple, elle est nécessaire à la théorie de l'évolution proposée par Darwin (1883) même si ce dernier n'a pas détaillé le mécanisme par lequel les individus héritait des attributs de leurs ancêtres. Les expériences de fertilisation sur les plantes de Mendel (1865) ont suggéré que les caractères étaient passés aux progénitures par l'entremise d'un ensemble d'unités héréditaires aujourd'hui appelés gènes.

La génétique est devenue un domaine de recherche en mathématiques au début des années 1900, culminant notamment avec l'ouvrage de Fisher (1930). Aujourd'hui, la génétique mathématique est un domaine en plein essor, en particulier son aspect statistique avec l'analyse de données génomiques de plus en plus précises et volumineuses.

Ce mémoire est un essai de génétique mathématique abstraite concernant la modélisation de la transmission de matériel génétique. Sa lecture ne présuppose aucune connaissance préalable en génétique médicale ; des références à des termes spécifiques concernant l'aspect biologique ou moléculaire sont glissées pour le lecteur familier avec le domaine, mais peuvent être ignorées sans conséquences pour la compréhension générale. Principalement, ce sont les concepts de base en probabilités et en processus stochastiques qui sont nécessaires. Le livre de Ross (1996), par exemple, introduit tous les concepts et définitions de base utilisées dans ce mémoire.

Le but principal de ce mémoire est de proposer un premier modèle intégrant des

éléments de la génétique familiale dans une approche de génétique des populations. Les figures présentées font partie de la narration ; elles sont toutes faites sur mesure afin de faciliter la compréhension. Les dérivations mathématiques sont inspirées de la littérature pour les résultats connus, mais elles sont toutes originales et dans une forme aussi complète que possible.

C'est le concept de généalogie qui est au centre des préoccupations de chacun des chapitres. Il s'agit d'une représentation sous forme d'arbre donnant l'ensemble des relations ancestrales entre plusieurs gènes. Un «arbre» familial où sont détaillées les relations de parenté ne constitue pas une généalogie au sens entendu dans ce mémoire. Le terme employé dans la littérature de génétique mathématique pour l'ensemble des relations de parenté entre plusieurs individus est *pedigree*; le concept de pedigree sera aussi discuté de manière extensive dans ce mémoire.

En génétique mathématique, la généalogie décrit plutôt les relations d'hérédité au niveau du gène. Même si un individu possède deux parents, si l'attention est restreinte à un seul de ses gènes, celui-ci est hérité d'un seul des deux parents. Lequel des deux parents est l'unique ancêtre du gène est une question importante pour construire une généalogie; il s'agit cependant d'une variable inconnue. Cet exemple motive une interprétation probabiliste du concept de généalogie.

En dehors du cadre familial, il existe une autre approche de modélisation en génétique : la génétique des populations. Cette dernière modélise les généalogies d'un point de vue plus macroscopique. Elle permet, entre autres, d'expliquer les liens d'hérédité entre les membres d'une population à un niveau qui excède les relations de parenté connues en imaginant des généalogies très grandes qui permettent de relier des gènes à leur distant ancêtre commun.

Le défi du projet de recherche qui a conduit à ce mémoire est de proposer une nouvelle approche en génétique des populations capable de modéliser, pour un

ensemble de gènes, à la fois les relations de parenté fortes et les relations ancestrales plus distantes. Une des principales difficultés rencontrées est de travailler avec les deux échelles de temps différentes sur lesquelles sont construites les généalogies en génétique familiale et en génétique des populations.

Les chapitres 1 à 3 introduisent des concepts fondamentaux tirés de la littérature dans une notation unifiée à travers les différents domaines. Le premier chapitre exprime l'hérédité génétique en termes de graphes dirigés acycliques et quelques résultats de base y sont démontrés. Le second chapitre expose le modèle de Wright-Fisher en génétique des populations et l'approche à rebours qui en découle : le processus de coalescence. Au troisième chapitre, les bases de la génétique familiale sont discutées dans un langage permettant de faire le lien avec les deux premiers chapitres.

Le chapitre 4 constitue le cœur du mémoire. Il est entièrement original et présente le nouveau paradigme de modélisation proposé : le p-coalescent. Au dernier chapitre, le phénomène de mutation en génétique est intégré au modèle du chapitre 4 comme un premier exemple illustrant la nature particulière de cette nouvelle approche.

CHAPITRE I

GÉNÉTIQUE ET GRAPHES

L'objectif de ce chapitre est de formuler les bases biologiques de l'hérédité génétique en termes de graphes. Un ensemble minimal de définitions et résultats, tirés de la littérature, qui concernent les graphes dirigés sera nécessaire au développement d'une description générale du concept de généalogie. Une compréhension intuitive de ces concepts devrait être suffisante pour la lecture de ce mémoire.

1.1 Graphe dirigé acyclique

Les graphes dirigés peuvent être représentés par un ensemble de sommets reliés par des flèches (voir figure 1.1). Une ressource étoffée sur cette classe de graphes est le livre de Bang-Jensen et Gutin (2008).

Définition 1.1. *Un graphe dirigé* $\mathbf{D}$ *est une paire* $(\mathbf{V}, \mathbf{A})$ *:* $\mathbf{V}$ *est un ensemble fini de sommets et* $\mathbf{A}$ *est un ensemble de paires ordonnées de sommets représentant les arêtes.*

Un chemin sur un graphe dirigé doit être compris dans le sens intuitif, un parcours sur le graphe en suivant l'orientation des flèches. De plus, un chemin ne parcourt pas deux fois la même arête.

Définition 1.2. *Sur un graphe dirigé* $\mathbf{D} = (\mathbf{V}, \mathbf{A})$*, un chemin de* $\mathbf{v_0}$ *vers* $\mathbf{v_l}$ *est*

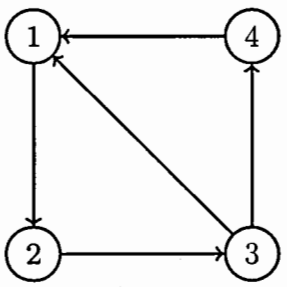


Figure 1.1 : Un graphe dirigé avec sommets $\mathbf{V} = \{1, 2, 3, 4\}$ et arêtes $\mathbf{A} = \{[1, 2], [2, 3], [3, 1], [3, 4], [4, 1)]\}$.

une séquence d'arêtes distinctes $\mathbf{a}_1 = [\mathbf{v}_0, \mathbf{v}_1], \mathbf{a}_2 = [\mathbf{v}_1, \mathbf{v}_2], \ldots, \mathbf{a}_l = [\mathbf{v}_{l-1}, \mathbf{v}_l]$. *Le nombre d'arêtes* l *est la longueur du chemin.*

Les liens entre la génétique et les graphes pourront être exprimés en restreignant l'attention à une classe des graphes dirigés : les graphes dirigés acycliques (DAG). Ce sont les graphes dirigés sur lesquels on ne retrouve pas de chemins retournant à leur point de départ (cycle).

Définition 1.3. *Un graphe dirigé acyclique est un graphe dirigé* $\mathbf{D} = (\mathbf{V}, \mathbf{A})$ *ne comportant pas de chemin d'un sommet* $\mathbf{v}$ *vers lui-même.*

Un DAG est représenté à la figure 1.2. En revanche, le graphe de la figure 1.1 n'est pas un DAG. L'orientation des flèches sera parfois omise pour les représentations de DAG avec la convention que celles-ci sont toujours dirigées vers le bas.

Une certaine terminologie sera utilisée pour définir des ensembles de sommets à partir d'un sommet $\mathbf{v} \in \mathbf{V}$. Ces concepts sont illustrés sur les figures 1.3 et 1.4.

- Les parents de $\mathbf{v}$ sont tous les sommets $\mathbf{u}$ tels qu'il existe une arête $[\mathbf{u}, \mathbf{v}]$.
- Les ancêtres de $\mathbf{v}$ sont tous les sommets $\mathbf{u}$ tels qu'il existe un chemin de $\mathbf{u}$ vers $\mathbf{v}$.

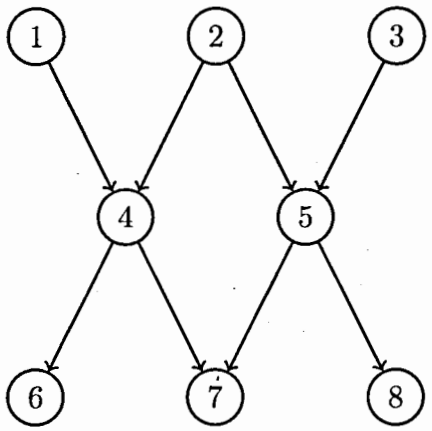


Figure 1.2 : Un graphe dirigé acyclique avec les sommets $\mathbf{V} = \{1, 2, \ldots, 8\}$ dans un ordre topologique (résultat 1.1).

- Les enfants de $\mathbf{v}$ sont tous les sommets $\mathbf{u}$ tels qu'il existe une arête $[\mathbf{v}, \mathbf{u}]$.
- Les descendants de $\mathbf{v}$ sont tous les sommets $\mathbf{u}$ tels qu'il existe un chemin de $\mathbf{v}$ vers $\mathbf{u}$.

Le vocabulaire est emprunté du lexique des relations familiales. Cela suggère que les liens de parenté peuvent être exprimés par un DAG ; cette conception sera discutée plus en détails au chapitre 3. Cependant, dans ce mémoire, ces termes feront toujours référence à des relations sur un graphe sauf lorsque indiqué autrement.

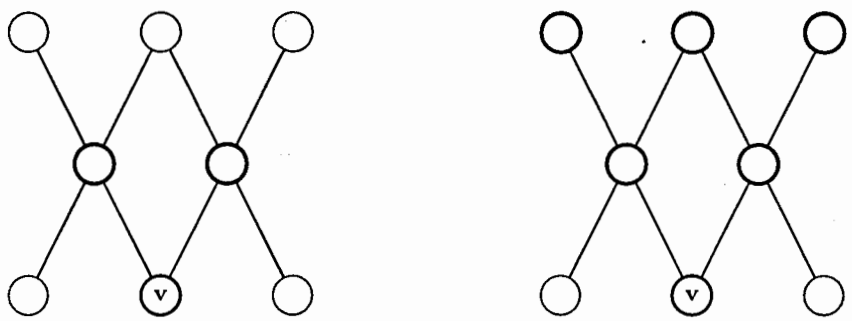


Figure 1.3 : En rouge les parents du sommet $\mathbf{v}$; en bleu, les ancêtres.

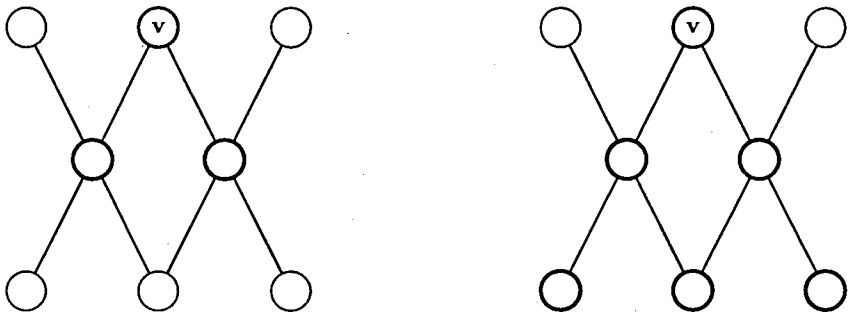


Figure 1.4 : En rouge les enfants du sommet $\mathbf{v}$; en bleu, les descendants.

Si $\pi(\mathbf{v})$ correspond à l'ensemble des parents d'un sommet $\mathbf{v} \in \mathbf{V}$, un DAG (et plus généralement un graphe dirigé) peut être défini comme la paire $(\mathbf{V}, \pi)$. Cette représentation est équivalente et sera parfois préférée. La fonction π prend ses valeurs dans l'ensemble $2^{\mathbf{V}}$ des parties de $\mathbf{V}$.

Le lemme suivant servira à la démonstration du résultat 1.1.

Lemme 1.1. *Pour un DAG* $\mathbf{D} = (\mathbf{V}, \pi)$, *il existe toujours un sommet* $\mathbf{v}$ *sans parents :* $\pi(\mathbf{v}) = \emptyset$.

Démonstration. Soit $|\mathbf{V}| = n$, le nombre de sommets. Supposons que $\pi(\mathbf{v}) \neq \emptyset$ pour tout $\mathbf{v} \in \mathbf{V}$. Autrement dit, pour un sommet $\mathbf{v}_1$, il existe un sommet $\mathbf{v}_2$ tel que $(\mathbf{v}_2, \mathbf{v}_1)$ est une arête du DAG. En itérant ainsi, il est possible de construire une séquence d'arêtes

$$[\mathbf{v}_{n+1}, \mathbf{v}_n], \ldots, [\mathbf{v}_3, \mathbf{v}_2], [\mathbf{v}_2, \mathbf{v}_1].$$

Comme il n'y a que n sommets, il existe au moins un sommet $\mathbf{v}$ qui se répète dans cette séquence. Il est donc possible d'en extraire un chemin de $\mathbf{v}$ vers $\mathbf{v}$ ce qui est une contradiction puisque $\mathbf{D}$ est un DAG. □

Il sera pratique dans plusieurs développements d'avoir l'ensemble des sommets $\mathbf{V}$

dans un ordre particulier, appelé ordre topologique, pour lequel tous les sommets sont plus grands que leurs parents. Les sommets sur le DAG de la figure 1.2 sont dans un ordre topologique. Le résultat suivant démontre l'existence d'un tel ordre.

Résultat 1.1. *Pour un DAG* $\mathbf{D} = (\mathbf{V}, \pi)$*, il est toujours possible de déterminer une relation d'ordre sur* $\mathbf{V}$ *telle que tout sommet* $\mathbf{v}$ *soit précédé par ses parents :* $\mathbf{u} < \mathbf{v}$ *pour tout* $\mathbf{v} \in \mathbf{V}, \mathbf{u} \in \pi(\mathbf{v})$*. Une telle relation est appelée ordre topologique et n'est pas nécessairement unique.*

Démonstration. Soit $|\mathbf{V}| = n$, le nombre de sommets. Un ordre topologique

$$\mathbf{v}_1 < \mathbf{v}_2 < \cdots < \mathbf{v}_n$$

peut être construit. Le point de départ est de choisir $\mathbf{v}_1 = \mathbf{u}$ avec $\mathbf{u}$ un sommet sans parents ; un tel sommet existe par le lemme 1.1. Ensuite, le sommet $\mathbf{v}_1$ est retiré du graphe avec toutes les arêtes qui le contiennent.

Le nouveau graphe est toujours un DAG, mais ne contient plus le sommet $\mathbf{v}_1$. Un sommet $\mathbf{u}$ sans parents peut ainsi être assigné à $\mathbf{v}_2$ et ce dernier retiré du graphe à son tour. La procédure est itérée jusqu'à ce que tout les sommets $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ aient été déterminés. Chacun des sommets est clairement retiré après ses parents alors $\mathbf{v}_1 < \mathbf{v}_2 < \cdots < \mathbf{v}_n$ est un ordre topologique. □

1.2 Arbres enracinés et génétique

Toutes les espèces vivantes possèdent un bagage génétique et celui-ci a la caractéristique de pouvoir se dupliquer. Le matériel génétique de chaque nouvel individu est obtenu d'un ou plusieurs autres individus vivants par duplication ; ce phénomène se nomme l'hérédité. Dans ce mémoire, le mot *gène* sera défini de la manière suivante.

Définition 1.4. *Un gène est une composante indivisible du bagage génétique dans le sens que celui-ci est hérité d'un seul individu.*

Un nucléotide, élément de base de la molécule d'ADN (Campbell *et al.*, 2009), est un exemple de gène dans l'esprit de la définition 1.4. Les concepts mathématiques qui seront élaborés se produisent au niveau du gène et décrivent les relations d'hérédité entre les individus. Le langage employé par Dawkins (1976) est particulièrement instructif; il parle du gène comme d'un duplicateur et de l'individu comme d'un véhicule.

Le concept de graphe dirigé acyclique permet de transcrire la notion d'hérédité en termes mathématiques : les sommets représentent les individus et les arêtes la relation héréditaire. Concrètement, un flèche $[\mathbf{u}, \mathbf{v}]$ signifie que le gène de l'individu $\mathbf{v}$ est hérité de l'individu $\mathbf{u}$.

Le nombre d'enfants d'un gène $\mathbf{u}$ est un nombre entier noté $v(\mathbf{u})$. Chacun de ses enfants se dupliquent également un nombre entier de fois. Le graphe engendré par ce processus (voir figure 1.5) est un DAG avec une structure particulière : un arbre enraciné en $\mathbf{u}$.

Définition 1.5. *Un arbre enraciné en* $\mathbf{u}$ *est un DAG* $\mathbf{D} = (\mathbf{V}, \alpha)$ *sur lequel il existe un seul chemin partant de la racine* $\mathbf{u}$ *vers chacun des sommets* $\mathbf{v} \in \mathbf{V}$.

Chacun des sommets $\mathbf{v}$ d'un arbre enraciné en $\mathbf{u}$ possède un seul parent, noté $\alpha(\mathbf{v})$, à l'exception de la racine $\mathbf{u}$ qui n'en possède aucun. Les sommets sans enfants sont les feuilles de l'arbre. Seuls les arbres enracinés sont discutés dans ce mémoire; le simple terme *arbre* sera parfois utilisé pour désigner cette classe de graphes.

De plus, une version un peu plus sommaire d'un arbre enraciné sera parfois illustrée pour représenter les liens d'hérédité. Elle sert à représenter les longues séquences

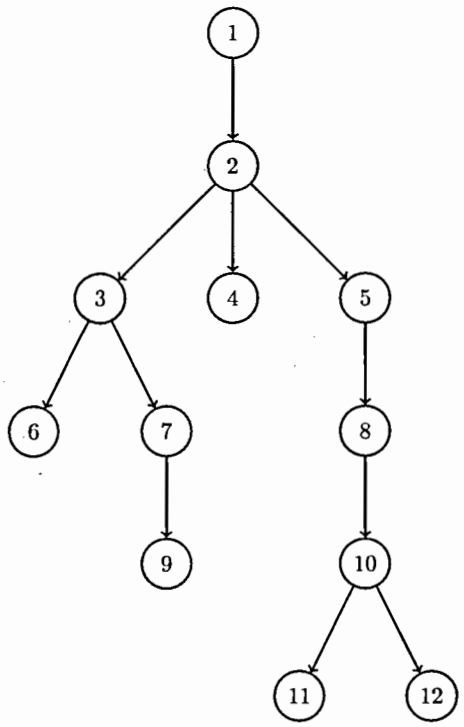


Figure 1.5 : Un arbre enraciné en 1. Les feuilles sont les sommets $4, 6, 9, 11, 12$.

de duplication de gènes de manière plus succincte. La différence est purement cosmétique.

À un arbre enraciné (par exemple celui sur la figure 1.5) correspond une arbre réduit pondéré (celui sur la figure 1.6). Il est obtenu à partir de l'arbre original en retirant les sommets n'ayant qu'un seul enfant (sauf la racine). Les chemins sont bonifiés d'une longueur afin de synthétiser les sommets perdus de la manière suivante : s'il existe un chemin

$$[\mathbf{u}_0, \mathbf{u}_1], [\mathbf{u}_1, \mathbf{u}_2], \ldots, [\mathbf{u}_{l-1}, \mathbf{u}_l]$$

de longueur l et que les sommets intermédiaires ont chacun exactement $v(\mathbf{u}_1) = v(\mathbf{u}_2) = \cdots = v(\mathbf{u}_{l-1}) = 1$ enfant, ses arêtes sont remplacées par une branche :

une arête pondérée $([\mathbf{u}_0, \mathbf{u}_l], l)$.

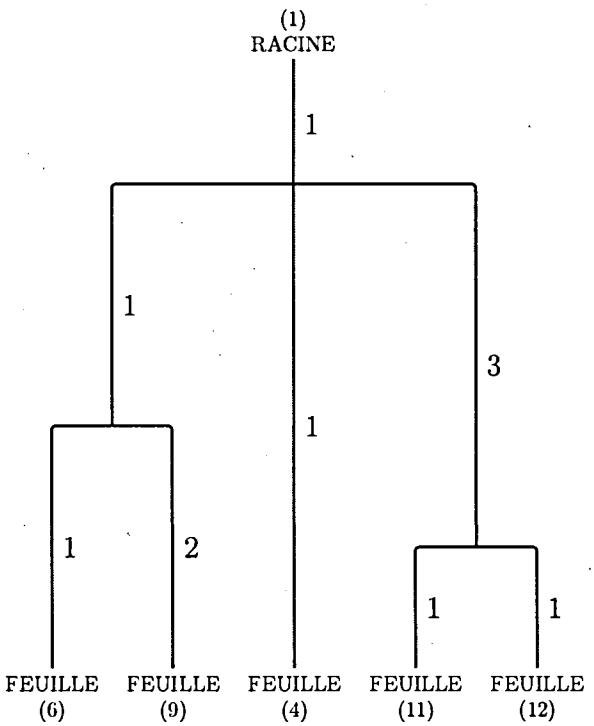


Figure 1.6 : La structure de l'arbre de la figure 1.5 est préservée et les branches sont pondérées par des longueurs.

Le concept d'arbre minimal permet de restreindre l'attention à un sous-ensemble d'individus (sommets). Il est illustré sur la figure 1.7.

Définition 1.6. *Soit* $\mathbf{D} = (\mathbf{V}, \alpha)$ *un arbre enraciné. L'arbre minimal contenant un sous-ensemble* $\mathbf{E} \subset \mathbf{V}$ *de sommets est l'arbre avec le nombre de sommets minimal parmi les sous-graphes de* $\mathbf{D}$ *qui contiennent* $\mathbf{E}$.

La taille d'un arbre sera une quantité d'intérêt. Les arbres seront mesurés par le nombre d'arêtes (ou la somme des longueurs de branches).

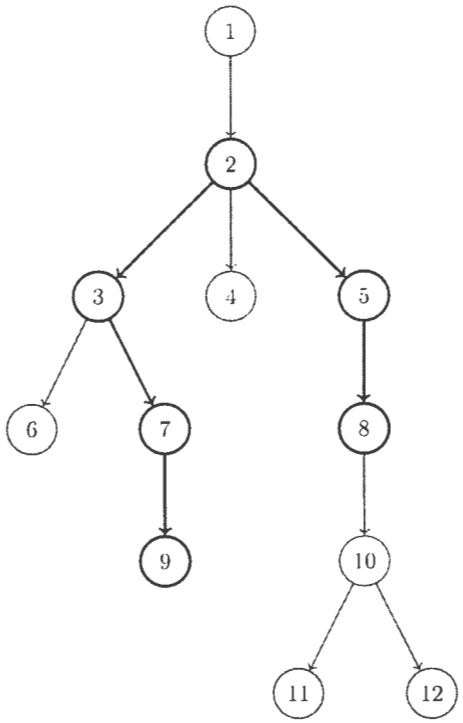


Figure 1.7 : L'arbre minimal contenant les sommets 8 et 9 est en bleu. La taille de l'arbre ou la distance $d(8,9)$ est 5.

Définition 1.7. *La taille d'un arbre* $\mathbf{D} = (\mathbf{V}, \mathbf{A})$ *est la cardinalité de l'ensemble* $\mathbf{A}$ *:* $|\mathbf{A}|$ *ou le nombre d'arêtes.*

La distance (illustrée sur la figure 1.7) entre deux sommets sur un arbre peut être définie sans équivoque à cause de la propriété d'unicité des chemins. Pour deux sommets $\mathbf{u}$ et $\mathbf{v}$, la distance se retrouve à partir de l'arbre minimal les contenant.

Définition 1.8. *La distance* $d(\mathbf{u}, \mathbf{v})$ *entre deux sommets sur un arbre enraciné est la taille de l'arbre minimal contenant* $\mathbf{u}$ *et* $\mathbf{v}$.

La fonction $d : \mathbf{V} \times \mathbf{V} \to \mathbb{N}$ est bien une distance au sens mathématique.

CHAPITRE II

MODÈLE DE WRIGHT-FISHER ET PROCESSUS DE COALESCENCE

Ce chapitre introduit le modèle de Wright-Fisher et le processus de coalescence, deux concepts en génétique des populations, avec quelques propriétés qui seront utiles dans les chapitres subséquents.

Le modèle de Wright-Fisher (Fisher, 1930 ; Wright, 1931) est un modèle idéalisé de reproduction de gènes pour des populations d'individus. D'une réalisation de ce modèle il est possible d'extraire une généalogie d'un échantillon d'individus : un arbre décrivant les relations d'hérédité entre ceux-ci.

Le processus de coalescence est une chaîne de Markov à temps continu qui construit des généalogies. Il décrit leur caractère aléatoire sous le modèle de Wright-Fisher à la limite lorsque le nombre d'individus dans la population est très grand.

2.1 Modèle de Wright-Fisher

Considérons une population d'individus du point de vue d'un gène (entité héréditaire indivisible, définition 1.4). Dans le modèle de Wright-Fisher, le processus de reproduction se produit à temps discret avec la convention que ce temps $\tau \in \mathbb{N} = \{0, 1, 2, \ldots\}$ est mesuré à rebours ($\tau = 0$ correspond au présent). La population au temps τ est appelée génération τ, notée $\mathcal{G}_\tau$, et toutes les générations

contiennent N individus. En d'autres mots, la grandeur de la population reste constante dans le temps ; les générations sont composées des individus 1 à N : $\mathcal{G}_\tau = \{1, 2, \ldots, N\}$.

La reproduction de la génération τ à $\tau - 1$ peut être représentée par un graphe dirigé (biparti) sur $\mathcal{G}_\tau$ et $\mathcal{G}_{\tau-1}$. Tous les sommets de $\mathcal{G}_{\tau-1}$ ont un seul parent ; chaque gène j de la génération $\tau - 1$ est hérité d'un seul gène $\alpha(j)$ de la génération précédente. Le graphe est donc composé des N arêtes

$$\{[\alpha(j), j] : 1 \leqslant j \leqslant N\}.$$

Dans le modèle de Wright-Fisher, le choix des parents est aléatoire ; il est supposé que ce choix se fait selon une loi uniforme discrète :

$$\alpha(j) \overset{\text{i.i.d.}}{\sim} \mathcal{U}\{1, 2, \ldots, N\}, \quad j = 1, 2, \ldots, N.$$

Plus simplement, tous les membres de la génération précédente τ peuvent être le parent d'un gène de la génération $\tau - 1$ avec équiprobabilité et indépendamment des parents des autres gènes de la génération $\tau - 1$. La figure 2.1 représente un graphe dirigé avec une telle structure.

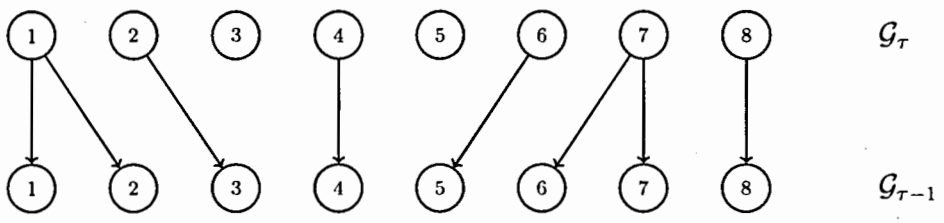


Figure 2.1 : Exemple du processus de reproduction entre la génération τ et $\tau - 1$ sous le modèle de Wright-Fisher pour une population de $N = 8$ individus.

Soit $v(j)$ le nombre d'enfants dans la génération $\mathcal{G}_{\tau-1}$ d'un gène j de la génération

$\mathcal{G}_\tau$. Le vecteur $(v(1), v(2), \ldots, v(N))$ suit une loi multinomiale

$$\mathcal{M}(N; (1/N, \ldots, 1/N)).$$

En effet, chacun des N parents disponibles est choisi avec probabilité $1/N$ lors de N essais indépendants.

Le processus de reproduction est indépendant et identiquement distribué pour toutes les générations successives $\mathcal{G}_\tau$ et $\mathcal{G}_{\tau-1}$ avec $\tau \geqslant 1$. Il est d'ailleurs échangeable sur les membres de $\mathcal{G}_\tau$ et $\mathcal{G}_{\tau-1}$ dans le sens que la probabilité d'une certaine configuration est constante sous une permutation des étiquettes 1 à N identifiant les individus. La figure 2.2 présente une réalisation du modèle de Wright-Fisher sur plusieurs générations.

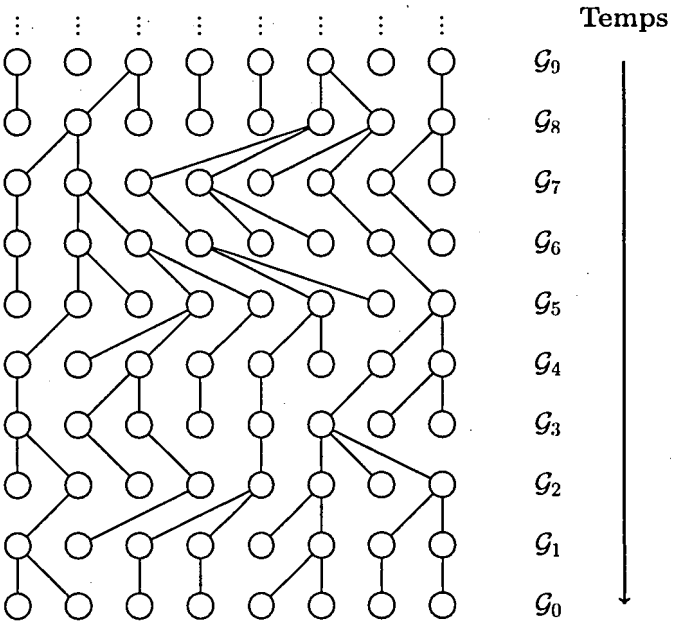


Figure 2.2 : Exemple d'une réalisation du modèle de Wright-Fisher sur 10 générations pour une population de $N = 8$ individus.

Pour la suite, l'attention sera restreinte à un échantillon composé d'un sous-

ensemble de n gènes de la génération $\mathcal{G}_0$ (présent).

2.1.1 Généalogie sous Wright-Fisher

Pour une réalisation du modèle de Wright-Fisher, la généalogie d'un échantillon de n gènes $\gamma_n \equiv \{g_1, g_2, \ldots, g_n\}$ de la génération $\mathcal{G}_0$ est, dans sa forme la plus fondamentale, l'arbre minimal contenant γ_n. Un exemple de généalogie se trouve à la figure 2.3.

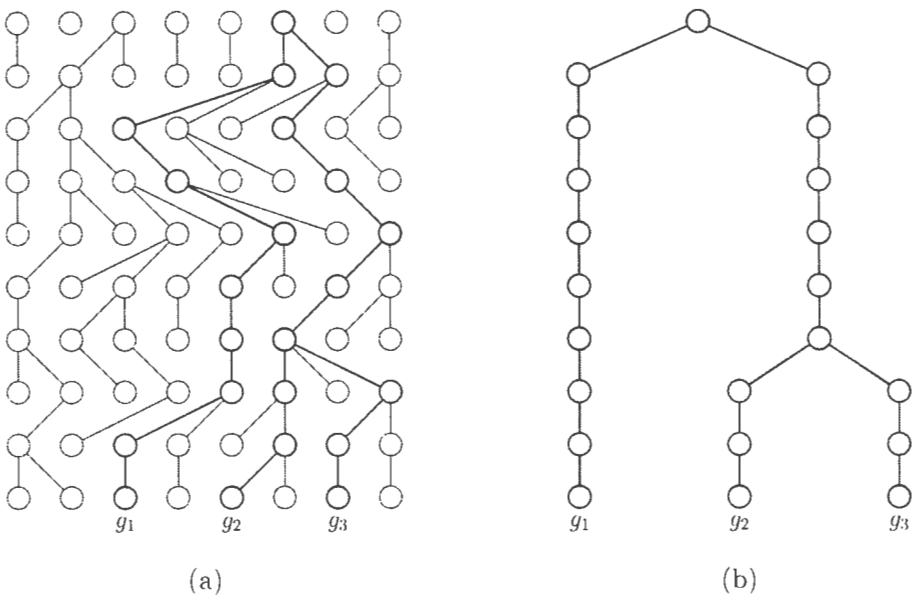


Figure 2.3 : (a) La généalogie de g_1, g_2 et g_3 est superposée sur une réalisation du modèle de Wright-Fisher. (b) La généalogie (redressée) est un arbre enraciné.

Les membres de l'échantillon γ_n se trouvent sur les feuilles de la généalogie. Les sommets avec deux enfants ou plus sur la généalogie correspondent à des évènements de coalescence : lorsque plusieurs gènes trouvent un ancêtre commun dans le passé ; ils ne forment désormais qu'une seule lignée. Les sommets avec un seul enfant ne seront généralement pas représentés ; ils sont éliminés au profit de lon-

gueurs de branches en nombre d'évènements de reproduction. L'aboutissement est une représentation d'arbre réduit pondéré comme celui de la figure 2.4a. Les longueurs de branches peuvent être résumées en notant simplement les temps entre les évènements de coalescence comme sur la figure 2.4b. Si l'on se restreint à des arbres binaires, comme pour le processus de coalescence décrit à la section 2.2, le nombre de ces événements est $n-1$.

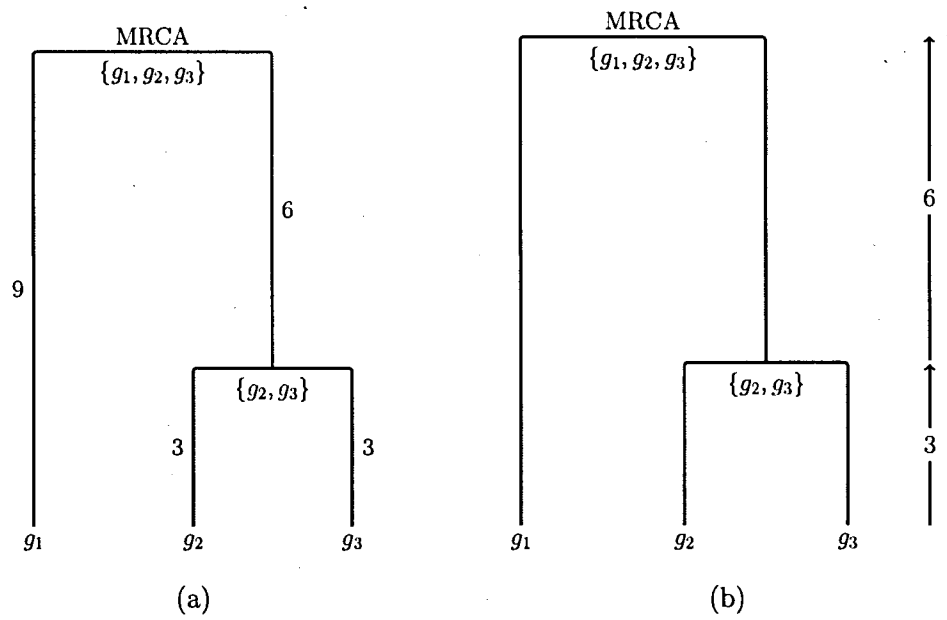


Figure 2.4 : (a) Une généalogie sous forme d'arbre réduit pondéré. (b) Les temps entre les évènements de coalescence résument les longueurs de branches.

2.1.2 Temps de coalescence et distance

Pour un échantillon de deux gènes $\gamma_2 = \{a, b\}$ de la génération $\mathcal{G}_0$, le temps de coalescence $\mathcal{T}_{ab}$ est défini comme le nombre de générations dans le passé avant que a et b trouvent un ancêtre commun (coalescent) sur la généalogie.

La probabilité que deux gènes a et b ne coalescent pas en une seule génération sous

le modèle de Wright-Fisher est la probabilité de choisir deux parents distincts à la génération précédente :

$$P(\alpha(a) \neq \alpha(b)) = \frac{N-1}{N} = 1 - \frac{1}{N}.$$

Comme cette probabilité est constante à chacune des générations, la probabilité que le temps de coalescence $\mathcal{T}_{ab}$ excède un temps $t \in [0, \infty)$ est

$$P(\mathcal{T}_{ab} > t) = P(\alpha^{\lfloor t \rfloor}(a) \neq \alpha^{\lfloor t \rfloor}(b)) = \left(1 - \frac{1}{N}\right)^{\lfloor t \rfloor}.$$

Le temps $\mathcal{T}_{ab}$ est donc distribué selon une loi géométrique de paramètre $1 - 1/N$.

La distance (définition 1.8) entre a et b sur la généalogie sera notée $\delta(a, b)$ pour le modèle de Wright-Fisher. Elle vaut deux fois le temps de coalescence et, donc, possède la fonction de survie

$$P(\delta(a,b) > t) = P(2\mathcal{T}_{ab} > t) = \left(1 - \frac{1}{N}\right)^{\lfloor \frac{t}{2} \rfloor}, \qquad t \geqslant 0. \tag{2.1}$$

La notion de distance entre deux gènes sur une généalogie permet de mesurer le degré de «dépendance» (héréditaire) ; la dépendance est plus forte entre deux gènes proches.

2.1.3 Généalogies exprimées en termes de partitions

À une généalogie sous le modèle de Wright-Fisher correspond une fonction r qui associe à chaque temps $t \in [0, \infty)$ une partition de γ_n. Plus précisément, si $\mathcal{C}_n$ est l'ensemble des partitions de γ_n, alors

$$r : [0, \infty) \to \mathcal{C}_n$$

avec la contrainte que pour $t' > t$, $r(t')$ est une partition plus grossière ou égale à $r(t)$.

Chacune des partitions correspond à une relation d'équivalence notée $\overset{r(t)}{=}$. Deux gènes a et b font partie de la même classe d'équivalence à partir du moment où ils ont coalescé sur la généalogie. La fonction r est donc construite de la façon suivante : pour chaque paire de gènes $\{a, b\} \subset \gamma_n$,

$$a \overset{r(t)}{=} b \quad \Leftrightarrow \quad t \geqslant \mathcal{T}_{ab}.$$

Par exemple, la fonction correspondant à la généalogie tracée à la figure 2.4 est

$$r(t) = \begin{cases} \{\{g_1\}, \{g_2\}, \{g_3\}\} & 0 \leqslant t < 3, \\ \{\{g_1\}, \{g_2, g_3\}\} & 3 \leqslant t < 9, \\ \{\{g_1, g_2, g_3\}\} & 9 \leqslant t. \end{cases}$$

C'est sous cette forme que sont décrites les généalogies par le processus de coalescence présenté à la section 2.2.

2.1.4 Processus ancestral

Un échantillon γ_n de la génération $\mathcal{G}_0$ d'une population de Wright-Fisher est composé de n individus distincts représentant n lignées. En remontant dans le passé, des évènements de coalescence regroupent les gènes et le nombre de lignées diminue jusqu'à valoir 1 lorsque le plus récent ancêtre commun est atteint.

Le processus ancestral (Tavaré, 2004, section 2.2) $\{A_n(\tau) : \tau \in \mathbb{N}\}$ est une chaîne de Markov donnant le nombre de lignées à la génération τ pour un échantillon de taille n : $A_n(0) = n$. Il prend ses valeurs dans $\{1, 2, \ldots, n\}$, il est décroissant et est éventuellement absorbé à l'état 1. Soit $\mathbf{M}$ sa matrice de transition, $\mathbf{m}_{kk'}$ représente la probabilité de passer de k à $k' \leqslant k$ lignées en une seule génération.

Supposons que le processus ancestral est à l'état k au temps τ. Noter que les parents des k gènes représentant les k lignées sont choisis uniformément et in-

dépendamment parmi les N membres de la population à la génération $\tau + 1$, la génération précédente. Ainsi, pour une assignation particulière des parents des k lignées : $1 \leqslant i_1, \ldots, i_k \leqslant N$,

$$P\left(\bigcap_{j=1}^{k} \{\alpha(j) = i_j\}\right) = \frac{1}{N^k}.$$

Pour calculer $\mathbf{m}_{kk'}$, il faut compter au numérateur le nombre de manières de choisir $i_1, \ldots, i_k$ tels qu'il existe exactement k' valeur distinctes :

$$\left|\bigcup_{j=1}^{k} i_j\right| = k'.$$

Il s'agit d'une quantité combinatoire qui peut être raisonnée de la façon suivante :

1. On choisit k' parents parmi les N membres de la population : $\binom{N}{k'}$.
2. Le nombre de façon d'agencer les k lignées parmi les k' parents de façon à ce que chacun ait au moins un enfant correspond au nombre de fonctions surjectives de $\{1, \ldots, k\}$ vers $\{1, \ldots, k'\}$: $S(k, k')k'!$.

Le coefficient $S(k, k')$ est un nombre de Stirling du second type qui représente le nombre de manières distinctes de partitioner un ensemble de cardinalité k en k' classes. Un exemple est donné à la figure 2.5.

Alors, la probabilité que k gènes aient k' parents distincts vaut

$$\mathbf{m}_{kk'} = \frac{S(k, k')k'!\binom{N}{k'}}{N^k} = \frac{S(k, k')(N)_{k'}}{N^k}, \qquad 1 \leqslant k' \leqslant k \leqslant n,$$

avec

$$(N)_{k'} = \prod_{j=0}^{k'-1} (N - j) = N^{k'} \prod_{j=0}^{k'-1} \left(1 - \frac{j}{N}\right),$$

le factoriel descendant.

Cette probabilité est dérivée de manière similaire dans Tavaré (2004, section 2.2). Elle peut également être raisonnée en tant que distribution d'occupation. En effet, elle équivaut à la probabilité de laisser exactement $N - k'$ boîtes vides en

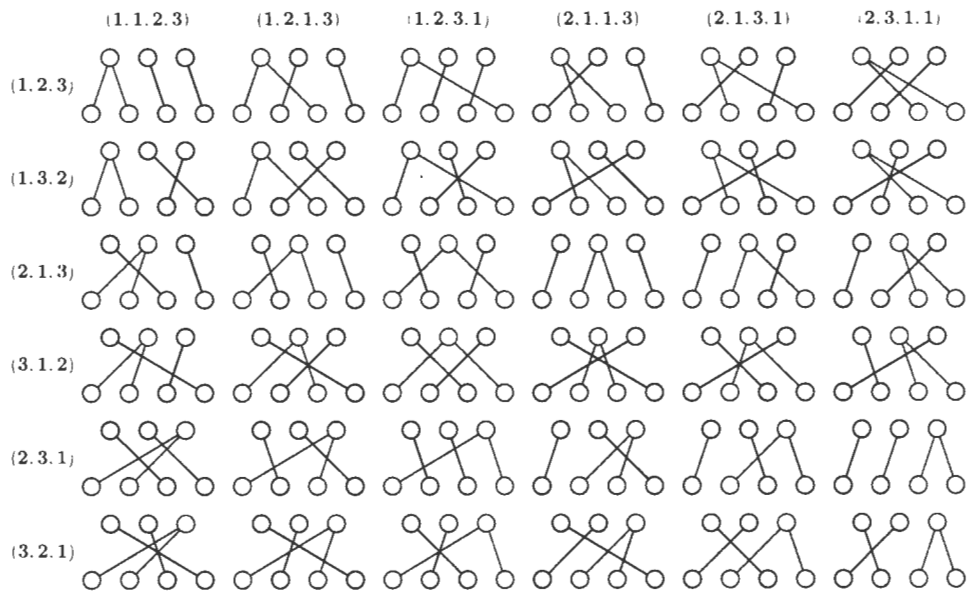


Figure 2.5 : Il existe $S(4,3)3! = \binom{4}{2}3! = 36$ fonctions surjectives de $\{1,2,3,4\}$ vers $\{1,2,3\}$. Chacune implique une seule coalescence d'exactement deux lignées.

plaçant k objets avec probabilité uniforme parmi N boîtes (Johnson *et al.*, 2005, section 10.4.1).

En particulier, pour n fixé et $k \leqslant n$, l'ordre asymptotique de la probabilité $\mathbf{m}_{kk'}$ en fonction de N est décrit par l'équation suivante :

$$\mathbf{m}_{kk'} = \frac{S(k,k')}{N^{k-k'}} \prod_{j=0}^{k'-1} \left(1 - \frac{j}{N}\right) = \mathcal{O}\left(\frac{1}{N^{k-k'}}\right). \tag{2.2}$$

Pour $k' = k$, noter que $S(k,k) = 1$ et donc la probabilité qu'il n'y ait aucune coalescence pour k lignées en une génération s'exprime

$$\begin{aligned}\mathbf{m}_{kk} = \prod_{j=0}^{k-1} \left(1 - \frac{j}{N}\right) &= 1 - \frac{\sum_{j=1}^{k-1} j}{N} + \mathcal{O}\left(\frac{1}{N^2}\right) \\ &= 1 - \frac{\binom{k}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right).\end{aligned}$$

Lorsque $k' = k-1$, le nombre de Stirling $S(k, k-1) = \binom{k}{2}$. En effet, choisir une partition de taille $k-1$ d'un ensemble de taille k correspond à mettre deux de ses éléments dans la même classe en laissant les autres distincts. Alors, la probabilité de passer de k lignées à $k-1$ s'exprime

$$\mathbf{m}_{k,k-1} = \frac{S(k,k-1)}{N} \prod_{j=0}^{k-2} \left(1 - \frac{j}{N}\right) = \frac{S(k,k-1)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$
$$= \frac{\binom{k}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right).$$

Finalement, pour $k' \leqslant k-2$, $\mathbf{m}_{kk'} = \mathcal{O}(1/N^2)$ par l'équation (2.2). Les ordres asymptotiques pour les probabilités de transition $\mathbf{m}_{kk'}$ seront très importants pour dériver le processus de coalescence comme modèle limite pour les généalogies dans une population de Wright-Fisher. En particulier, il sera démontré que la probabilité de voir plus que deux lignées coalescer en même temps est négligeable lorsque le nombre d'individus dans la population est très grand.

2.1.5 Propriétés des généalogies sous Wright-Fisher

Certaines propriétés des généalogies qui surviennent dans le cadre du modèle de Wright-Fisher peuvent être dérivées à partir du processus ancestral. En particulier, deux concepts sont développés : la hauteur et la taille d'une généalogie. Cette dernière sera utile au chapitre 5 pour analyser le degré de polymorphisme d'un échantillon de gènes.

Premièrement, la hauteur de la généalogie $\mathcal{T}^{(n)}_{\text{MRCA}}$ pour un échantillon de taille n est le numéro de la génération dans laquelle se trouve l'ancêtre commun le plus récent (MRCA) de l'échantillon :

$$\mathcal{T}^{(n)}_{\text{MRCA}} = \min\{\tau \in \mathbb{N} : A_n(\tau) = 1\}.$$

Le temps $T^{(n)}_{\text{MRCA}}$ est un temps d'absorption puisque la chaîne est constante à

partir du moment ou l'état 1 est atteint ; le nombre de lignées reste 1 pour n'importe quel temps $\tau \geqslant T_{\text{MRCA}}$. Une analyse des premiers pas permet d'exprimer (entre autres) l'espérance d'un temps d'absorption comme la solution d'un système d'équations linéaires. La procédure générale est détaillée dans le livre de Gallager (2011, section 3.5.1).

En procédant à une analyse des premiers pas sur $A_n(\tau)$, l'espérance de la hauteur de la généalogie peut être exprimée comme

$$E[\mathcal{T}^{(n)}_{\text{MRCA}}] = 1 + \sum_{k=1}^{n} E[\mathcal{T}^{(k)}_{\text{MRCA}}]\mathbf{m}_{nk}.$$

En substituant les entrées de $\mathbf{M}$, la matrice de transition du processus ancestral (section 2.1.4), la relation suivante est obtenue :

$$\frac{\binom{n}{2}}{N} E[\mathcal{T}^{(n)}_{\text{MRCA}}] = 1 + \frac{\binom{n}{2}}{N} E[\mathcal{T}^{(n-1)}_{\text{MRCA}}] + \mathcal{O}\left(\frac{1}{N^2}\right) \sum_{k=1}^{n} E[\mathcal{T}^{(k)}_{\text{MRCA}}]. \tag{2.3}$$

En multipliant par $N/\binom{n}{2}$, la relation (2.3) peut s'exprimer dans la forme récursive suivante :

$$\left[1 + \mathcal{O}\left(\frac{1}{N}\right)\right] E[\mathcal{T}^{(n)}_{\text{MRCA}}] = \frac{N}{\binom{n}{2}} + E[\mathcal{T}^{(n-1)}_{\text{MRCA}}] + \mathcal{O}\left(\frac{1}{N}\right) \sum_{k=1}^{n-1} E[\mathcal{T}^{(k)}_{\text{MRCA}}]. \tag{2.4}$$

Intuitivement, plus le nombre N d'individus dans la population est grand, plus il faut attendre avant d'atteindre le MRCA. Pour un échantillon de taille $n = 1$, le MRCA est atteint à la génération zéro :

$$E[\mathcal{T}^{(1)}_{\text{MRCA}}] = 0 = \mathcal{O}(N).$$

Ainsi, par la récurrence (2.4),

$$E[\mathcal{T}^{(k)}_{\text{MRCA}}] = \mathcal{O}(N)$$

pour tout $k \in \{1, 2, \ldots, n\}$.

Alors, l'espérance du temps au plus récent ancêtre commun sous le modèle de Wright-Fisher s'exprime ainsi :

$$\begin{aligned} E[\mathcal{T}_{\mathrm{MRCA}}^{(n)}] &= N \sum_{k=2}^{n} \frac{1}{\binom{k}{2}} + \mathcal{O}(1) \\ &= N \sum_{k=2}^{n} \frac{2}{k(k-1)} + \mathcal{O}(1) \\ &= 2N \sum_{k=2}^{n} \left[\frac{1}{k-1} - \frac{1}{k} \right] + \mathcal{O}(1) \\ &= 2N \left(1 - \frac{1}{n} \right) + \mathcal{O}(1). \end{aligned}$$

D'autre part, la taille de la généalogie ou le temps total pour un échantillon de taille n se définit comme le nombre d'arêtes sur l'arbre. Lorsque le nombre de lignées $A_n(\tau) = k$, chacun des ancêtres des k lignées choisit un parent dans la génération précédente $(\tau + 1)$; un total de k arêtes est ajouté au graphe. La généalogie est complète une fois que le plus récent ancêtre commun est atteint. Le temps total s'exprime alors en fonction de $\mathcal{T}_{\mathrm{MRCA}}^{(n)}$ comme

$$\mathcal{T}_{\mathrm{TOT}}^{(n)} = \sum_{\tau=0}^{\mathcal{T}_{\mathrm{MRCA}}^{(n)}-1} A_n(\tau).$$

Une analyse des premiers pas pour l'espérance du temps total mène à

$$E[\mathcal{T}_{\mathrm{TOT}}^{(n)}] = n + \sum_{k=1}^{n} E[\mathcal{T}_{\mathrm{TOT}}^{(k)}] \mathbf{m}_{nk}.$$

Une démarche identique à celle pour le MRCA indique que le temps total est d'ordre asymptotique $\mathcal{O}(N)$ comme fonction de N. De plus, le temps total s'ex-

prime comme

$$\begin{aligned} E[\mathcal{T}_{\mathrm{TOT}}^{(n)}] &= N \sum_{k=2}^{n} \frac{k}{\binom{k}{2}} + \mathcal{O}(1) \\ &= N \sum_{k=2}^{n} \frac{2}{k-1} + \mathcal{O}(1) \\ &= 2N \sum_{k=1}^{n-1} \frac{1}{k} + \mathcal{O}(1). \end{aligned} \tag{2.5}$$

Les résultats présentés dans cette section sont cohérents avec les propriétés limites connues du modèle de Wright-Fisher telles que données par le processus de coalescence. Les propriétés de la hauteur et de la taille de la généalogie sous le processus de coalescence sont discutées à la section 2.2.1. Il est à noter qu'ils sont obtenus ici à partir d'une analyse des premiers pas sur le processus ancestral, ce qui constitue une démarche originale ne s'appuyant pas sur le processus de coalescence.

2.2 Processus de coalescence

Le processus de coalescence (Kingman, 1982a) est un processus stochastique qui construit des généalogies du présent vers le passé. Il peut être dérivé d'une grande classe de modèles de population à reproduction échangeable (Kingman, 1982b) incluant le modèle de Wright-Fisher présenté à la section 2.1.

Définition 2.1. *Le processus de coalescence* $\{R_t, t \geqslant 0\}$ *est une chaîne de Markov à temps continu sur l'ensemble* $\mathcal{C}_n$ *des partitions de* γ_n *avec générateur* $\mathbf{Q}$ *:*

$$\mathbf{q}_{\xi\eta} = \begin{cases} -\binom{|\xi|}{2} & \text{si } \xi = \eta, \\ 1 & \text{si } \xi \prec \eta, \\ 0 & \text{sinon,} \end{cases}$$

où $\xi \prec \eta$ *lorsque* η *est une partition obtenue en joignant deux classes de* ξ*.*

Le processus de coalescence génère des généalogies sous forme d'arbres binaires. La chaîne démarre à l'état $\Delta = \{\{g_1\}, \{g_2\}, \ldots, \{g_n\}\}$ et tous les états sont transitoires sauf $\Theta = \{\{g_1, \ldots, g_n\}\}$ où la chaîne est éventuellement absorbée en exactement $n-1$ transitions. La relation d'équivalence donnée par une partition est appelée *identité par descendance* (IBD) (Thompson, 2013).

Le nombre de classes d'une partition ξ est noté $|\xi|$ et correspond au nombre de lignées distinctes. La chaîne fait toujours des transitions en regroupant deux classes d'équivalences. La partition obtenue de ξ en joignant les classes i et j avec $i < j$ est notée $\xi^{(i,j)}$. Le nombre de classes diminue de un à chaque transition : pour $\xi \prec \eta$, $|\eta| = |\xi| - 1$. Le nombre de lignées $|R_t|$ forme par ailleurs un processus de mort pur sur $\{1, 2, \ldots, n\}$ avec taux $d_k = \binom{k}{2}$ pour $2 \leqslant k \leqslant n$.

2.2.1 Propriétés du processus de coalescence

Certaines propriétés du processus de coalescence comme l'espérance de la hauteur et de la taille de la généalogie sont présentées ici et seront utiles à la présentation de résultats dans les chapitres subséquents.

Le temps T_k avant une coalescence lorsqu'il y a $|R_t| = k$ lignées est distribué selon une loi exponentielle de paramètre $\binom{k}{2}$. Ainsi,

$$E[T_k] = \frac{1}{\binom{k}{2}} = \frac{2}{k(k-1)}, \qquad k = 2, 3, \ldots, n.$$

La hauteur de la généalogie, ou le temps jusqu'au plus récent ancêtre commun, T_{MRCA} s'exprime comme le temps d'attente avant que le nombre de lignées ne soit plus que un :

$$T_{\text{MRCA}} = \min\{t \geqslant 0 : |R_t| = 1\}$$

Pour le processus de coalescence, le nombre de lignées $|R_t|$ est un processus de

mort pur ; donc, T_{MRCA} est la somme des $n-1$ variables aléatoires exponentielles

$$T_{\text{MRCA}} = \sum_{k=2}^{n} T_k,$$

de sorte que

$$E[T_{\text{MRCA}}] = \sum_{k=2}^{n} E[T_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n} \left[\frac{1}{k-1} - \frac{1}{k}\right] = 2\left(1 - \frac{1}{n}\right).$$

En moyenne, le temps de coalescence lorsqu'il ne reste que deux lignées représente plus de la moitié du temps au plus récent ancêtre commun :

$$1 = E[T_2] > \frac{E[T_{\text{MRCA}}]}{2} = 1 - \frac{1}{n}.$$

La taille de la généalogie ou le temps total pour le processus de coalescence se développe à partir des temps de coalescence ; puisqu'il y a k branches sur la généalogie durant un temps T_k, le temps total s'exprime

$$T_{\text{TOT}} = \sum_{k=2}^{n} kT_k.$$

Ainsi,

$$E[T_{\text{TOT}}] = \sum_{k=2}^{n} kE[T_k] = \sum_{k=2}^{n} k\frac{2}{k(k-1)} = 2\sum_{k=1}^{n-1} \frac{1}{k}. \tag{2.6}$$

De plus, toutes les suites de partitions $\eta_0, \eta_1, \ldots, \eta_{n-1}$ telles que

$$\Delta = \eta_0 \prec \eta_1 \prec \cdots \prec \eta_{n-1} = \Theta$$

sont équiprobables sous le processus de coalescence ; celles-ci donnent la forme de l'arbre. La figure 2.6 représente une généalogie typique.

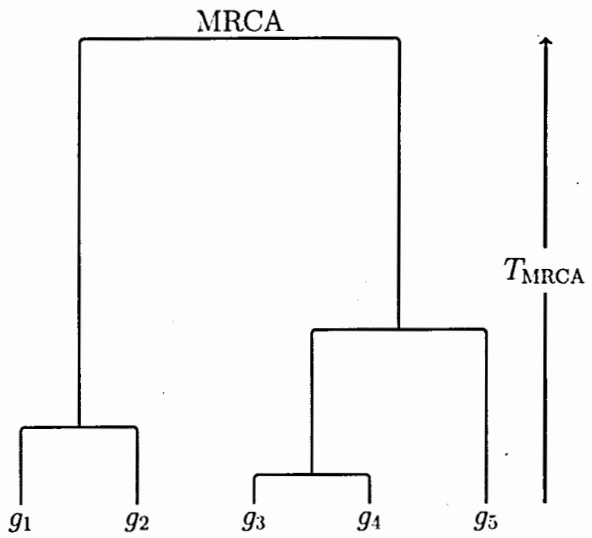


Figure 2.6 : Une réalisation moyenne du processus de coalescence avec $T_k = E[T_k]$ pour $k = 2, \ldots, 5$ $(n = 5)$.

2.2.2 Simulation de généalogies

Des généalogies qui sont des réalisations du processus de coalescence se simulent aisément. La procédure de simulation est dérivée d'une construction des chaînes de Markov à temps continu (Durrett, 2012, section 4.1). Ce type de procédure se base sur le concept de chaîne intégrée. L'idée générale est présentée ici sans développements mathématiques.

La suite d'états $i_0, i_1, i_2, \ldots$ visitée par une chaîne de Markov à temps continu est une chaîne de Markov à temps discret appelée chaîne intégrée. Si $\mathbf{Q}$ est le générateur de la chaîne à temps continu, les transitions, pour la chaîne intégrée, se font d'un état i vers un état j avec probabilité

$$-\frac{\mathbf{q}_{ij}}{\mathbf{q}_{ii}}. \tag{2.7}$$

Le temps passé à l'état i est la variable exponentielle

$$T_i \sim \mathcal{E}\mathrm{xp}(-\mathbf{q}_{ii})$$

et est indépendant de l'état j vers lequel la chaîne intégrée fait une transition.

Cela suggère un méthode itérative pour générer les valeurs d'une chaîne de Markov à temps continu. Si la chaîne démarre à l'état i, le temps T_i peut être simulé indépendamment de la nouvelle valeur j pour la chaîne intégrée qui, elle, est tirée en accord avec (2.7). La procédure est ensuite répétée un nombre arbitraire de fois à partir du nouvel état j.

Dans le cas du processus de coalescence R_t, il a la particularité d'être absorbé à l'état

$$\Theta = \{\{g_1, \ldots, g_n\}\} \in \mathcal{C}_n$$

après exactement $n-1$ transitions. Autrement dit, $\mathbf{q}_{\Theta\Theta} = 0$ ce qui doit être interprété comme $T_\Theta = \infty$; après $n-1$ transitions, la chaîne est coincée à Θ durant un temps infini.

Les transitions se font toujours d'une partition ξ vers une partition η avec $\xi \prec \eta$. La probabilité de transition vers un état $\eta = \xi^{(i,j)}$ obtenu en joignant les classes d'équivalence i et j de ξ est uniforme sur toutes les paires $i, j = 1, 2, \ldots, |\xi|$ telles que $i < j$.

Le taux de transition, lorsque R_t se trouve à l'état ξ, vaut

$$-\mathbf{q}_{\xi\xi} = \binom{|\xi|}{2}$$

ce qui correspond au paramètre de la loi exponentielle du temps de coalescence $T_{|\xi|}$. Il y a $n-1$ de ces temps correspondant aux transitions entre les états

$$\Delta = \eta_0 \prec \eta_1 \prec \cdots \prec \eta_{n-1} = \Theta.$$

Au départ, la chaîne R_t est à l'état $\Delta = \{\{g_1\}, \{g_2\}, \ldots, \{g_n\}\}$ et avec $|\Delta| = n$. Le nombre de classes diminue de un à chacune des transitions. Le temps passé à l'état η_τ, l'état de la chaîne après τ sauts, est donc une variable aléatoire exponentielle de taux

$$-\mathbf{q}_{\eta_\tau \eta_\tau} = \binom{n-\tau}{2}, \qquad \tau = 0, 1, \ldots, n-2.$$

Pour le processus de coalescence, le taux de transition $-\mathbf{q}_{\eta_\tau \eta_\tau}$ pour le $(\tau+1)^{\text{ième}}$ saut ne dépend pas de la partition η_τ visitée. Les temps de coalescence

$$T_n, T_{n-1}, \ldots, T_2$$

correspondent à ces variables exponentielles avec paramètres respectifs

$$\binom{n}{2}, \binom{n-1}{2}, \ldots, \binom{2}{2}$$

et sont indépendants de la chaîne intégrée du processus de coalescence R_t.

L'algorithme 2.1 procède en générant les $n-1$ temps de coalescence et ensuite la séquence des partitions visitées. Au dernier passage dans la boucle, $i = 1$, $j = 2$ et les deux dernières classes sont regroupées pour former l'état $\Theta = \{\{g_1, \ldots, g_n\}\}$.

2.2.3 Dérivation du processus de coalescence à partir du modèle de Wright-Fisher

Le processus de coalescence a été découvert comme processus limite du modèle de Wright-Fisher par Kingman (1982b). Le développement présenté dans cette section est entièrement calqué sur la section 2 de cet article. Il est cependant exposé ici dans une forme plus détaillée.

Soit $\mathcal{R}_\tau$ le processus de généalogie donnant la partition à la génération $\tau \in \mathbb{N}$ mesuré dans le passé d'un échantillon γ_n de n gènes choisis à la génération $\mathcal{G}_0$ sous le modèle de Wright-Fisher. Ce processus prend ses valeurs dans l'ensemble des partitions de γ_n, $\mathcal{C}_n$, et est markovien en τ ; sa matrice de transition est notée $\mathbf{P}_N$.

$\eta_0 = \Delta$ (Partition initiale)

$n = |\eta_0|$ (Taille de l'échantillon)

POUR $k = n, n-1, \ldots, 2$

  $t_k \sim \mathcal{E}\mathrm{xp}(1)$

  $T_k = t_k / \binom{k}{2}$

FIN **POUR**

$\xi = \eta_0$ (Partition courante)

$k = n$ (Nombre de lignées)

POUR $\tau = 1, \ldots, n-1$

  $(i,j) \sim \mathcal{U}\{i,j : 1 \leqslant i < j \leqslant k\}$ (Tirage sans remise)

  $\eta_\tau = \xi^{(i,j)}$

  $\xi = \eta_\tau$ (Actualisation de la partition)

  $k = k - 1$ (Actualisation du nombre de lignées)

FIN **POUR**

Algorithme 2.1 : Simulation d'une généalogie pour un échantillon γ_n selon le processus de coalescence.

Le processus ancestral, introduit à la section 2.1.4, $A_n(\tau) = |\mathcal{R}_\tau|$ donne le nombre total de lignées à la génération τ. La matrice de transition de $\mathcal{R}_\tau$, le processus sur les partitions, peut être dérivée à partir du processus ancestral et sa matrice de transition $\mathbf{M}$.

La probabilité qu'il n'y ait aucune coalescence pour k lignées en une génération s'exprime

$$\mathbf{m}_{kk} = 1 - \frac{\binom{k}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right).$$

Alors, pour $\mathbf{P}_N$, la matrice de transition du processus $\mathcal{R}_\tau$ sur les partitions,

$$\mathrm{p}_{\xi\xi} = \mathbf{m}_{|\xi||\xi|} = 1 - \frac{\binom{|\xi|}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right), \tag{2.8}$$

pour tout $\xi \in \mathcal{C}_n$.

Le nombre de lignées diminue de un en passant de k à $k-1$ avec probabilité

$$\mathbf{m}_{k,k-1} = \frac{\binom{k}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right).$$

Cela correspond pour $\mathcal{R}_\tau$ à une transition d'une partition ξ vers une partition η telle que $\xi \prec \eta$. Pour une partition ξ donnée, il existe $\binom{|\xi|}{2}$ de ces partitions vers lesquelles les transitions sont équiprobables, une conséquence de l'échangeabilité du modèle de Wright-Fisher. Donc, pour $\xi \prec \eta$,

$$\mathrm{p}_{\xi\eta} = \frac{\mathbf{m}_{|\xi|,|\xi|-1}}{\binom{|\xi|}{2}} = \frac{1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{2.9}$$

Les transitions vers d'autres partitions que celles décrites par (2.8) et (2.9) se font avec probabilité $\mathcal{O}(1/N^2)$ puisque pour $k' \leqslant k-2$, $\mathbf{m}_{kk'} = \mathcal{O}(1/N^2)$; la probabilité de voir plus de deux lignées coalescer en une seule génération est de l'ordre de $1/N^2$.

La matrice de transition $\mathbf{P}_N$ s'exprime donc

$$\mathrm{p}_{\xi\eta} = \begin{cases} 1 - \frac{\binom{|\xi|}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right) & \text{si } \xi = \eta, \\ \frac{1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right) & \text{si } \xi \prec \eta, \\ \mathcal{O}\left(\frac{1}{N^2}\right) & \text{sinon.} \end{cases}$$

De plus, la matrice peut s'écrire sous la forme

$$\mathbf{P}_N = \mathbf{I} + \frac{1}{N}\mathbf{Q} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{2.10}$$

avec $\mathbf{Q}$ le générateur du processus de coalescence (définition 2.1).

Le Lemme suivant sur les normes de matrices servira à la démonstration du résultat 2.1.

Lemme 2.1. *Soit* $\mathbf{P}$ *et* $\mathbf{Q}$ *deux matrices stochastiques* $n \times n$, $\mathbf{A}$ *une matrice* $n \times n$ *quelconque et la norme*

$$||\mathbf{A}|| = \max_{i=1}^{n} \sum_{j=1}^{n} |\mathbf{a}_{ij}|.$$

Alors,

$$||\mathbf{P}^N - \mathbf{Q}^N|| \leqslant N||\mathbf{P} - \mathbf{Q}||. \tag{2.11}$$

Démonstration. Premièrement, pour $\mathbf{A}$ une matrice $n \times n$ quelconque, en appliquant l'inégalité du triangle,

$$\begin{aligned} ||\mathbf{PA}|| &= \max_{i=1}^{n} \sum_{j=1}^{n} \left| \sum_{k=1}^{n} \mathbf{p}_{ik}\mathbf{a}_{kj} \right| \\ &\leqslant \max_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} |\mathbf{p}_{ik}\mathbf{a}_{kj}| \\ &= \max_{i=1}^{n} \sum_{k=1}^{n} \mathbf{p}_{ik} \sum_{j=1}^{n} |\mathbf{a}_{kj}| \\ &\leqslant \max_{k=1}^{n} \sum_{j=1}^{n} |\mathbf{a}_{kj}| \\ &= ||\mathbf{A}||. \end{aligned} \tag{2.12}$$

La deuxième inégalité est due au fait que $\sum_{k=1}^{n} \mathbf{p}_{ik} \sum_{j=1}^{n} |\mathbf{a}_{kj}|$ est une combinaison convexe.

Ensuite, l'inégalité du triangle est appliquée une fois de plus pour montrer que

$$\begin{aligned}
||\mathbf{AP}|| &= \max_{i=1}^{n} \sum_{j=1}^{n} \left| \sum_{k=1}^{n} \mathbf{a}_{ik}\mathbf{p}_{kj} \right| \\
&\leqslant \max_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} |\mathbf{a}_{ik}\mathbf{p}_{kj}| \\
&= \max_{i=1}^{n} \sum_{k=1}^{n} |\mathbf{a}_{ik}| \sum_{j=1}^{n} \mathbf{p}_{kj} \\
&= \max_{i=1}^{n} \sum_{k=1}^{n} |\mathbf{a}_{ik}| \\
&= ||\mathbf{A}||. \qquad (2.13)
\end{aligned}$$

Cela est valide car pour une matrice stochastique $\mathbf{P}$, $\sum_{j=1}^{n} \mathbf{p}_{kj} = 1$.

La propriété 2.11 est vérifiée pour $N = 1$. S'il est supposé qu'elle est vraie jusqu'à $N - 1$, en appliquant l'inégalité du triangle,

$$\begin{aligned}
||\mathbf{P}^N - \mathbf{Q}^N|| &= ||\mathbf{P}(\mathbf{P}^{N-1} - \mathbf{Q}^{N-1}) + (\mathbf{P} - \mathbf{Q})\mathbf{Q}^{N-1}|| \\
&\leqslant ||\mathbf{P}(\mathbf{P}^{N-1} - \mathbf{Q}^{N-1})|| + ||(\mathbf{P} - \mathbf{Q})\mathbf{Q}^{N-1}|| \\
&\leqslant ||(\mathbf{P}^{N-1} - \mathbf{Q}^{N-1})|| + ||(\mathbf{P} - \mathbf{Q})|| \\
&\leqslant N||(\mathbf{P} - \mathbf{Q})||.
\end{aligned}$$

La deuxième inégalité est vraie en vertu des équations (2.12) et (2.13). Finalement, la dernière ligne est obtenue en invoquant l'hypothèse de récurrence. □

Résultat 2.1. *(Kingman, 1982b) En mesurant le temps en unité de N générations, le processus de généalogie $\{\mathcal{R}_{\lfloor Nt \rfloor}, t \geqslant 0\}$ dérivé du modèle de Wright-Fisher tend en distribution vers le processus de coalescence lorsque $N \to \infty$:*

$$\mathbf{P}_N^{\lfloor Nt \rfloor} \underset{N \to \infty}{\longrightarrow} \exp(t\mathbf{Q}).$$

Démonstration. Selon le lemme 2.1,

$$\begin{aligned}\left\|\mathbf{P}_N^{\lfloor Nt\rfloor} - \exp\left(\frac{1}{N}\lfloor Nt\rfloor\mathbf{Q}\right)\right\| \leqslant & Nt\left\|\mathbf{P}_N - \exp\left(\frac{1}{N}\mathbf{Q}\right)\right\| \\ = & Nt\left\|\mathbf{I} + \frac{1}{N}\mathbf{Q} + \mathcal{O}\left(\frac{1}{N^2}\right) - \exp\left(\frac{1}{N}\mathbf{Q}\right)\right\| \\ = & \mathcal{O}\left(\frac{1}{N}\right),\end{aligned}$$

en vertu de l'équation (2.10). □

Il existe une version du modèle de Wright-Fisher tenant compte de la nature diploïde et sexuée de certains organismes vivants (incluant les humains). Möhle (1998) a montré que les généalogies se comportaient également comme le processus de coalescence à la limite lorsque le nombre d'individus dans la population est très grand.

CHAPITRE III

GÉNÉTIQUE FAMILIALE

Les modèles probabilistes en génétique se distinguent en deux grandes spécialités : la génétique familiale et la génétique des populations. Le modèle de Wright-Fisher présenté au chapitre 2 est un exemple de cette dernière approche. Le chapitre présent traite des bases probabilistes de la génétique familiale.

Cette présentation est originale puisque l'emphase est mise sur la construction de généalogies sous forme d'arbres. Plusieurs concepts de base de la génétique des familles interviennent dans les développements et ils sont présentés au fur et à mesure.

3.1 Pedigree

Les relations de parenté à l'intérieur d'une famille peuvent être décrites par un pedigree : un graphe dirigé acyclique (Lauritzen et Sheehan, 2003) où une flèche $a \to b$ signifie a est parent de b. Un exemple est donné à la figure 3.1. Chacun des sommets possède deux parents ou zéro ; ces derniers individus sont appelés fondateurs. De plus, les flèches seront parfois omises avec la convention que les parents sont toujours placés plus haut que leurs enfants.

Définition 3.1. *Un pedigree est un graphe dirigé acyclique décrit par une paire*

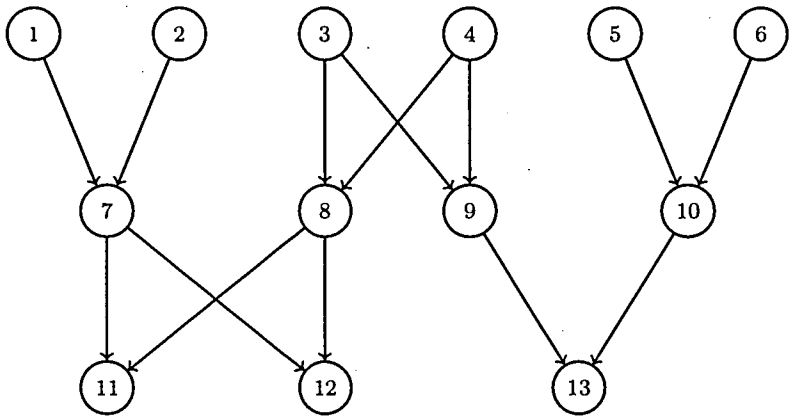


Figure 3.1 : Pedigree sur trois générations illustrant la relation entre les frères et soeurs 11 et 12 et leur cousin 13. Les fondateurs sont les individus 1 à 6.

$(\mathcal{I}, \pi)$: $\mathcal{I}$ est un ensemble de gènes chacun appartenant soit aux fondateurs $\mathcal{F}$, soit à leurs descendants (non-fondateurs) $\mathcal{D}$ de sorte que $\mathcal{I} = \mathcal{F} \cup \mathcal{D}$ et

$$\pi : \mathcal{D} \to \mathcal{I} \times \mathcal{I}$$

est la fonction associant à chaque membre $i \in \mathcal{D}$ ses deux parents $\pi_0(i)$ et $\pi_1(i)$. De plus, $\mathcal{I}$ est dans un ordre topologique : $\pi_0(i), \pi_1(i) < i$ pour tout $i \in \mathcal{D}$.

Il est à noter que la définition 3.2 s'applique également lorsqu'il y a plusieurs graphes disjoints.

Puisqu'un pedigree est considéré comme un graphe fini ($|\mathcal{I}| < \infty$), il existe une borne sur la longueur des chemins sur celui-ci. Cette borne sera appelée la hauteur du pedigree et sera invoquée dans plusieurs développements concernant les distances sur un pedigree. Le concept est illustré à la figure 3.2.

Définition 3.2. *La hauteur $\Lambda(\mathcal{P})$ d'un pedigree $\mathcal{P} = (\mathcal{I}, \pi)$ est la longueur maximale parmi l'ensemble des chemins reliant un des fondateurs à un autre membre du pedigree.*

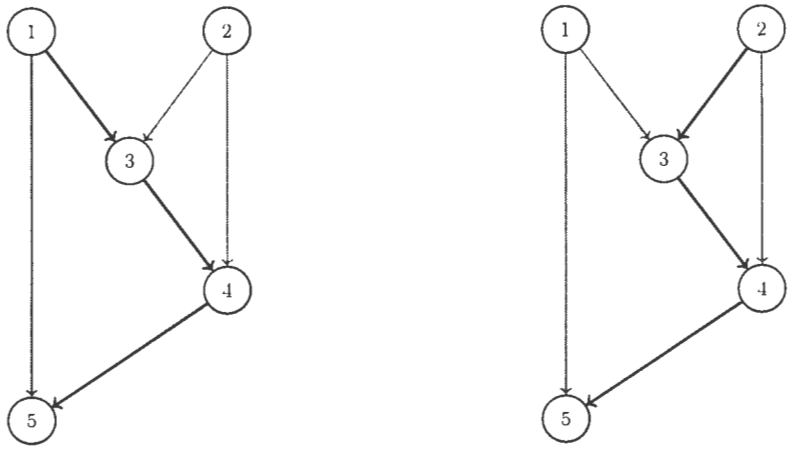


Figure 3.2 : Un exemple de pedigree de hauteur $\Lambda(\mathcal{P}) = 3$. Les deux chemins de longueur maximale sont en bleu.

3.1.1 Ploïdie

Afin de situer l'approche décrite dans ce chapitre, il est nécessaire de faire une courte parenthèse et de mentionner brièvement le principe de ploïdie en biologie. Les contraintes liées à la ploïdie pour la reproduction de gènes sont ignorées dans ce mémoire contrairement à l'approche courante en génétique familiale. Des arguments sont donnés pour justifier en quoi une généralisation à la reproduction diploïde et sexuée est assez immédiate.

Tous les individus d'une espèce diploïde et sexuée, par exemple les humains, possèdent deux copies de chaque gène, chacune provenant d'un seul des deux parents qui sont de sexe opposé (mère, père). La littérature de la génétique des familles traite nécessairement de ce type d'espèce puisque la contrainte d'avoir deux parents est intimement liée au caractère diploïde. Si les sommets sur le pedigree de la figure 3.1 représentent des individus, ils seraient remplacés dans une approche tenant compte de la nature diploïde par leurs deux gènes et leurs relations de parenté respectives comme à la figure 3.3.

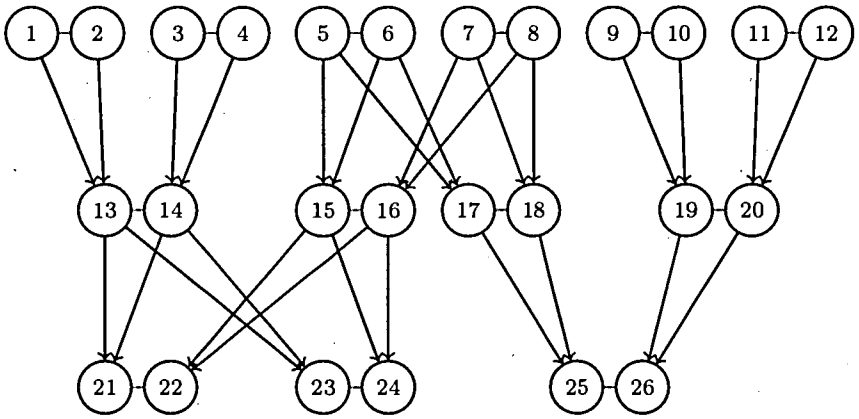


Figure 3.3 : Pedigree montrant les relations familiales entre les gènes pour des individus diploïdes. Les liens en gris sont symboliques ; ils montrent l'appartenance de deux gènes à un même individu, mais ne font pas partie du graphe. Par exemple, les parents du gène 13 sont les gènes 1 et 2 qui appartiennent au même individu.

La structure du pedigree du point de vue des gènes est la même que les DAGs décrits par la définition 3.2 : chaque sommet possède deux parents ou zéro. Le terme pedigree peut être entendu comme ce type de graphe sur les gènes.

Cependant, dans ce développement, la nécessité d'avoir un nombre pair de gènes et l'aspect sexué de la reproduction sont ignorés ; les pedigrees auront une structure moins contrainte et plus générale dans l'objectif d'utiliser des définitions plus succinctes. Le graphe de la figure 3.1 servira parfois d'exemple de pedigree malgré qu'il possède un nombre impair de sommets. Cette approche simplifiée permettra d'«arrimer» la génétique familiale au modèle de Wright-Fisher standard, présenté au chapitre 2, pour introduire le modèle unifié du chapitre 4.

3.1.2 Indicateurs de méiose

La propagation d'entités génétiques indivisibles (gènes) sur un pedigree est entièrement décrite par la première loi de Mendel (1865) : un gène est hérité d'un seul de ses parents et le choix se fait avec équiprobabilité. Cette loi est formalisée en introduisant des variables aléatoires appelées indicateurs de méiose (Thompson, 2000).

Définition 3.3. *À chacun des membres non-fondateurs $i \in \mathcal{D}$ d'un pedigree est associé un indicateur de méiose S_i. Les variables aléatoires S_i sont indépendantes et identiquement distribuées de loi Bernoulli$(1/2)$. Elles décrivent la provenance d'un gène parmi les deux parents.*

L'ensemble des indicateurs de méiose est noté $\mathbf{S} = \{S_i : i \in \mathcal{D}\}$. Il encode tout le caractère aléatoire en génétique familiale. En ce sens, toute variable aléatoire sur le pedigree est une fonction des indicateurs de méiose.

Lorsqu'ils sont réalisés, chaque gène, à l'exception des fondateurs, ne possède qu'un seul parent comme dans le modèle de population de Wright-Fisher ; l'indicateur de méiose S_i associé à un non-fondateur $i \in \mathcal{D}$ détermine l'unique parent qui transmet son matériel génétique à i :

$$\alpha(i) = \pi_{S_i}(i). \tag{3.1}$$

La variable aléatoire $\alpha(i)$ est une fonction de l'indicateur de méiose S_i. Une réalisation de $\mathbf{S}$ est montrée à la figure 3.4.

3.2 Généalogies

Lorsque tous les indicateurs de méiose $\mathbf{S}$ pour un pedigree $\mathcal{P} = (\mathcal{I}, \pi)$ sont réalisés, à tout membre $i \in \mathcal{I}$ correspond un et un seul fondateur. Autrement dit, il existe

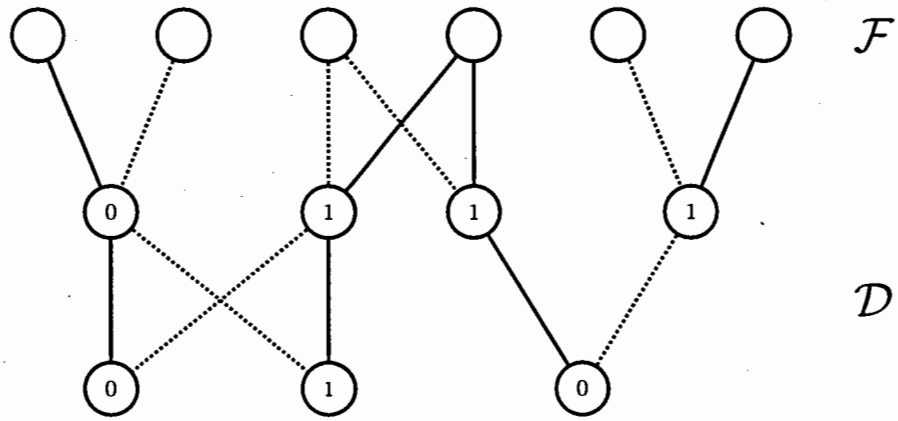


Figure 3.4 : Une réalisation des indicateurs de méiose détermine le parent duquel (arête en noir) un gène est hérité (sauf pour les fondateurs). Sur la figure, la valeur 0 ou 1 correspond à choisir, respectivement, le parent de gauche ou de droite.

$\lambda_i \in \{0, 1, \ldots, \Lambda(\mathcal{P})\}$ tel que la composition $\alpha^{\lambda_i}(i) \in \mathcal{F}$. Cela suggère la fonction

$$\varphi_{\mathbf{S}} : \mathcal{I} \to \mathcal{F}$$

associant à chacun des gènes son unique fondateur. La relation sera exprimée en notation abrégée φ, mais est une fonction des indicateurs de méiose.

À un fondateur $i \in \mathcal{F}$ correspond lui même : $\varphi(i) = \alpha^0(i) = i$. Pour l'exemple de la figure 3.5 la fonction φ se décrit par

$$\varphi(i) = \begin{cases} i & i \in \mathcal{F} = \{1, \ldots, 6\}, \\ 1 & i = 7, 11, \\ 4 & i = 8, 9, 12, 13, \\ 6 & i = 10. \end{cases}$$

La généalogie d'un échantillon γ_n provenant d'un pedigree est, dans sa forme la plus fondamentale, un ensemble d'arbres entièrement déterminé par les indicateurs de méiose $\mathbf{S}$. Elle est un sous graphe du pedigree constitué de tous les chemins

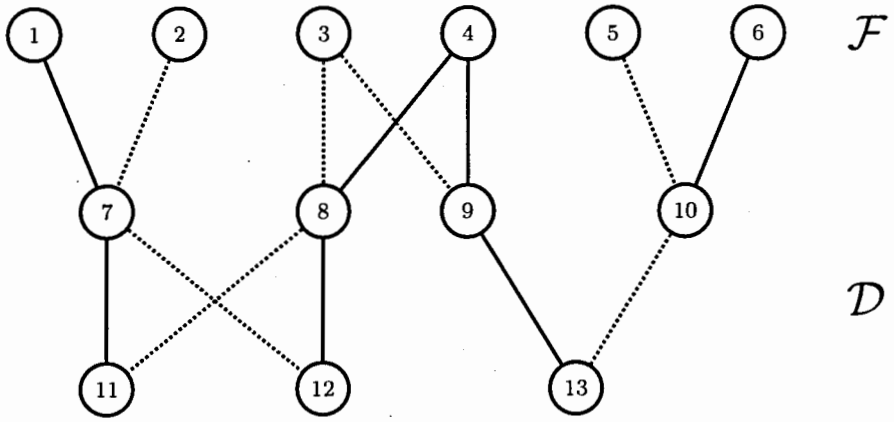


Figure 3.5 : Pour un membre du pedigree, il existe un seul chemin reliant un fondateur à celui-ci lorsque les indicateurs de méiose $\mathbf{S}$ sont réalisés.

reliant γ_n aux fondateurs $\mathcal{F}$. Ces chemins sont formés des arêtes

$$\mathbf{g}_i(\mathbf{S}) = \left\{[\varphi(i) = \alpha^{\lambda_i}(i), \alpha^{\lambda_i - 1}(i)], \ldots, [\alpha^2(i), \alpha(i)], [\alpha(i), i]\right\}$$

pour $i \in \gamma_n$. Un exemple est donné à la figure 3.6. La généalogie est la réunion de tous ces chemins qui relient chacun des $i \in \gamma_n$ à son fondateur :

$$\mathbf{G}_{\gamma_n}(\mathbf{S}) = \bigcup_{i \in \gamma_n} \mathbf{g}_i(\mathbf{S}).$$

Seuls l'échantillon γ_n, les fondateurs et les évènements de coalescence (sommets avec deux enfants ou plus) sont informatifs. Comme à la section 2.1.1, les autres sommets seront parfois omis sur les représentations de généalogies, remplacés par la longueur des branches en nombre d'évènements de reproduction. La structure de la généalogie sera résumée par un ensemble d'arbres avec l'échantillon sur les feuilles et les racines provenant des fondateurs. Des exemples sont illustrés à la figure 3.7.

Les sommets de l'échantillon peuvent également représenter des évènements de coalescence. Cela peut se produire lorsque l'on échantillonne des gènes avec cer-

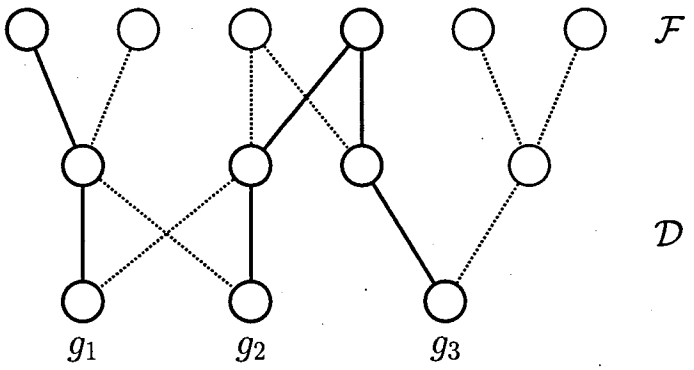


Figure 3.6 : Un sous-graphe du pedigree est en rouge. Celui-ci correspond à une réalisation d'une généalogie.

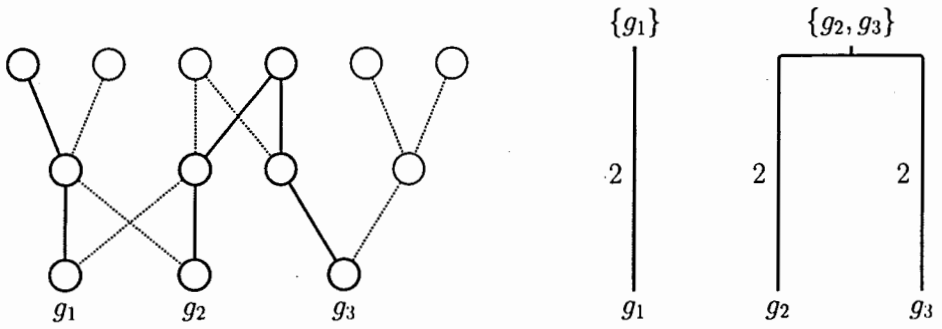
tains de leurs ancêtres sur le pedigree comme illustré à la figure 3.8. Dans ces cas, les gènes appartenant à l'échantillon sont tout de même placés sur les feuilles de la généalogie sur une branche de longueur zéro pour garder un représentation unifiée avec les généalogies présentées à la section 2.1.1 qui surviennent dans le cadre du modèle de Wright-Fisher.

Contrairement au processus de coalescence, les généalogies de génétique familiale ne sont pas limitées à des arbres binaires ; par exemple, trois gènes peuvent trouver un ancêtre commun au même endroit. Cela se produit potentiellement lorsque plusieurs gènes ont un parent en commun sur le pedigree comme ceux illustrés à la figure 3.9.

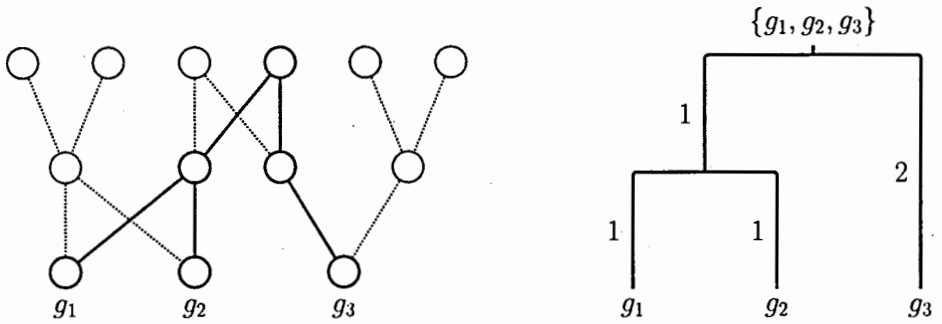
3.2.1 Simulation de généalogies

Lorsque tous les indicateurs de méiose $\mathbf{S}$ sont réalisés, la fonction

$$\alpha : \mathcal{D} \to \mathcal{I}$$



(a)



(b)

Figure 3.7 : Deux exemples de généalogies comme des sous-graphes d'un même pedigree et les arbres qui leurs correspondent. (a) La généalogie est composée de deux arbres disjoints correspondant à deux fondateurs distincts. (b) Toutes les lignées coalescent sur le pedigree et la généalogie est composée d'un seul arbre.

associant à chacun des non-fondateurs son unique parent est complètement spécifiée. Pour chacun des membres $g \in \gamma_n$ de l'échantillon, il existe un unique chemin partant d'un fondateur $\varphi(g) \in \mathcal{F}$ vers g. La réunion de ces chemins est un ensemble d'arbres qui représente la généalogie.

Cependant, pour construire la généalogie, tous les indicateurs de méiose n'ont pas

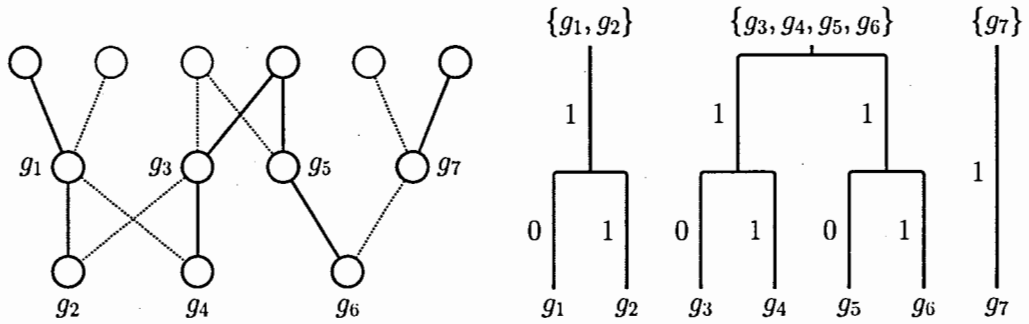


Figure 3.8 : Généalogie d'un échantillon de sept gènes composé des trois gènes $\{g_2, g_4, g_6\}$ et de l'ensemble $\{g_1, g_3, g_5, g_7\}$ de leurs parents sur le pedigree.

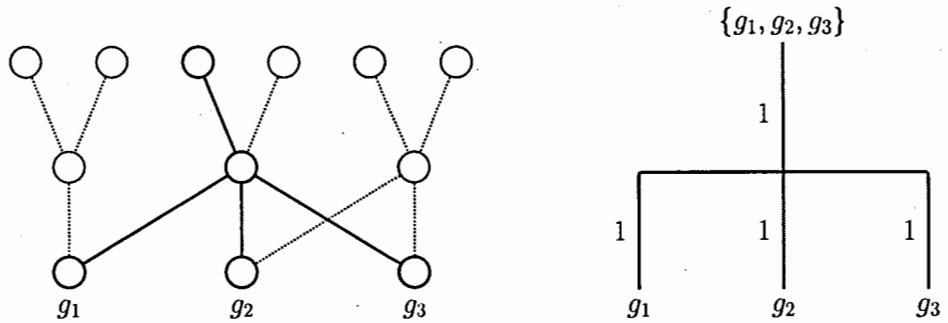


Figure 3.9 : Les trois gènes de l'échantillon coalescent dans le même ancêtre.

besoin d'être observés. Par exemple, il est naturel de penser que la généalogie est indépendante d'un indicateur de méiose S_i associé à un gène i qui n'est ancêtre (sur le pedigree) d'aucun membre de l'échantillon.

Une façon de raisonner cette dernière affirmation est de prendre une réalisation quelconque $\{\mathbf{S} = \mathbf{s}\}$ des indicateurs de méiose. Si l'on modifie la valeur d'un indicateur de méiose S_i pour un gène i qui n'est pas ancêtre d'un gène $g \in \gamma_n$, aucun chemin reliant un fondateur à un membre de l'échantillon n'est modifié. Autrement dit, la généalogie n'est pas une fonction de ces indicateurs de méiose.

Soit $g_{\max} = \max \gamma_n$, le gène maximal de l'échantillon pour un certain ordre topolo-

gique sur le pedigree (résultat 1.1). Une généalogie $\mathbf{G}_{\gamma_n}(\mathbf{S})$ peut être décomposée en une arête $[\alpha(g_{\max}), g_{\max}]$ et une généalogie sur un nouvel échantillon composé de γ_n sans $g_{\max}$ et incluant $\alpha(g_{\max})$:

$$\mathbf{G}_{\gamma_n}(\mathbf{S}) = \mathbf{G}_{\gamma_n \setminus \{g_{\max}\} \cup \{\alpha(g_{\max})\}}(\mathbf{S}) \cup [\alpha(g_{\max}), g_{\max}].$$

La finesse de cette décomposition provient du fait que $g_{\max}$ ne se retrouve pas sur cette nouvelle généalogie. En effet, le gène $g_{\max}$ n'est ancêtre d'aucun autre membre de γ_n ni de $\alpha(g_{\max})$ par la propriété d'ordre topologique. La nouvelle généalogie est donc indépendante de l'indicateur de méiose $S_{g_{\max}}$.

L'algorithme 3.1 exploite cette relation récursive ; il fixe $\alpha(g_{\max})$ en simulant l'indicateur $S_{g_{\max}}$ en premier. La procédure est ensuite itérée sur la nouvelle généalogie. Toutes les compositions $\alpha^j(g)$ pour tout membre $g \in \gamma_n$ de l'échantillon sont déterminées jusqu'à atteindre les fondateurs. La suite d'indicateurs de méiose simulés dépend de l'ordre topologique qui n'est pas forcément unique. Une application de l'algorithme 3.1 est illustrée à la figure 3.10. De plus, l'utilité de l'algorithme 3.1 est discutée dans l'appendice A présentant une série d'expériences liées au concept de généalogie en génétique familiale.

3.2.2 État IBD

Deux gènes sont identiques par descendance (IBD) s'ils ont un ancêtre commun sur la généalogie (ils coalescent sur le pedigree). Cela se produit si et seulement si ils proviennent d'un même fondateur.

Définition 3.4. *Pour un échantillon de n gènes $\gamma_n \subset \mathcal{I}$ provenant d'un pedigree, l'état IBD* $\mathbf{J}(\mathbf{S})$ *(Thompson, 2000) est la partition de γ_n formée à partir de la relation d'équivalence*

$$a \overset{IBD}{=} b \quad \Leftrightarrow \quad \varphi(a) = \varphi(b)$$

$E = \gamma_n \setminus \mathcal{F}$ (Les fondateurs sont écartés)

TANT QUE $E \neq \emptyset$

$g_{\max} = \max E$ (E est ordonné)

$S_{g_{\max}} \sim \mathcal{B}\text{ernoulli}\,(1/2)$

$\alpha(g_{\max}) = \pi_{S_{g_{\max}}}(g_{\max})$

$\varphi(g_{\max}) \to \varphi(\alpha(g_{\max}))$

$E = E \setminus \{g_{\max}\}$ ($g_{\max}$ est retiré)

si $\alpha(g_{\max}) \in \mathcal{F}$

$\varphi(\alpha(g_{\max})) = \alpha(g_{\max})$

sinon

$E = E \cup \{\alpha(g_{\max})\}$ ($\alpha(g_{\max})$ peut être déjà dans E)

FIN **si**

FIN **TANT QUE**

Algorithme 3.1 : Simulation d'une généalogie pour un échantillon γ_n provenant d'un pedigree $\mathcal{P} = (\mathcal{I}, \pi)$. L'algorithme donne aussi explicitement l'état IBD en calculant φ pour les membres de l'échantillon γ_n. Une utilisation de l'algorithme 3.1 est présentée dans l'appendice A.

pour $a, b \in \gamma_n$.

Alors, $\mathbf{J(S)}$ est une variable aléatoire sur l'espace $\mathcal{C}_n$ des partitions de γ_n entièrement déterminée par $\mathbf{S}$. Sa fonction de masse

$$P(\mathbf{J(S)} = \xi), \qquad \xi \in \mathcal{C}_n,$$

est difficile à caractériser car elle dépend fortement de la structure du pedigree.

L'état IBD est lié à la généalogie du point de vue suivant : pour deux gènes a et b, $\varphi(a) = \varphi(b)$ si et seulement si a et b appartiennent au même arbre sur la

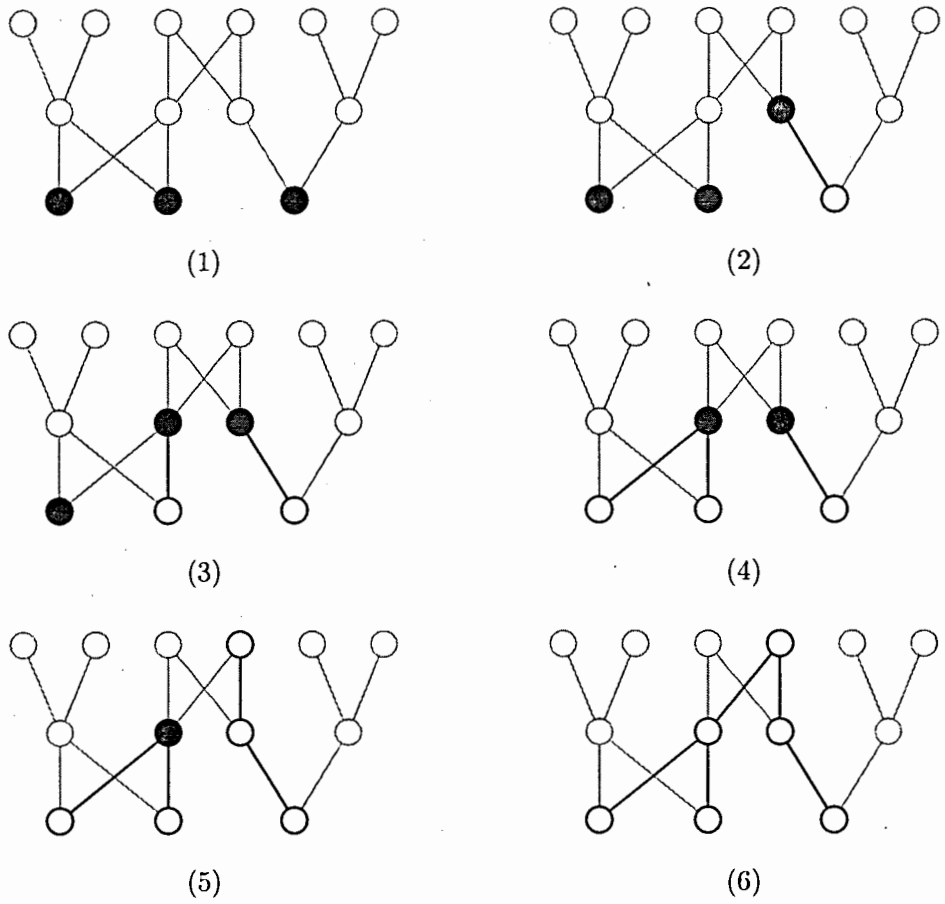


Figure 3.10 : Un exemple du déroulement de l'algorithme 3.1. Les sommets pleins représentent l'ensemble E à chacun des passage dans la boucle TANT QUE.

généalogie. En ce sens, $\mathbf{J(S)}$ est connu lorsque la généalogie $\mathbf{G}_{\gamma_n}(\mathbf{S})$ est observée.

Comme il a été discuté à la section 3.2.1, les gènes qui ne sont pas ancêtres d'au moins un des membres de l'échantillon n'ont aucune influence sur la généalogie de ceux-ci et donc, par extension, sur l'état IBD. Ces sommets peuvent être retirés sans conséquences pour la loi de probabilité de $\mathbf{J(S)}$.

De plus, certains sommets jouent un rôle comparable à celui d'un entonnoir sur le pedigree. Tous les chemins partant d'un ancêtre d'un «entonnoir» $e \in \mathcal{D}$ vers

un membre $g \in \gamma_n$ de l'échantillon passent par e. Le principe est illustré à la figure 3.11.

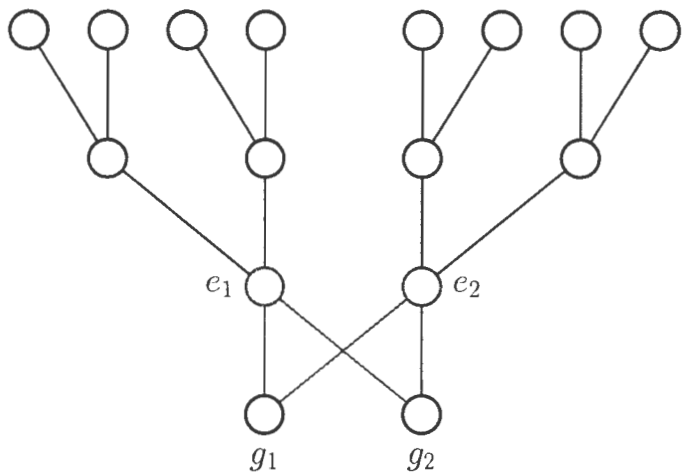


Figure 3.11 : Les sommets e_1 et e_2 sont des entonnoirs par rapport à l'échantillon $\{g_1, g_2\}$; ils peuvent jouer le rôle de fondateur car si g_1 et g_2 sont IBD, ils le sont déjà à partir de e_1 ou e_2. La partie du pedigree en bleu peut être élaguée sans modifier la loi de probabilité de l'état IBD $\mathbf{J}(\mathbf{S})$.

Dans ces cas, si à des membres de l'échantillon correspond un fondateur ancêtre de e, ils sont IBD au moment où e est atteint. Le sommet «entonnoir» e peut donc jouer le rôle de fondateur et ses ancêtres sont élagués.

Un façon d'argumenter cette dernière conclusion est en partant d'une réalisation quelconque des indicateurs de méiose $\{\mathbf{S} = \mathbf{s}\}$ spécifiant une généalogie $\mathbf{G}_{\gamma_n}(\mathbf{s})$. Soit S_i, un indicateur pour i un «entonnoir» e ou un des ses ancêtres sur le pedigree.

Si i est sur un arbre de la généalogie, cet arbre ne contient que les membres de l'échantillon descendants de e. En modifiant $\mathbf{S}_i$, un nouvel arbre est créé, mais celui-ci contient toujours exclusivement les membres de l'échantillon descendants

de e car tous les chemins des ancêtres de e passent par lui pour rejoindre l'échantillon. La relation d'équivalence (définition 3.4) est donc inchangée lorsque l'on modifie la valeur de S_i ; l'état IBD n'est pas une fonction d'un tel indicateur de méiose.

3.2.3 Probabilité de coalescence

Comme constaté au début de la section 3.2 consacrée aux généalogies, tous les gènes d'un échantillon ne sont pas obligés de coalescer sur le pedigree. Pour une paire de gènes a et b, la probabilité qu'ils trouvent un ancêtre commun sur la généalogie sera notée f_{ab}.

L'état IBD pour un échantillon de $n = 2$ gènes prend ses valeurs dans les deux partitions de $\{a, b\}$ possibles : $\{\{a\}, \{b\}\}$ et $\{\{a, b\}\}$. La quantité f_{ab} correspond à la probabilité

$$P(\mathbf{J}(\mathbf{S}) = \{\{a, b\}\})$$

pour un échantillon $\gamma_2 = \{a, b\}$. Autrement dit, si les deux gènes ont coalescé, ils appartiennent à la même classe d'équivalence.

Cette probabilité dépend donc de la structure du pedigree $\mathcal{P}$ et des indicateurs de méiose $\mathbf{S} = \{S_i : i \in \mathcal{D}\}$ associés. Puisqu'elle s'exprime comme une probabilité sur l'état IBD, elle peut être retrouvée à partir du pedigree élagué des sommets qui ne sont pas ancêtres de a ou b et des sommets qui sont ancêtre d'un «entonnoir» par rapport à l'échantillon $\{a, b\}$ comme décrit à la section 3.2.2.

Deux gènes coalescent sur le pedigree si et seulement si ils proviennent d'un même fondateur. La probabilité d'une certaine réalisation $\{\mathbf{S} = \mathbf{s}\}$ des indicateurs de méiose est $1/2^{|\mathcal{D}|}$; alors, en sommant sur l'ensemble $\mathbf{S}$ des indicateurs de méiose,

la probabilité f_{ab} s'exprime comme

$$f_{ab} = P(\varphi(a) = \varphi(b)) = \frac{1}{2^{|\mathcal{D}|}} \sum_{\mathbf{s} \in \{0,1\}^{|\mathcal{D}|}} P(\varphi(a) = \varphi(b) | \mathbf{S} = \mathbf{s}). \tag{3.2}$$

Puisque la fonction φ est déterminée par $\mathbf{S}$, chaque terme de la somme de l'équation (3.2) prend la valeur 0 ou 1.

La valeur f_{ab} est un analogue haploïde (un seul gène par individu) du coefficient de parenté (Malécot, 1948) ; elle peut également être calculée de manière récursive tel que démontré par le résultat suivant.

Résultat 3.1. *Puisque la probabilité f_{ab} est symétrique pour a et b, il est supposé, sans perte de généralité, que $a \geqslant b$. Si $a = b$, a et b proviennent du même fondateur avec probabilité 1 : $f_{ab} = 1$. Si $a > b$ et que a est un fondateur ($a \in \mathcal{F}$), alors $f_{ab} = 0$. Sinon, le coefficient f_{ab} s'exprime dans une forme récursive en fonction de b et des parents de a sur le pedigree :*

$$f_{ab} = \frac{1}{2}(f_{\pi_0(a)b} + f_{\pi_1(a)b}).$$

Démonstration. Premièrement, pour le cas $a = b$, la probabilité de coalescence s'exprime

$$f_{ab} = P(\varphi(a) = \varphi(a)) = 1.$$

Autrement dit, il est considéré qu'un gène coalesce avec lui-même avec probabilité 1.

Ensuite, si $a > b$ et a est un fondateur, alors a est son propre fondateur : $\varphi(a) = a$. À cause de la propriété d'ordre topologique (résultat 1.1), $a > b$ implique que a n'est pas ancêtre de b sur le pedigree. Il est donc impossible que a soit le fondateur de b :

$$f_{ab} = P(a = \varphi(b)) = 0.$$

Sinon, lorsque $a > b$ et $a \notin \mathcal{F}$, la stratégie est d'exprimer f_{ab} en sommant sur les valeurs possibles de l'indicateur de méiose S_a. Les deux parents potentiels pour a sur la généalogie sont $\pi_0(a)$ et $\pi_1(a)$ et un des deux est choisi avec probabilité uniforme $1/2$:

$$\begin{aligned} f_{ab} &= P(\varphi(a) = \varphi(b)) \\ &= \frac{1}{2}\left[P(\varphi(a) = \varphi(b)|S_a = 0) + P(\varphi(a) = \varphi(b)|S_a = 1)\right]. \end{aligned} \tag{3.3}$$

Lorsque l'indicateur de méiose S_a est fixé à une valeur $s \in \{0, 1\}$, le parent de a sur la généalogie est $\alpha(a) = \pi_s(a)$; ils proviennent donc du même fondateur :

$$\varphi(a) = \varphi(\pi_s(a)).$$

Alors,

$$P(\varphi(a) = \varphi(b)|S_a = s) = P(\varphi(\pi_s(a)) = \varphi(b)|S_a = s), \quad s \in \{0, 1\}. \tag{3.4}$$

De plus, les choix des fondateurs $\varphi(\pi_s(a))$ et $\varphi(b)$ sont indépendants de S_a car a ne fait pas partie des ancêtres de $\pi_s(a)$ ou de b par la propriété d'ordre topologique $(\pi_s(a), b < a)$. Ainsi, son indicateur de méiose n'a aucune influence sur la généalogie de ces derniers.

Donc, la probabilité que $\pi_s(a)$ et b soient IBD est indépendante de S_a :

$$P(\varphi(\pi_s(a)) = \varphi(b)|S_a = s) = P(\varphi(\pi_s(a)) = \varphi(b)), \quad s \in \{0, 1\}. \tag{3.5}$$

Par les équations (3.4) et (3.5),

$$P(\varphi(a) = \varphi(b)|S_a = s) = P(\varphi(\pi_s(a)) = \varphi(b)), \quad s \in \{0, 1\}, \tag{3.6}$$

et la forme récursive du résultat 3.1 est retrouvée en substituant (3.6) dans (3.3).

□

Afin de calculer f_{ab} pour toutes les paires (a, b) parmi les membres $\mathcal{I}$ d'un pedigree, il serait inutilement onéreux de les calculer successivement car plusieurs calculs de coefficients intermédiaires se répètent. La stratégie optimale est de les calculer dans un ordre astucieux et de les garder en mémoire afin d'éviter les répétitions. Par exemple, lorsque la forme récursive

$$f_{ab} = \frac{1}{2}(f_{\pi_0(a)b} + f_{\pi_1(a)b})$$

est employée, il serait profitable d'avoir calculé $f_{\pi_0(a)b}$ et $f_{\pi_1(a)b}$ au préalable. L'algorithme 3.2 est une adaptation de celui proposé par Lange (2003, section 5.2) pour calculer les coefficients de parenté. À chacun des passages dans la boucle, si la forme récursive doit être utilisée, les coefficients intermédiaires sont déjà calculés, par les trois inégalités $\pi_0(a), \pi_1(a), b < a$.

POUR $a = 1, 2, \ldots, I$
  POUR $b = 1, 2, \ldots, a$
    si $a = b$
      $f_{ab} = 1$
    sinon si $a \in \mathcal{F}$
      $f_{ab} = 0$
    sinon
      $f_{ab} = \frac{1}{2}(f_{\pi_0(a)b} + f_{\pi_1(a)b})$
    FIN **si**
  FIN **POUR**
FIN **POUR**

Algorithme 3.2 : Calcul de la probabilité de coalescence f_{ab} pour toutes les paires $a, b \in \mathcal{I} = \{1, 2, \ldots, I\}$ sur un pedigree $\mathcal{P} = (\mathcal{I}, \pi)$.

3.3 Temps et distance sur la généalogie

La généalogie en génétique familiale est décrite par un ensemble d'arbres comme présenté à la section 3.2. Comme pour le modèle de Wright-Fisher, les temps sur la généalogie sont mesurés en nombre d'évènements de reproduction (arêtes sur les arbres). Ces temps sont toujours bornés d'une certaine façon par la hauteur finie du pedigree.

3.3.1 Temps total

Le nombre d'arêtes $T_{\mathrm{TOT}}^{(\mathcal{P})}$ ou la somme des tailles des arbres de la généalogie d'un échantillon γ_n provenant d'un pedigree $\mathcal{P}$ est analogue au temps total pour le processus de coalescence ou le modèle de Wright-Fisher. Sa loi de probabilité dépend fortement de la structure du pedigree.

Comme décrit plus en détails à la section 3.2, la généalogie en génétique familiale est un ensemble d'arbres composé de tous les chemins reliant un des fondateurs à un membre de l'échantillon. Chacun de ces chemins a une longueur plus petite ou égale à $\Lambda(P)$, la hauteur du pedigree, tel que décrit par la définition 3.2.

Le nombre d'arêtes sur la réunion de tous ces chemins est plus petit ou égal à la somme des longueurs des chemins car ils ne sont pas forcément disjoints. Ainsi, le temps total $T_{\mathrm{TOT}}^{(\mathcal{P})}$ pour une généalogie sur un pedigree $\mathcal{P}$ d'un échantillon de n gènes est borné :

$$T_{\mathrm{TOT}}^{(\mathcal{P})} \leqslant n\Lambda(\mathcal{P}),$$

où $\Lambda(P)$ est la hauteur du pedigree.

3.3.2 Distance

Soit deux gènes a et b qui coalescent sur le pedigree. Comme constaté à la section 3.2 sur les généalogies, les temps aux évènements de coalescence ne sont pas forcément symétriques pour a et b dans le sens que la distance entre a et l'évènement de coalescence de a et b n'est pas forcément égale à la distance entre b et ce même évènement de coalescence.

Par exemple, des gènes a et $b = \pi_0(a)$, un des deux parents de a, coalescent si l'indicateur de méiose $S_a = 0$. Le parent de a sur la généalogie est alors $\alpha(a) = b$. Du point de vue de b, le temps de coalescence est zéro alors qu'il a fallu un évènement de reproduction pour a avant de retrouver b.

Puisque le temps de coalescence est difficilement défini sur un pedigree, c'est la notion de distance qui sera développée. La distance $d_{\mathcal{P}}(a, b)$ entre deux gènes est définie lorsque a et b sont IBD sur le pedigree, c'est-à-dire qu'ils appartiennent au même arbre sur la généalogie. La distance pour deux sommets sur un arbre est décrite par la définition 1.8 ; il s'agit du nombre d'arêtes sur l'arbre minimal contenant a et b. La notion de distance sur une généalogie provenant d'un pedigree est illustrée à la figure 3.12.

Par exemple, à un gène g sur la généalogie correspond un unique fondateur

$$\varphi(g) = \alpha^{\lambda_g}(g).$$

Les deux gènes sont IBD et leur distance est bornée par la hauteur du pedigree :

$$d_{\mathcal{P}}(g, \varphi(g)) = \lambda_g \leqslant \Lambda(P). \tag{3.7}$$

En général, la distance entre deux gènes a et b, s'ils sont IBD, est plus petite ou égale à la longueur de deux chemins sur le pedigree ou au temps total si l'on

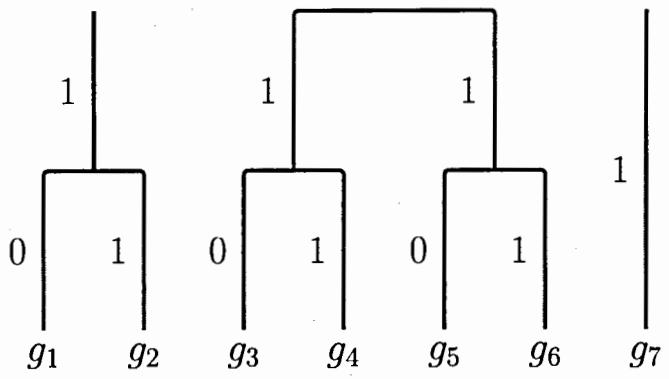


Figure 3.12 : Les arbres minimaux contenant respectivement $\{g_1, g_2\}$ et $\{g_3, g_6\}$ sont en bleu. Les distances $d_{\mathcal{P}}(g_1, g_2)$ et $d_{\mathcal{P}}(g_3, g_6)$ valent respectivement 1 et 3.

considère l'échantillon $\gamma_2 = \{a, b\}$. La distance $d_{\mathcal{P}}(a, b)$ est donc bornée lorsque définie :

$$d_{\mathcal{P}}(a, b) \leqslant 2\Lambda(P). \tag{3.8}$$

CHAPITRE IV

P-COALESCENT

En génétique des populations, les relations familiales sont typiquement ignorées malgré qu'elles existent entre les individus. Une des raisons est que les relations familiales distantes ne sont tout simplement pas très bien documentées dans la plupart des cas.

Si le pedigree de toute la population était disponible, il serait possible, du moins théoriquement, d'appliquer les concepts du chapitre 3 pour construire des généalogies jusqu'au plus récent ancêtre commun d'un échantillon. Par exemple, cela permettrait de comparer certaines propriétés des généalogies sous Wright-Fisher avec celles prédites par la génétique familiale.

Wakeley *et al.* (2012) explorent cette idée en comparant, à l'aide de simulations, les temps de coalescence prévus sous le modèle de Wright-Fisher à ceux obtenus en traçant des généalogies à partir d'indicateurs de méiose. Pour ce faire, ils imaginent un grand pedigree regroupant l'ensemble d'une population. Ils arrivent à la conclusion que les temps avant le plus récent ancêtre commun à travers le pedigree pour une paire de gènes $\{a, b\}$ choisie aléatoirement dans une grande population ne sont pas si différents des temps exponentiels prédits par le processus de coalescence.

Cependant, si a et b sont plutôts choisis arbitrairement sur un pedigree avec un coefficient de parenté f_{ab} élevé, la loi de probabilité de leur temps de coalescence devrait être grandement affectée. L'approche décrite dans ce chapitre, le p-coalescent, permet de modéliser l'effet des liens de parenté dans un modèle de population comme celui de Wright-Fisher.

Plus généralement, le p-coalescent construit des généalogies dans l'esprit de la génétique des populations et, en particulier, du processus de coalescence à partir d'un échantillon de gènes γ_n provenant d'un pedigree. Ce sont les fondateurs du pedigree qui seront considérés en tant que membres d'une population. L'idée générale est représentée à la figure 4.1.

Supposition 4.1. *Le p-coalescent suppose que les fondateurs $\mathcal{F}$ du pedigree proviennent de la génération $\mathcal{G}_0$ d'une population composée de $N \to \infty$ individus se reproduisant selon le modèle de Wright-Fisher.*

La supposition 4.1 permettra de développer une nouvelle approche à rebours dans le temps pour construire des généalogies. Intuitivement, les généalogies familiales s'arrêtent lorsque les fondateurs sont atteints ; il serait envisageable de compléter la généalogie des fondateurs à partir des principes énoncés au chapitre 2 sur le modèle de Wright-Fisher.

4.1 Généalogies

La généalogie sous le p-coalescent est un arbre enraciné dans le plus récent ancêtre commun de l'échantillon γ_n. Elle est composée de deux parties distinctes : une dérivée de l'approche en génétique familiale et l'autre basée sur le modèle de Wright-Fisher et le processus de coalescence.

En génétique familiale, la généalogie est un ensemble d'arbres chacun enraciné

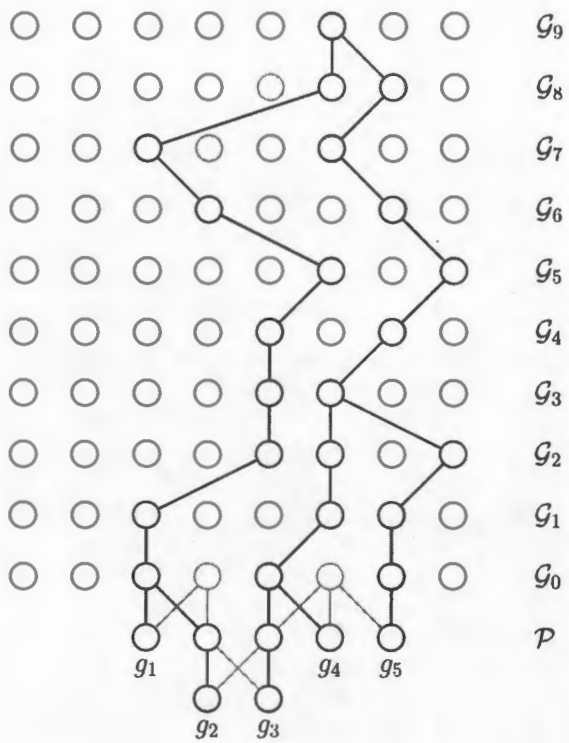


Figure 4.1 : Les gènes $g_1, g_2, \ldots, g_5$ proviennent d'un pedigree tandis que les fondateurs appartiennent à une population se reproduisant selon le modèle de Wright-Fisher.

dans un fondateur. En particulier, lorsque le pedigree est composé de plusieurs graphes disjoints, à chacun des sous graphes correspond au moins un arbre. Une généalogie sur un tel pedigree est considérée à la figure 4.2.

Un objectif primordial du p-coalescent est de modéliser la généalogie d'un échantillon γ_n provenant d'un pedigree comme un seul arbre. Les fondateurs qui forment les racines des arbres sur le pedigree ont une généalogie qui excède celui-ci.

Si ces fondateurs proviennent d'une grande population se reproduisant selon le modèle de Wright-Fisher (supposition 4.1), le processus de coalescence permet de

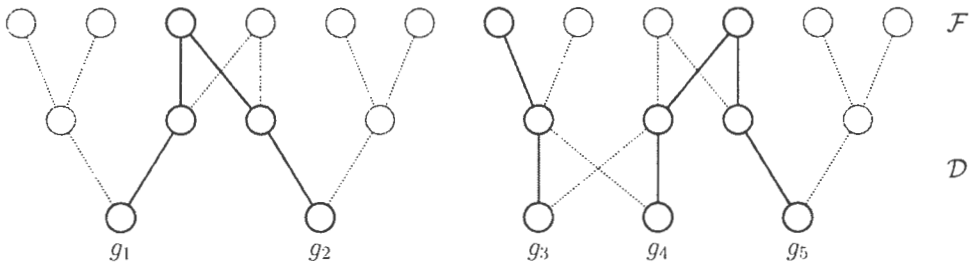


Figure 4.2 : Une généalogie sur un pedigree composé de deux sous graphes disjoints.

modéliser leur généalogie jusqu'au plus récent ancêtre commun (MRCA). Ainsi, un arbre binaire joint les fondateurs qui sont les racines de la généalogie familiale comme dans la représentation de la figure 4.3.

4.1.1 Construction de la généalogie

La généalogie $\mathbf{G}_{\gamma_n}(\mathbf{S})$ en génétique familiale, telle que décrite à la section 3.2, est formée par la réunion de tous les chemins qui partent des fondateurs $\mathcal{F}$ vers l'échantillon γ_n.

Le pedigree contraint et complexifie la généalogie, mais contient plus d'information que le modèle de Wright-Fisher. Entre autres, le nombre de parents potentiels sur la généalogie pour chacun des gènes est de deux à la place de N, le nombre d'individus dans la population. La généalogie des non-fondateurs sur le pedigree sera modélisée par la première loi de Mendel et les indicateurs de méiose.

Pour une certaine réalisation $\mathbf{G}_{\gamma_n}(\mathbf{s})$, l'état IBD est $\mathbf{J}(\mathbf{s})$ et les racines des arbres de la généalogie sur le pedigree forment un ensemble aléatoire

$$\varphi(\gamma_n) \equiv \bigcup_{i \in \gamma_n} \varphi(i)$$

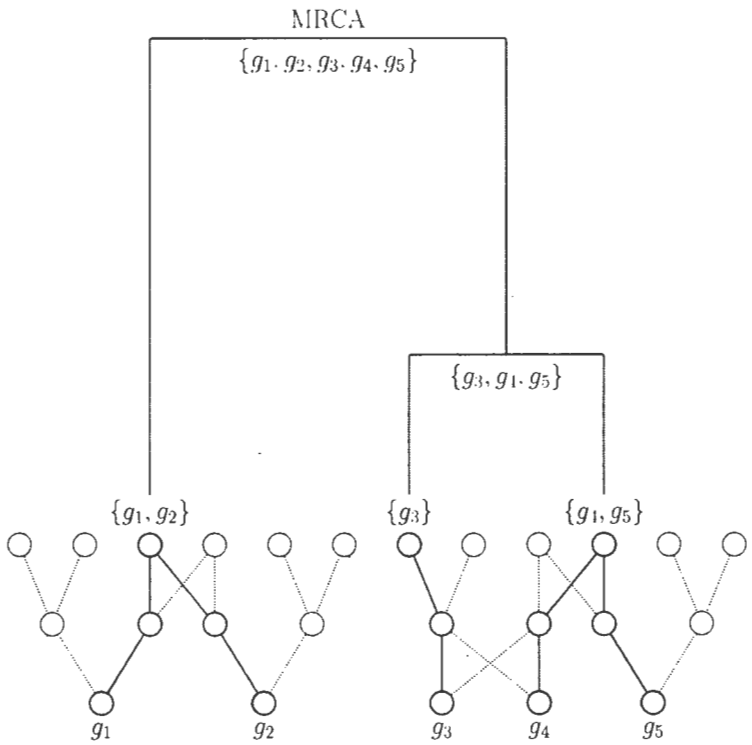


Figure 4.3 : Un arbre (binaire) de coalescence complète la généalogie provenant du pedigree à partir de l'état IBD $\{\{g_1, g_2\}, \{g_3\}, \{g_4, g_5\}\}$.

composé des fondateurs qui correspondent à au moins un membre de l'échantillon.

En vertu de la supposition 4.1, l'ensemble de fondateurs $\varphi(\gamma_n)$ provient de la génération $\mathcal{G}_0$ d'une population de Wright-Fisher. Puisque le modèle de Wright-Fisher est échangeable sur les membres de la population, le choix des fondateurs n'affecte pas la distribution de leur généalogie. Si la taille de la population est très grande, l'arbre est distribué selon le processus de coalescence sur un échantillon de $|\mathbf{J}(\mathbf{s})|$ gènes soit le nombre de classes d'équivalence de l'état IBD ou le nombre de fondateurs dans l'ensemble $\varphi(\gamma_n)$. Intuitivement, le processus de coalescence est appliqué sur un ensemble $\varphi(\gamma_n)$ aléatoire de gènes représentant l'état IBD de l'échantillon.

Conditionnellement à la généalogie familiale $\mathbf{G}_{\gamma_n}(\mathbf{s})$, le premier évènement du processus de coalescence reliant les fondateurs se produit après un temps

$$T_{|\mathbf{J}(\mathbf{s})|} \sim \mathcal{E}\mathrm{xp}\left[\binom{|\mathbf{J}(\mathbf{s})|}{2}\right]$$

mesuré en unité de N générations alors que deux fondateurs dans $\varphi(\gamma_n)$ coalescent. Deux arbres de $\mathbf{G}_{\gamma_n}(\mathbf{s})$ sont alors réunis sur la généalogie globale et, par le fait même, deux classes de $\mathbf{J}(\mathbf{s})$ n'en forment désormais qu'une seule. L'état IBD effectue donc une transition au temps $T_{|\mathbf{J}(\mathbf{s})|}$ vers un état η avec probabilité uniforme

$$\frac{1}{\binom{|\mathbf{J}(\mathbf{s})|}{2}}$$

pour $\mathbf{J}(\mathbf{s}) \prec \eta$.

Le processus $R_t^{(\mathbf{S})}$ reliant les classes d'équivalence de $\mathbf{J}(\mathbf{S})$ sur la généalogie est une chaîne de Markov à temps continu sur l'espace $\mathcal{C}_n$ des partitions de γ_n avec le même générateur $\mathbf{Q}$ que le processus de coalescence (définition 2.1). Il a la particularité de démarrer à l'état IBD $\mathbf{J}(\mathbf{S}) \in \mathcal{C}_n$ et de compléter la généalogie en $|\mathbf{J}(\mathbf{S})| - 1$ transitions.

Le processus $R_t^{(\mathbf{S})}$ est dépendant de la généalogie familiale $\mathbf{G}_{\gamma_n}(\mathbf{S})$ et, plus généralement, des indicateurs de méiose $\mathbf{S}$ puisque sa distribution initiale est celle de l'état IBD :

$$P(R_0^{(\mathbf{S})} = \xi) = P(\mathbf{J}(\mathbf{S}) = \xi), \quad \xi \in \mathcal{C}_n.$$

4.1.2 Simulation de généalogies

Le dessous de l'arbre est composé de chemins sur le pedigree. Cette généalogie familiale $\mathbf{G}_{\gamma_n}(\mathbf{S})$ est réalisée avant la partie générée par le processus de coalescence sur les fondateurs. L'algorithme 3.1 permet de simuler ces généalogies et d'en extraire l'état IBD.

Par contre, le dessus de l'arbre dépend de $\mathbf{G}_{\gamma_n}(\mathbf{S})$, mais seulement à travers l'état IBD $\mathbf{J}(\mathbf{S})$. L'algorithme 2.1 de simulation du processus de coalescence a été pensé afin de pouvoir démarrer à une partition arbitraire parmi $\mathcal{C}_n$. Conditionnellement à $\mathbf{G}_{\gamma_n}(\mathbf{S})$, il peut être initialisé à $\eta_0 = \mathbf{J}(\mathbf{S})$ afin de simuler le reste de la généalogie.

4.2 Temps sous le p-coalescent

Les temps sous le p-coalescent sont définis sur l'arbre correspondant à une réalisation de la généalogie comme celle illustrée à la figure 4.4. En mesurant les temps en unité de N reproductions, la longueur des branches qui proviennent du pedigree sont négligeables. Ce fait intuitif est exprimé formellement au résultat 4.1.

4.2.1 Distance

Le temps de coalescence pour deux gènes distincts a et b de la génération $\mathcal{G}_0$ d'une population se reproduisant selon le modèle de Wright-Fisher est distribué selon une loi $\mathcal{G}\text{éo}(1 - 1/N)$ comme discuté à la section 2.1.2. La distance $\delta(a, b)$ entre a et b (définition 1.8) est une variable aléatoire qui vaut deux fois le temps de coalescence sous le modèle de Wright-Fisher. Sa fonction de survie (équation (2.1)) converge vers celle d'une loi $\mathcal{E}\text{xp}(1/2)$ lorsque le temps $t \geqslant 0$ est mesuré en unité de N générations :

$$P(\delta(a, b) > Nt) = \left(1 - \frac{1}{N}\right)^{\left\lfloor \frac{Nt}{2} \right\rfloor} \underset{N \to \infty}{\longrightarrow} \exp\left(-\frac{t}{2}\right). \tag{4.1}$$

Il s'agit d'une conséquence du résultat 2.1 sur la convergence vers le processus de coalescence ; la distance entre deux gènes converge en loi vers deux fois T_2, le temps de coalescence de loi $\mathcal{E}\text{xp}(1)$ pour un échantillon de deux gènes.

Pour le p-coalescent, à cause de la partie de la généalogie sur le pedigree, le temps de coalescence d'une paire de gènes n'est pas forcément symétrique pour a et b,

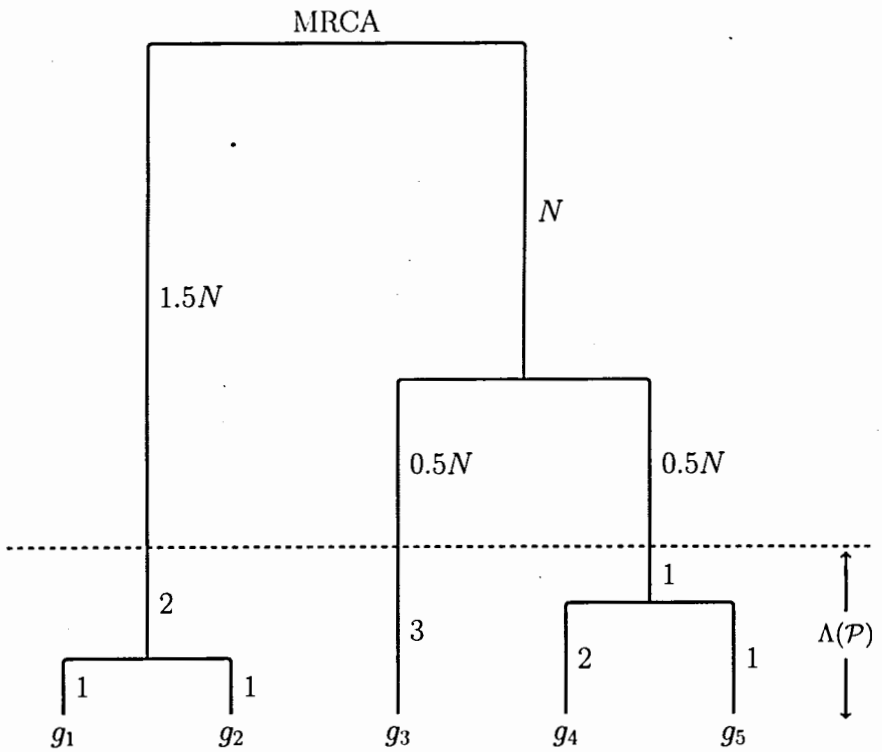


Figure 4.4 : Un arbre représentant la généalogie sous le p-coalescent. En bleu, les arbres minimaux contenant respectivement $\{g_1, g_2\}$ et $\{g_3, g_5\}$.

c'est-à-dire que la distance entre a et l'évènement de coalescence de a et b n'est pas nécessairement égale à la distance entre b et ce même évènement.

Ainsi, comme en génétique familiale, il sera plus facile de décrire la distance $d(a, b)$: la taille de l'arbre minimal contenant la paire. Deux arbres minimaux sur une généalogie du p-coalescent sont illustrés sur la figure 4.4 ; celui reliant g_3 et g_5 n'est pas symétrique à cause de la partie de la généalogie sur le pedigree.

Pour la coalescence de deux gènes a et b sous le p-coalescent, deux situations complémentaires peuvent se produire. La figure 4.5 contient des exemples de distances sous chacune des alternatives.

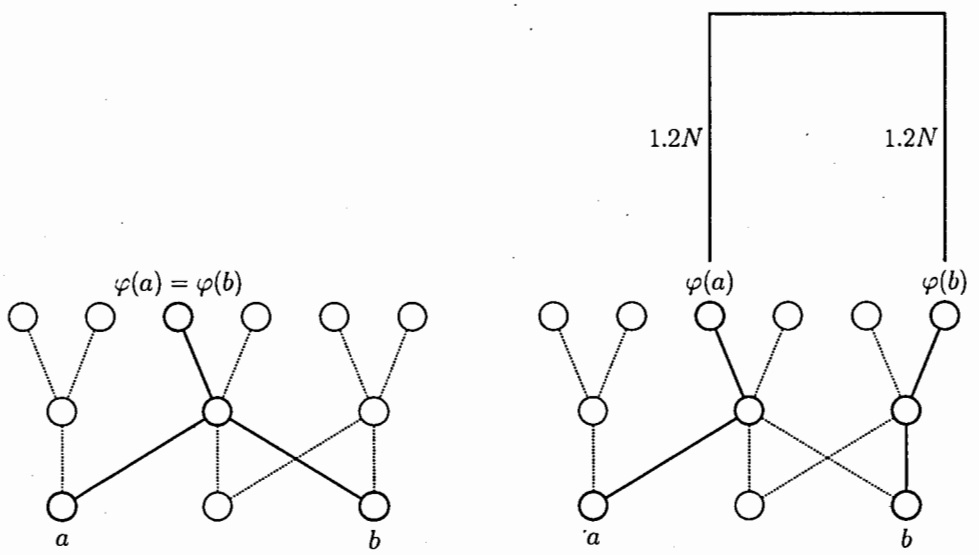


Figure 4.5 : (1) La coalescence se produit sur le pedigree : $d(a,b) = 2$. (2) La coalescence excède le pedigree : $d(a,b) = 2.4N + 4$.

1. Si a et b sont IBD sur le pedigree, c'est-à-dire que $\varphi(a) = \varphi(b)$, l'arbre minimal contenant les deux gènes est un sous graphe du pedigree et leur distance $d(a,b)$ est notée $d_{\mathcal{P}}(a,b)$ comme à la section 3.3.2. Dans ce cas, la distance $\delta(\varphi(a),\varphi(b))$ entre les fondateurs est identiquement zéro ; il s'agit de la distance entre un gène et lui-même.
2. Autrement, si les deux gènes proviennent de fondateurs distincts, l'arbre minimal contenant a et b est complété à partir de $\varphi(a)$ et $\varphi(b)$ dans une population de Wright-Fisher. La taille $d(a,b)$ de cet arbre vaut deux fois le temps de coalescence des fondateurs $2\mathcal{T}_{\varphi(a)\varphi(b)} = \delta(\varphi(a),\varphi(b))$ plus la longueur des chemins sur le pedigree reliant $\varphi(a)$ à a et $\varphi(b)$ à b. Ces longueurs s'expriment en termes de distances sur une généalogie familiale comme $d_{\mathcal{P}}(a,\varphi(a))$ et $d_{\mathcal{P}}(b,\varphi(b))$.

Dans tous les cas, la portion de $d(a,b)$, sous le p-coalescent, provenant du pedigree est toujours bornée par deux fois la hauteur $\Lambda(\mathcal{P})$ du pedigree (voir les inégalités (3.7) et (3.8)) comme le temps total $T_{\text{TOT}}^{(\mathcal{P})}$ en génétique familiale pour un échantillon de taille deux. Intuitivement, la somme des longueurs de deux chemins sur le pedigree est plus petite que deux fois la hauteur de celui-ci par la définition 3.2. Puisque $\delta(\varphi(a),\varphi(b)) = 0$ lorsque $\varphi(a) = \varphi(b)$, la distance $d(a,b)$ est toujours sujette, avec probabilité 1, à l'inégalité

$$\delta(\varphi(a),\varphi(b)) \leqslant d(a,b) \leqslant \delta(\varphi(a),\varphi(b)) + 2\Lambda(\mathcal{P}). \tag{4.2}$$

Pour le p-coalescent, la loi de probabilité de la distance $d(a,b)$ est paramétrée par la probabilité f_{ab} de coalescence sur le pedigree. En mesurant le temps en unité de N générations, $d(a,b)$ converge en loi vers une variable aléatoire exponentielle avec un point de masse à zéro.

Résultat 4.1. *Lorsque $N \to \infty$, la fonction de survie de la distance $d(a,b)$ sous le p-coalescent, mesurée par le temps $t \in \mathbb{R}$ en unité de N reproductions, converge vers*

$$\mathbb{1}_{(-\infty,0)}(t) + \mathbb{1}_{[0,\infty)}(t)[1 - f_{ab}] \exp\left(-\frac{t}{2}\right) \tag{4.3}$$

sauf en 0.

Démonstration. La démonstration est basée sur l'inégalité (4.2). En particulier, elle implique l'inégalité suivante sur les fonctions de survie :

$$P(\delta(a,b) > Nt) \leqslant P(d(a,b) > Nt) \leqslant P(\delta(a,b) + 2\Lambda(\mathcal{P}) > Nt). \tag{4.4}$$

Premièrement, il est à noter que la loi de probabilité de la distance entre les deux fondateurs $\delta(\varphi(a),\varphi(b))$ se comporte déjà comme le mélange (4.3) lorsque N est très grand.

En effet, en envisageant les deux scénarios complémentaires où les gènes coalescent soit sur le pedigree, soit dans la population, la fonction de survie de $\delta(\varphi(a), \varphi(b))$ s'exprime

$$P(\delta(\varphi(a),\varphi(b)) > Nt) = P(\delta(\varphi(a),\varphi(b)) > Nt | \varphi(a) = \varphi(b))\, f_{ab} + P(\delta(\varphi(a),\varphi(b)) > Nt | \varphi(a) \neq \varphi(b))\, [1 - f_{ab}].$$

La distance $\delta(a, b)$ vaut zéro lorsque $a = b$; alors, la fonction de survie sous la première alternative est identiquement zéro à partir de $Nt = 0$:

$$P(\delta(\varphi(a),\varphi(b)) > Nt | \varphi(a) = \varphi(b)) = \mathbb{1}_{(-\infty,0)}(Nt).$$

Lorsque $\varphi(a)$ et $\varphi(b)$ sont distincts, la distance $\delta(\varphi(a), \varphi(b))$ est celle d'un échantillon de taille deux provenant d'une population de Wright-Fisher. Sa fonction de survie est décrite par l'équation (2.1). Ainsi, conditionnellement à $\{\varphi(a) \neq \varphi(b)\}$, la fonction de survie de $\delta(\varphi(a), \varphi(b))$ s'exprime

$$P(\delta(\varphi(a),\varphi(b)) > Nt | \varphi(a) \neq \varphi(b)) = \mathbb{1}_{(-\infty,0)}(Nt) + \mathbb{1}_{[0,\infty)}(Nt) \left(1 - \frac{1}{N}\right)^{\lfloor \frac{Nt}{2} \rfloor}.$$

Globalement, la fonction de survie de $\delta(\varphi(a), \varphi(b))$ avec le temps $t \in \mathbb{R}$ mesuré en unité de N reproductions est

$$\begin{aligned} P(\delta(\varphi(a),\varphi(b)) > Nt) &= \mathbb{1}_{(-\infty,0)}(Nt) + \mathbb{1}_{[0,\infty)}(Nt)[1 - f_{ab}] \left(1 - \frac{1}{N}\right)^{\lfloor \frac{Nt}{2} \rfloor} \\ &= \mathbb{1}_{(-\infty,0)}(t) + \mathbb{1}_{[0,\infty)}(t)[1 - f_{ab}] \left(1 - \frac{1}{N}\right)^{\lfloor \frac{Nt}{2} \rfloor} \end{aligned} \tag{4.5}$$

et converge vers le mélange (4.3) à la limite lorsque $N \to \infty$. La deuxième égalité est obtenue en exprimant les indicatrices comme fonctions de t.

De plus, la distance $d(a, b)$ sous le p-coalescent est au plus une translation par $2\Lambda(\mathcal{P})$ de $\delta(\varphi(a), \varphi(b))$. La fonction de survie de $\delta(\varphi(a), \varphi(b)) + 2\Lambda(\mathcal{P})$ (voir (4.5))

est

$$\mathbb{1}_{(-\infty,0)}(Nt - 2\Lambda(\mathcal{P})) + \mathbb{1}_{[0,\infty)}(Nt - 2\Lambda(\mathcal{P}))[1 - f_{ab}]\left(1 - \frac{1}{N}\right)^{\left\lfloor \frac{Nt}{2} - \Lambda(\mathcal{P})\right\rfloor}.$$

En exprimant les indicatrices comme fonctions de t, elle s'écrit

$$\mathbb{1}_{\left(-\infty,\frac{2\Lambda(\mathcal{P})}{N}\right)}(t) + \mathbb{1}_{\left[\frac{2\Lambda(\mathcal{P})}{N},\infty\right)}(t)[1 - f_{ab}]\left(1 - \frac{1}{N}\right)^{\left\lfloor \frac{Nt}{2} - \Lambda(\mathcal{P})\right\rfloor}.$$

La fonction de survie

$$P(\delta(\varphi(a), \varphi(b)) + 2\Lambda(\mathcal{P}) > Nt)$$

converge partout sauf en $t = 0$, un point de discontinuité, vers le mélange (4.3) à la limite lorsque $N \to \infty$. Par l'inégalité (4.4), $d(a, b)$ converge donc en loi vers une variable exponentielle avec un point de masse f_{ab} à zéro. □

Le résultat 4.1 reflète la différence des échelles de temps. La généalogie sur le pedigree est considérée un évènement de reproduction à la fois. Sous le modèle de Wright-Fisher, la probabilité de coalescence est trop faible dans une grande population ; le temps doit être mesuré en unité de N générations pour construire la généalogie.

De plus, les distances ne dépendent que de la distribution marginale de $R_t^{(\mathbf{S})}$, le processus de coalescence sur les classes d'équivalence de $\mathbf{J}(\mathbf{S})$, avec $d(a, b) = 0$ si $a \overset{\text{IBD}}{=} b$ pour la relation d'équivalence $R_0^{(\mathbf{S})} = \mathbf{J}(\mathbf{S})$; cela se produit avec probabilité f_{ab}. Sinon, la distance $d(a, b)$ est de loi $\mathcal{E}\text{xp}(1/2)$; elle vaut deux fois le temps de coalescence des deux fondateurs représentant les classes d'équivalence distinctes de a et b. La figure 4.6 montre la généalogie de la figure 4.4 avec les temps à l'échelle.

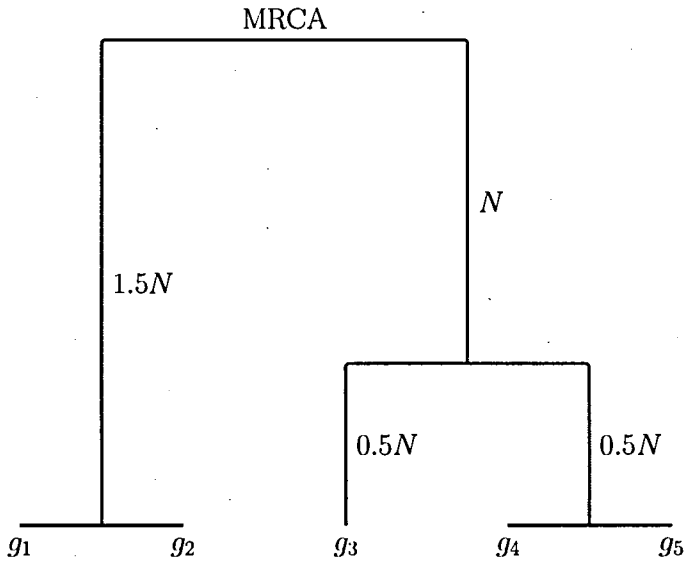


Figure 4.6 : Une généalogie sous le p-coalescent à l'échelle de N évènements de reproduction. Les distances $d(g_1, g_2)$ et $d(g_3, g_5)$ valent respectivement zéro et N.

4.2.2 Temps total

Le temps total T_{TOT} est le nombre d'évènements de reproduction (arêtes sur l'arbre) ou la taille de la généalogie d'un échantillon γ_n sous le p-coalescent. Il est constitué de la somme du temps total de la généalogie sur le pedigree et du temps total des $|\mathbf{J(S)}|$ fondateurs, représentant les classes d'équivalence de l'état IBD, dans la population de Wright-Fisher :

$$T_{\text{TOT}} = T_{\text{TOT}}^{(\mathcal{P})} + \mathcal{T}_{\text{TOT}}^{(|\mathbf{J(S)}|)}.$$

Le temps $T_{\text{TOT}}^{(\mathcal{P})}$ représente la portion de la généalogie sur le pedigree. Il est borné par $n\Lambda(\mathcal{P})$ comme argumenté à la section 3.3.1 ; les temps mesurés sur le pedigree sont négligeables en unité de N reproductions.

La taille $\mathcal{T}_{\text{TOT}}^{(|\mathbf{J(S)}|)}$ de la partie de la généalogie provenant de la population de

Wright-Fisher est de l'ordre de N comme démontré à la section 2.1.5. Elle dépend de la généalogie sur le pedigree, mais seulement à travers l'état IBD. Par l'équation (2.5), le nombre d'arêtes sur la généalogie provenant de la population vaut en moyenne

$$E[\mathcal{T}_{\mathrm{TOT}}^{(|\mathbf{J}(\mathbf{S})|)}] = E[E[\mathcal{T}_{\mathrm{TOT}}^{(|\mathbf{J}(\mathbf{S})|)}|\mathbf{J}(\mathbf{S})]] = E\left[2N \sum_{k=1}^{|\mathbf{J}(\mathbf{S})|-1} \frac{1}{k}\right] + \mathcal{O}(1),$$

où la dernière espérance est sur l'ensemble $\mathbf{S}$ des indicateurs de méiose.

De plus, comme le nombre d'arêtes provenant du pedigree est borné par $n\Lambda(\mathcal{P})$ selon l'équation (3.7), le temps total T_{TOT} sur la généalogie du p-coalescent est sujet à l'inégalité

$$\mathcal{T}_{\mathrm{TOT}}^{(|\mathbf{J}(\mathbf{S})|)} \leqslant T_{\mathrm{TOT}} \leqslant \mathcal{T}_{\mathrm{TOT}}^{(|\mathbf{J}(\mathbf{S})|)} + n\Lambda(\mathcal{P}).$$

Soit T_{TOT}^* le temps mesuré en unité de N générations :

$$T_{\mathrm{TOT}}^* = \frac{T_{\mathrm{TOT}}}{N}.$$

L'espérance de T_{TOT}^*, à la limite lorsque le nombre d'individus dans la population est très grand, est

$$\lim_{N\to\infty} E[T_{\mathrm{TOT}}^*] = E\left[2 \sum_{k=1}^{|\mathbf{J}(\mathbf{S})|-1} \frac{1}{k}\right]. \tag{4.6}$$

L'espérance est sur l'ensemble $\mathbf{S}$ des indicateurs de méiose et, plus particulièrement, sur la variable $\mathbf{J}(\mathbf{S})$; elle peut être approchée par la méthode de Monte-Carlo en utilisant l'algorithme 3.1 de simulation de généalogies sur un pedigree. Une telle procédure est utilisée dans l'appendice A.

Dans le cas d'un échantillon $\gamma_2 = \{a, b\}$ de taille deux, l'espérance vaut

$$2P(|\mathbf{J}(\mathbf{S})| = 2) = 2[1 - f_{ab}], \tag{4.7}$$

ce qui est en accord avec le résultat 4.1 ; le temps total ou la taille de la généalogie d'un échantillon de deux gènes est simplement la distance entre ceux-ci.

En conclusion, les généalogies sous le p-coalescent sont en moyenne plus «petites» que celles décrites par le processus de coalescence : puisque les gènes proviennent d'un pedigree, ils ont une probabilité non nulle de coalescer en un temps négligeable. Cela a des conséquences particulières lorsque vient le temps de mesurer le taux de mutation comme exploré au chapitre 5.

CHAPITRE V

MODÈLE DE MUTATION

Ce chapitre introduit un phénomène important concernant la duplication de gènes : les mutations. Les gènes ne se reproduisent pas toujours de manière exacte et c'est ce qui explique une partie du polymorphisme génétique entre les individus.

Le modèle de mutation proposé est inspiré de la littérature de la génétique des populations ; il possède la caractéristique fondamentale de se produire indépendamment du processus à partir duquel la généalogie est construite. Plusieurs résultats connus en lien avec le modèle de Wright-Fisher sont obtenus à partir d'une démarche originale basée sur la loi des évènements rares.

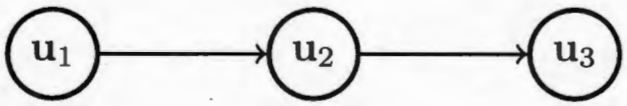
De plus, puisque ce modèle est formulé en termes de généalogies interprétées comme des graphes ou, plus précisément, des arbres, il est généralisable au modèle du p-coalescent décrit au chapitre 4.

5.1 Mutation et généalogie

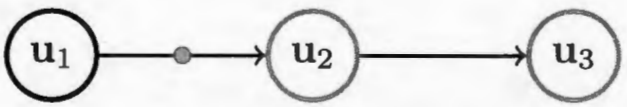
Les notions de duplication génétique et d'hérédité entre les individus ont été discutées au chapitre 1. Entre autres, une représentation sous forme d'arbres enracinés y est discutée. Une arête $[\mathbf{u}, \mathbf{v}]$ ou $\mathbf{u} \to \mathbf{v}$ représente la relation «$\mathbf{v}$ est hérité de $\mathbf{u}$».

En langage courant, une mutation est une erreur dans la duplication d'un gène qui

a pour conséquence d'observer l'enfant et le parent dans des formes distinctes. En l'absence de mutations subséquentes, la nouvelle forme est passée aux descendants du gène mutant par duplication. La figure 5.1 permet de visualiser le phénomène.



(a)



(b)

Figure 5.1 : (a) Aucun évènement de mutation : les gènes sont dupliqués dans une forme identique. (b) Une mutation sur l'arête $[\mathbf{u}_1, \mathbf{u}_2]$ a pour conséquence d'observer $\mathbf{u}_2$ et son descendant $\mathbf{u}_3$ dans une forme distincte de $\mathbf{u}_1$.

La structure globale du graphe représentant les relations d'hérédité est, dans sa forme la plus générale, un ensemble d'arbres enracinés appelé généalogie. Puisque les arêtes représentent les évènements de reproduction, chacune d'entre elles est une occasion indépendante pour un évènement de mutation.

Soit $\mathbf{G} = (\mathbf{V}, \mathbf{A})$ une généalogie fixe (un ensemble d'arbres quelconque), à un sommet $\mathbf{v}$ qui n'est pas une racine correspond un parent $\alpha(\mathbf{v})$ et une arête $\mathbf{a}_\mathbf{v} = [\alpha(\mathbf{v}), \mathbf{v}] \in \mathbf{A}$. Cette arête est porteuse d'une mutation avec probabilité μ indépendamment des autres arêtes. Cela est formalisé par la définition suivante.

Définition 5.1. *Une arête* $\mathbf{a}_\mathbf{v} = [\alpha(\mathbf{v}), \mathbf{v}]$ *sur une généalogie* $\mathbf{G}$ *fixe est porteuse d'une mutation si l'indicateur* $m(\mathbf{a}_\mathbf{v}) = 1$, *où*

$$m(\mathbf{a}_\mathbf{v}) \overset{i.i.d.}{\sim} \mathcal{B}ernoulli(\mu)$$

pour les gènes avec un parent : $\{\mathbf{v} : \alpha(\mathbf{v}) \neq \emptyset\}$.

Lorsque la généalogie $\mathbf{G}$ est considérée comme le résultat d'un processus aléatoire comme dans les modèles décrits aux chapitres 2, 3 et 4, l'ensemble $\mathbf{A}$ des arêtes n'est pas fixé d'avance. Les mutations seront considérées indépendantes de la généalogie dans le sens que lorsque cette dernière est observée, les mutations se produisent sur les arêtes comme indiqué par la définition 5.1.

Supposition 5.1. *Une mutation,* $m(\mathbf{a}) = 1$, *se produit sur une arête* $\mathbf{a}$ *d'une généalogie aléatoire* $\mathbf{G} = (\mathbf{V}, \mathbf{A})$ *indépendamment de celle-ci :*

$$m(\mathbf{a})|\mathbf{G} \overset{i.i.d.}{\sim} \mathcal{B}ernoulli(\mu)$$

pour $\mathbf{a} \in \mathbf{A}$.

Le nombre total d'évènements de mutations conditionnellement à une généalogie $\mathbf{G} = (\mathbf{V}, \mathbf{A})$ est la somme d'indicateurs

$$\mathcal{M}_{\text{TOT}} = \sum_{\mathbf{a} \in \mathbf{A}} m(\mathbf{a}).$$

La variable aléatoire $\mathcal{M}_{\text{TOT}}$ est distribuée, conditionnellement à la généalogie, selon une loi binomiale :

$$\mathcal{M}_{\text{TOT}}|\mathbf{G} \sim \mathcal{B}\text{inomiale}(|\mathbf{A}|, \mu) \tag{5.1}$$

avec $|\mathbf{A}|$ le nombre d'arêtes sur la généalogie $\mathbf{G}$ ou le nombre de gènes ayant un parent.

Pour le modèle de Wright-Fisher, la supposition 5.1 revient à dire que les mutations n'affectent pas (positivement ou négativement) la capacité d'un gène à se reproduire dans la population. Le terme généralement employé est *mutation sélectivement neutre* (Ewens, 1972).

Un contre-exemple serait une mutation létale ; celle-ci empêche complètement le gène mutant de se reproduire. Ce genre de mutation n'est pas indépendant de la généalogie au sens de la supposition 5.1 puisqu'il ne peut se produire sur l'arbre autrement que sur les arêtes qui portent les feuilles car seules celles-ci n'ont pas de descendants.

L'hypothèse d'indépendance permet de placer les mutations sur les arêtes d'une réalisation du modèle de Wright-Fisher en tirant un ensemble de variables aléatoires de Bernoulli. L'effet des mutations dans une population est illustré à la figure 5.2.

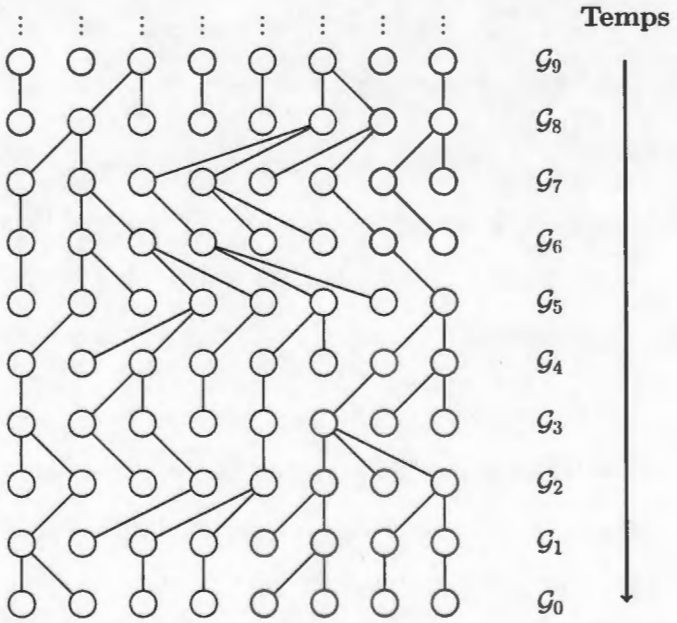


Figure 5.2 : Une mutation entre les générations $\mathcal{G}_6$ et $\mathcal{G}_5$ explique le polymorphisme des gènes de la génération G_0.

Le nombre de mutations sur une généalogie $\mathbf{G}$ fixe d'un échantillon γ_n provenant de la génération $\mathcal{G}_0$ d'une population de Wright-Fisher est distribué (équation (5.1)) selon une loi binomiale. Le nombre d'«essais» est le nombre d'arêtes

sur la généalogie ou le temps total $\mathcal{T}_{\text{TOT}}^{(n)}$ et la probabilité de «succès» est μ.

Globalement, le nombre moyen d'évènements de mutation sur la généalogie de γ_n sous le modèle de Wright-Fisher peut être trouvé en utilisant l'espérance itérée et l'équation (2.5) :

$$E[\mathcal{M}_{\text{TOT}}^{(n)}] = E[E[\mathcal{M}_{\text{TOT}}^{(n)}|\mathbf{G}]] = E[\mu\mathcal{T}_{\text{TOT}}^{(n)}] = \mu\left[2N\sum_{k=1}^{n-1}\frac{1}{k} + \mathcal{O}(1)\right]. \qquad (5.2)$$

La taille moyenne de la généalogie augmente avec le nombre N d'individus dans la population et, ainsi, le nombre moyen de mutations augmente. Par exemple, en vertu de (5.2), l'espérance du nombre de mutations pour un échantillon de taille $n = 2$ est de l'ordre de $2N\mu$. Intuitivement, dans une grande population, la probabilité de pouvoir distinguer une paire de gènes choisie au hasard serait proche de un.

Dans les faits, la probabilité de mutation μ par reproduction est plutôt faible et les échantillons de gènes ne présentent pas autant de polymorphisme même dans les grandes populations. Pour cela, il est usuel de considérer une probabilité de mutation de l'ordre de $1/N$.

Dans les mots de Kingman (1982b), cette dernière hypothèse est nécessaire pour atteindre un équilibre entre mutation et dérive. Pour simplifier, si la probabilité de mutation est trop faible, la population a tendance à être de plus en plus uniforme (pas de polymorphisme). Au contraire, s'il est trop élevé, les membres de la population ont tendance à tous être dans une forme génétique distincte.

Le taux de mutation θ est défini comme

$$\theta = 2N\mu \quad \Leftrightarrow \quad \mu = \frac{\theta}{2N}. \qquad (5.3)$$

Ce paramètre est constant par rapport au nombre N d'individus dans la population. La constante 2 est arbitraire ; elle permet d'interpréter θ comme le nombre

moyen de mutation sur la généalogie d'un échantillon de deux gènes lorsque le nombre N d'individus dans la population est très grand :

$$\lim_{N\to\infty} E[\mathcal{M}_{\mathrm{TOT}}^{(2)}] = \theta.$$

Cette dernière propriété est obtenue en vertu de l'équation (5.2) en exprimant la probabilité de mutation μ en fonction du taux de mutation θ.

Le paramètre θ peut être mesuré à partir des différences entre les membres d'un échantillon de gènes de la génération présente comme il sera discuté plus en détails à la section 5.3.

5.2 Sites polymorphes

Dans la suite, il sera considéré, au départ, que les gènes sont composés de κ sites, chacun se trouvant dans un état qui est identifiable. Le caractère indivisible des gènes est toutefois respecté; chacun des sites provient du même parent lors du processus de reproduction (définition 1.4). Il s'agit d'une représentation abstraite (Watterson, 1975) d'un cistron, une partie de la molécule d'ADN composée des bases nucléiques A, C ,T ou G.

Un évènement de mutation se produit sur une arête $\mathbf{a}_g = [\alpha(g), g]$ lorsque l'état d'un certain nombre de sites $\mathcal{M}_g > 0$ est modifié. Le gène parent et son enfant diffèrent en $\mathcal{M}_g$ sites; ils possèdent, entre les deux, $\mathcal{M}_g$ sites polymorphes. Un exemple d'évènement de mutation est illustré à la figure 5.3. En contrepartie, en l'absence d'évènements de mutation $\mathcal{M}_g$ vaut 0; l'état d'aucun site est modifié et les deux gènes sont identiques. Il est supposé que chacun des κ sites mute indépendamment des autres selon un processus de Bernoulli.

Définition 5.2. *Pour un gène g avec un parent $\alpha(g)$ sur la généalogie, un site $k \in \{1, 2, \ldots, \kappa\}$ est dans un état différent de celui correspondant sur son parent*

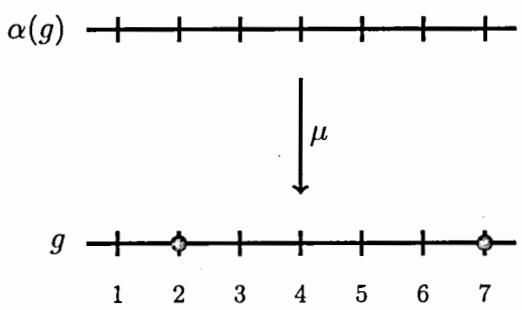


Figure 5.3 : Exemple d'une mutation causée par un changement d'état de $\mathcal{M}_g = 2$ sites parmi $\kappa = 7$.

$\alpha(g)$ *si* $m_k(\mathbf{a}_g) = 1$, *où*

$$m_k(\mathbf{a}_g) \overset{i.i.d.}{\sim} \mathcal{B}ernoulli(\nu), \quad k = 1, 2, \ldots, \kappa.$$

Le nombre de sites affectés est la somme de κ variable aléatoires de Bernoulli :

$$\mathcal{M}_g = \sum_{k=1}^{\kappa} m_k(\mathbf{a}_g)$$

Ainsi, M_g est distribué selon une loi $\mathcal{B}$inomiale(κ, ν). Puisque la probabilité d'observer aucune mutation est $1 - \mu$, les paramètres μ et ν sont liés à travers la relation suivante :

$$P(\mathcal{M}_g = 0) = (1 - \nu)^{\kappa} = 1 - \mu \qquad \Leftrightarrow \qquad \nu = 1 - (1 - \mu)^{1/\kappa}. \tag{5.4}$$

Il sera considéré que les gènes sont composés d'un grand nombre de sites ayant une faible probabilité de mutation. En gardant la probabilité μ d'observer au moins un site mutant fixe par rapport au nombre κ de sites, la loi des évènements rares permet de modéliser le nombre $\mathcal{M}_g$ de mutations comme une variable aléatoire de Poisson.

5.2.1 Modèle des sites infiniment nombreux

Le modèle des sites infiniment nombreux a été étudié, entre autres, par Watterson (1975) dans le cadre du modèle de Wright-Fisher. Il sera dérivé dans cette section à partir du processus de Bernoulli introduit à la définition 5.2 et la loi des évènements rares.

En effet, il est possible d'étudier le comportement du nombre $\mathcal{M}_g$ de sites mutants lorsque le nombre κ de sites tend vers l'infini. La probabilité ν de mutation par site décroit selon la contrainte (5.4) ; c'est la probabilité μ d'observer au moins un site mutant qui reste fixe. Premièrement, le lemme suivant sur une forme indéterminée servira à la démonstration du résultat 5.1

Lemme 5.1. *La limite lorsque* $n \to \infty$ *de la suite*

$$n(1-(1-x)^{1/n}), \tag{5.5}$$

pour $x \in (0,1)$, *est une forme indéterminée de type* $0/0$ *qui converge vers*

$$-\log(1-x).$$

Démonstration. D'abord, la forme indéterminée est manifeste lorsque l'équation (5.5) est exprimée comme le quotient

$$\frac{1-(1-x)^{1/n}}{1/n}.$$

Ensuite, il suffit de calculer les dérivées par rapport à n du numérateur et du dénominateur et d'appliquer la règle de l'Hôpital.

La dérivée du numérateur est

$$\frac{\mathrm{d}}{\mathrm{d}n}\left\{1-(1-x)^{1/n}\right\} = \frac{(1-x)^{1/n}\log(1-x)}{n^2}.$$

Pour le dénominateur,

$$\frac{\mathrm{d}}{\mathrm{d}n}\left\{\frac{1}{n}\right\} = -\frac{1}{n^2}.$$

Par la règle de l'Hôpital, la limite de (5.5) est la limite du quotient des dérivées :

$$\lim_{n\to\infty} n(1-(1-x)^{1/n}) = \lim_{n\to\infty} -(1-x)^{1/n}\log(1-x) = -\log(1-x).$$

□

Il est à noter que la variable aléatoire $\mathcal{M}_g$ suit une loi binomiale; sa loi limite lorsque $\kappa \to \infty$ est Poisson. Il s'agit d'un cas particulier de la loi des évènements rares démontré au résultat suivant.

Résultat 5.1. *En faisant tendre κ vers l'infini tout en contractant la probabilité de mutation par site selon l'équation* (5.4)*, la loi limite de $\mathcal{M}_g$ est*

$$\mathcal{P}oisson(-\log(1-\mu)).$$

Démonstration. La variable aléatoire $\mathcal{M}_g$ suit une loi $\mathcal{B}$inomiale(κ,ν) ; elle possède la fonction de masse

$$\begin{aligned} P(\mathcal{M}_g = m) &= \binom{\kappa}{m}\nu^m(1-\nu)^{\kappa-m} \\ &= \binom{\kappa}{m}(1-(1-\mu)^{1/\kappa})^m((1-\mu)^{1/\kappa})^{\kappa-m} \\ &= \frac{\kappa!}{(\kappa-m)!m!}(1-(1-\mu)^{1/\kappa})^m(1-\mu)^{1-m/\kappa}. \end{aligned}$$

De plus, le factoriel descendant est défini comme

$$(\kappa)_m = \frac{\kappa!}{(\kappa-m)!} = \prod_{j=0}^{m-1}(\kappa-j).$$

Ainsi, la fonction de masse de $\mathcal{M}_g$ peut donc être exprimée comme

$$P(\mathcal{M}_g = m) = \frac{(1-\mu)^{1-m/\kappa}}{m!}\prod_{j=0}^{m-1}(\kappa-j)(1-(1-\mu)^{1/\kappa}). \tag{5.6}$$

La limite lorsque κ tend vers l'infini de la première partie de l'équation (5.6) est

$$\lim_{\kappa\to\infty} \frac{(1-\mu)^{1-m/\kappa}}{m!} = \frac{1-\mu}{m!} = \frac{\exp(\log(1-\mu))}{m!}.$$

La limite de chacun des facteurs du produit en (5.6) vaut

$$\lim_{\kappa\to\infty} (\kappa - j)(1-(1-\mu)^{1/\kappa}) = -\log(1-\mu)$$

par le lemme 5.1.

Globalement, la fonction de masse de $\mathcal{M}_g$ converge vers celle d'une

$$\mathcal{P}\text{oisson}(-\log(1-\mu))$$

à la limite lorsque $\kappa \to \infty$:

$$\lim_{\kappa\to\infty} P(\mathcal{M}_g = m) = \frac{\exp(\log(1-\mu))(-\log(1-\mu))^m}{m!}.$$

□

La probabilité d'observer aucun site mutant sous le modèle de Poisson est toujours égale à $1-\mu$. Le résultat 5.1 est valide pour une valeur μ quelconque. Cependant, la probabilité μ de mutation est généralement considérée faible (de l'ordre de $1/N$) comme discuté à la section 5.1. Dans ce cas, le nombre moyen de sites mutants est $-\log(1-\mu) \approx \mu$.

5.2.2 Position des sites mutants

L'objectif de cette section est d'énoncer quelques résultats importants concernant la position des sites mutants. Les mutations sur les sites sont, à la base, modélisées par un processus de Bernoulli (définition 5.2) : les mutations se produisent sur chacun des sites indépendamment des autres sites et avec probabilité constante.

Intuitivement, sachant qu'il y a, par exemple, un seul nouveau site mutant, sa position s est donnée avec probabilité uniforme sur $\{1, 2, \ldots, \kappa\}$.

À partir du processus de Bernoulli modélisant les sites mutants sur un gène (définition 5.2) peut être défini un processus (binomial) à temps discret

$$\{\mathcal{N}_g(s), s = 0, 1, \ldots, \kappa\}$$

tel que

$$\mathcal{N}_g(s) = \sum_{k=1}^{s} m_k(\mathbf{a}_g), \quad m_k(\mathbf{a}_g) \overset{\text{i.i.d.}}{\sim} \mathcal{B}\text{ernoulli}(\nu).$$

Le processus $\mathcal{N}_g(s)$ modélise, pour un évènement de reproduction représenté par l'arête $[\alpha(g), g]$, le nombre de mutations parmi les s premiers sites pour $s \leqslant \kappa$, où κ est le nombre total de sites. Le processus démarre à zéro et prend ses valeurs parmi les entiers $\{0, 1, \ldots, \kappa\}$. Il est markovien en s : une transition de i vers $i+1$ se produit avec probabilité ν, la probabilité de mutation par site; autrement, il reste à i. Sa matrice de transition est donc décrite par

$$\mathbf{p}_{ij} = \begin{cases} 1 - \nu & j = i, \\ \nu & j = i + 1, \\ 0 & \text{sinon.} \end{cases}$$

La probabilité de transition d'un état i vers j en ℓ pas de $\mathcal{N}_g(s)$ sera notée $\mathbf{p}_{ij}(\ell)$. Il s'agit de la probabilité suivante pour un certain $s \leqslant \kappa - \ell$:

$$\begin{aligned} \mathbf{p}_{ij}(\ell) &= P(\mathcal{N}_g(s + \ell) = j | \mathcal{N}_g(s) = i) \\ &= P\left(\sum_{k=s+1}^{s+\ell} m_k(\mathbf{a}_g) = j - i \right). \end{aligned}$$

Les valeurs de j vers lesquelles une transition en ℓ pas en partant de i est possible sont les entiers $i, i + 1, \ldots, i + \ell$. Puisque les variables $m_k(\mathbf{a}_g)$ sont indépendantes

et de loi $\mathcal{B}\text{ernoulli}(\nu)$, la probabilité $\mathbf{p}_{ij}(\ell)$ s'exprime comme une loi binomiale :

$$\mathbf{p}_{ij}(\ell) = \binom{\ell}{j-i} \nu^{j-i}(1-\nu)^{\ell-j+i}.$$

Il est à noter que la probabilité ν de mutation par site est en relation avec le nombre de sites κ à travers l'équation (5.4). La probabilité de transition en ℓ pas s'exprime en fonction de κ et μ comme

$$\mathbf{p}_{ij}(\ell;\kappa) = \binom{\ell}{j-i} (1-(1-\mu)^{1/\kappa})^{j-i}((1-\mu)^{1/\kappa})^{\ell-j+i}.$$

Le processus $\mathcal{N}_g(s)$ peut être approximé par un processus de Poisson (spatial) sur l'intervalle $[0,1]$ lorsque le nombre de sites κ est assez grand, tel que décrit par le résultat suivant.

Résultat 5.2. *Lorsque le nombre de sites $\kappa \to \infty$, les probabilités de transition d'un état i vers un état $j \geqslant i$ du processus $\mathcal{N}_g(\lfloor \kappa x \rfloor)$, pour un point $x \in [0,1]$ mesuré en unité de κ sites, convergent vers celles d'un processus de Poisson homogène d'intensité $-\log(1-\mu)$:*

$$\mathbf{p}_{ij}(\lfloor \kappa x \rfloor;\kappa) \underset{\kappa\to\infty}{\longrightarrow} \frac{\exp(\log(1-\mu)x)(-\log(1-\mu)x)^{j-i}}{(j-i)!}.$$

La preuve est omise; il s'agit d'une généralisation directe de la démarche pour la démonstration du résultat 5.1 en remplaçant m par $j-i$ et κ par $\lfloor \kappa x \rfloor$. Le résultat 5.2 n'implique pas que les $\mathbf{p}_{ij}(\lfloor \kappa x \rfloor;\kappa)$ convergent uniformément pour tout $i, j \in \mathbb{N}$ tels que $i \leqslant j$.

La classe de processus de Bernoulli *rétrécissants* («shrinking») décrite dans Gallager (2011, section 2.2.5) présente certaines similarités avec le mécanisme de mutation à temps discret introduit à la définition 5.2 et sujet à la contrainte (5.4). Le résultat 5.2 est inspiré du théorème 2.2.4 et du corollaire 2.2.1 de la section citée, mais la démarche est originale.

Dans la suite, il sera supposé que, pour un évènement de reproduction, les mutations apparaissent sur le gène-enfant, représenté par l'intervalle $[0,1]$, selon un processus de Poisson $\{\tilde{\mathcal{N}}_g(x), x \in [0,1]\}$ homogène d'intensité $-\log(1-\mu)$. Ce nouveau modèle rejoint la littérature tel qu'expliqué plus bas.

Une mutation sur l'intervalle $[0,1]$ correspond à un saut de 1 pour le processus $\tilde{\mathcal{N}}_g(x)$. Sans entrer dans les détails, sachant qu'il est apparu $\mathcal{M}_g$ sites mutants lors de la reproduction $\alpha(g) \rightarrow g$ ou, de manière équivalente, $\tilde{\mathcal{N}}_g(1) = \mathcal{M}_g$, les positions des $\mathcal{M}_g$ nouveaux sites mutants sont des variables aléatoires indépendantes et identiquement distribuées de loi $\mathcal{U}[0,1]$. La position conditionnelle des sauts pour un processus de Poisson homogène est discutée de manière extensive dans Daley et Vere-Jones (1988, section 2.1).

Intuitivement, la probabilité qu'un même point de l'intervalle $[0,1]$ mute lors de deux évènements de reproduction indépendants est négligeable. Il s'agit de la propriété de disjonction des processus de Poisson (Kingman, 1992, section 2.2). Cela permet de supposer, dans la suite, qu'une nouvelle mutation se produit toujours sur un site qui n'était précédemment pas affecté, ce qui est l'essence du concept des site infiniment nombreux.

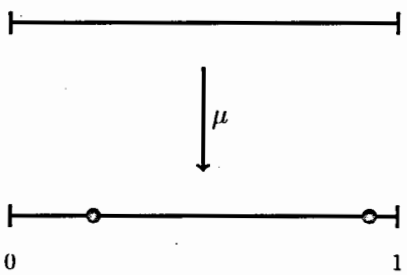


Figure 5.4 : Les mutations apparaissent en une reproduction selon un processus de Poisson sur l'intervalle $[0,1]$.

5.2.3 Nombre de sites polymorphes

Le nombre de nouveaux sites polymorphes par duplication suit approximativement une loi de Poisson comme discuté à la section 5.2.1. Plus spécifiquement, le mécanisme de mutation est modélisé par un processus de Poisson d'intensité $-\log(1-\mu)$ sur l'intervalle $[0,1]$ (résultat 5.2). L'objectif de cette section est d'étudier le phénomène sur une généalogie, plus particulièrement lorsque celle-ci est mesurée en unité de N (le nombre d'individus dans la population) reproductions en supposant un taux de mutation $\theta = 2N\mu$.

Premièrement, considérons une chaîne d'évènements de reproduction comme celle présentée à la figure 5.5. Lorsqu'une mutation apparaît sur un site, elle est passée à tous les gènes descendants.

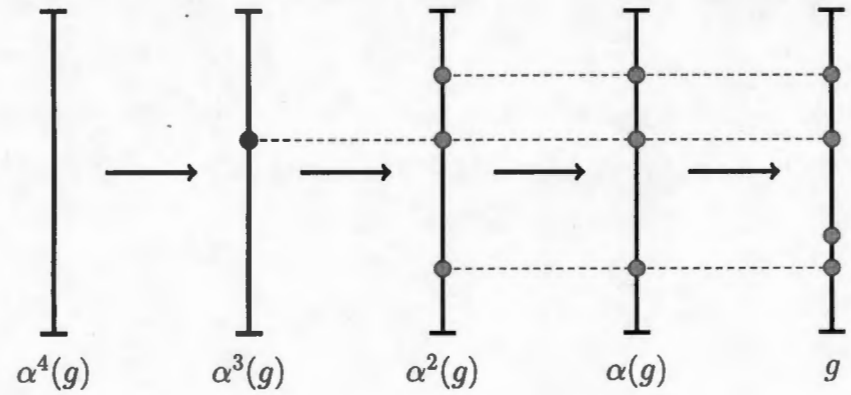


Figure 5.5 : Le nombre de sites polymorphes entre deux gènes séparés par $j = 4$ reproductions est la somme de 4 variables aléatoires de Poisson indépendantes.

Les sites polymorphes entre deux gènes séparés par j évènements de reproduction se comportent comme la superposition de j processus de Poisson homogènes (voir sur la figure 5.5). Le résultat est un processus de Poisson d'intensité $-j\log(1-\mu)$ sur $[0,1]$. Ceci est une propriété fondamentale des processus ponctuels de Poisson

qui est démontrée, entre autres, dans Kingman (1992, section 2.2).

Le nombre $\mathcal{M}_g^{(j)}$ de sites polymorphes entre deux gènes séparés par j reproductions est donc une variable aléatoire de loi de Poisson :

$$\mathcal{M}_g^{(j)} \sim \mathcal{P}\text{oisson}(-j\log(1-\mu)), \quad j \in \mathbb{N}. \tag{5.7}$$

Lorsque le temps est mesuré en unité de N reproductions et que la probabilité de mutation est à l'échelle de $1/N$, le résultat limite suivant est obtenu.

Résultat 5.3. *Soit $\mu = \theta/2N$ la probabilité de mutation exprimée en fonction du taux de mutation θ donné en (5.3). Le nombre de sites polymorphes entre deux gènes séparés par un chemin de longueur $t \geqslant 0$ mesuré en unité de N reproductions se comporte, lorsque $N \to \infty$, comme une variable aléatoire de Poisson de moyenne $\theta t/2$.*

Démonstration. En vertu de l'équation (5.7), le nombre de sites polymorphes sur l'intervalle $[0,1]$ entre un gène et son descendant $\lfloor Nt \rfloor$ reproductions plus tard est une variable aléatoire de loi de Poisson de moyenne

$$-\lfloor Nt \rfloor \log\left(1 - \frac{\theta}{2N}\right).$$

La fonction $-\log(1-x)$ peut être exprimée en tant que série de Taylor développée autour de 0 :

$$-\log(1-x) = x + \frac{1}{2}x^2 + \frac{1}{3}x^3 + \ldots$$

En utilisant un développement limité,

$$-\lfloor Nt \rfloor \log\left(1 - \frac{\theta}{2N}\right) = \lfloor Nt \rfloor \left[\frac{\theta}{2N} + \mathcal{O}\left(\frac{1}{N^2}\right)\right] \underset{N\to\infty}{\longrightarrow} \frac{\theta t}{2},$$

pour tout $t \geqslant 0$. □

Pour un ensemble γ_n de n gènes, le nombre K de sites polymorphes est défini comme le nombre de positions sur l'intervalle $[0, 1]$ pour lesquelles au moins deux des gènes diffèrent.

Premièrement, puisque le nombre de sites κ est très grand, il est supposé que, par la propriété de disjonction des processus de Poisson, les mutations se produisent toujours à des points distincts de l'intervalle $[0, 1]$. Ensuite, toutes les mutations qui précèdent le plus récent ancêtre commun (en mesurant le temps du présent vers le passé) sur la généalogie ne sont pas héritées par l'ensemble de l'échantillon. Les mutations sur une généalogie sont illustrées à la figure 5.6.

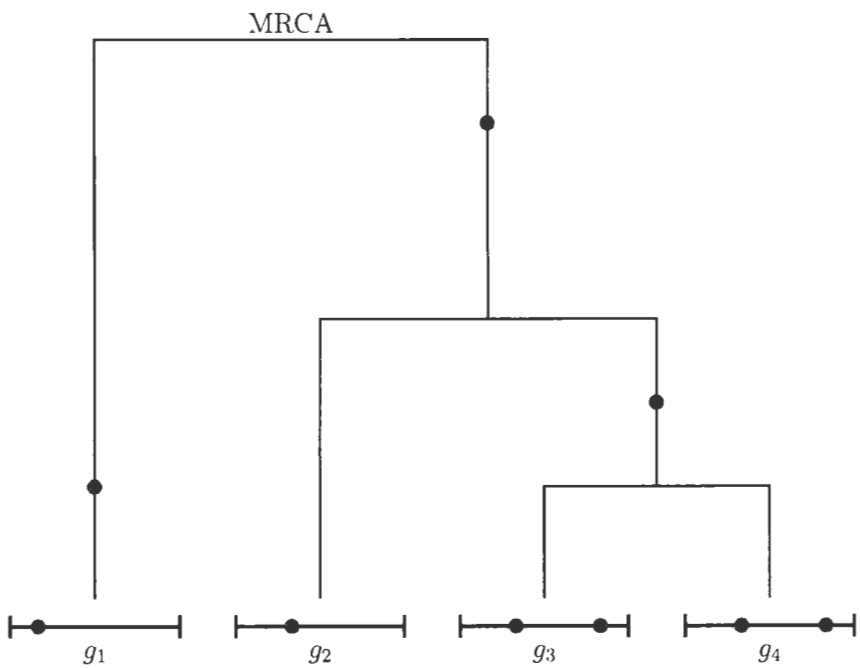


Figure 5.6 : Les évènements de mutation sur la généalogie expliquent le nombre $K = 3$ de sites polymorphes observés dans un échantillon de $n = 4$ gènes.

Donc, à chacune des mutations qui se produit sur la généalogie de γ_n jusqu'à leur plus récent ancêtre commun correspond un site polymorphe de l'échantillon. Le

nombre K de sites polymorphes pour un échantillon γ_n est ainsi la somme des sites mutants sur chacune des branches de la généalogie. Pour une généalogie $\mathbf{G}$ fixe, il s'agit d'une variable aléatoire $\mathcal{P}\text{oisson}\,(\theta T_{\text{TOT}}/2)$ avec T_{TOT}, la taille de $\mathbf{G}$ mesurée en unité de N reproductions.

5.3 Mesures du taux de mutation

Le taux de mutation θ peut être mesuré à partir d'un échantillon γ_n de gènes selon le nombre de sites polymorphes. Il sera supposé que l'état de chacun des sites sur $[0, 1]$ peut être comparé entre tous les individus.

En pratique, si les données sont composées d'une grande quantité de bases nucléiques identifiées par leurs formes A, C, T et G, il est généralement accepté que le modèle des sites infiniment nombreux est adéquat.

Intuitivement, si le taux de mutation θ est élevé, le nombre K de sites polymorphes d'un échantillon devrait augmenter en conséquence. Des estimateurs de θ peuvent être élaborés par la méthode des moments pour le modèle basé sur le processus de coalescence et le p-coalescent décrit au chapitre 4.

5.3.1 Processus de coalescence

Comme présenté à la section 5.2.3, lorsque la généalogie $\mathbf{G}$ est observée, le nombre de sites polymorphes d'un échantillon γ_n est une variable aléatoire de loi

$$\mathcal{P}\text{oisson}\left(\frac{\theta T_{\text{TOT}}}{2}\right)$$

avec T_{TOT}, la taille de la généalogie en unité de N reproductions. La généalogie d'un échantillon provenant d'une population se reproduisant selon le modèle de Wright-Fisher peut être approximée par le processus de coalescence lorsque le

nombre N d'individus dans la population est très grand comme démontré au chapitre 2.

L'espérance de la taille de la généalogie d'un échantillon de taille n mesurée en unité de N reproductions sous le processus de coalescence est dérivée à la section 2.2.1. L'espérance de K peut ainsi être obtenue en itérant sur la généalogie :

$$E[K] = E[E[K|\mathbf{G}]] = E\left[\frac{\theta T_{\text{TOT}}}{2}\right] = \theta \sum_{k=1}^{n-1} \frac{1}{k}. \tag{5.8}$$

L'équation (5.8) est en accord avec les développements successifs de Watterson (1975) et Ewens (1974) qui ont étudié le comportement du nombre K de sites polymorphes en utilisant des processus de diffusion pour approximer le modèle de Wright-Fisher à une époque qui précède le développement du processus de coalescence tel que décrit par Kingman (1982a).

Cela suggère que le nombre K de sites polymorphes peut servir à estimer θ. En effet, un estimateur basé sur K est donné par la méthode des moments :

$$\tilde{\theta} = \frac{K}{\sum_{k=1}^{n-1} \frac{1}{k}}.$$

Dans le cas $n = 2$, l'estimateur vaut K, le nombre de sites polymorphes entre les deux gènes.

Une autre approche (Tajima, 1983) pour l'estimation de θ à partir d'un échantillon de taille n est basée sur le nombre de sites polymorphes pour chacune des paires $\{a, b\} \subset \gamma_n$. Le nombre de sites polymorphes ou le nombre de différences par paire pour un couple de gènes $\{a, b\}$ est noté k_{ab}.

L'espérance marginale de k_{ab} est θ ; il s'agit du nombre de sites polymorphes entre $n = 2$ gènes (voir (5.8)). Pour un échantillon γ_n, la moyenne de ces différences par

paire s'exprime

$$\bar{k} = \frac{1}{\binom{n}{2}} \sum_{\{a,b\} \subset \gamma_n} k_{ab}.$$

L'espérance de la moyenne $\bar{k}$ vaut également θ :

$$E[\bar{k}] = \frac{1}{\binom{n}{2}} \sum_{\{a,b\} \subset \gamma_n} E\left[k_{ab}\right] = \theta.$$

Donc, $\bar{k}$ est un autre estimateur de θ, noté $\hat{\theta}$, obtenu par la méthode des moments.

Les estimateurs $\tilde{\theta}$ et $\hat{\theta}$ basés respectivement sur K et $\bar{k}$ sont présentés ici afin de les comparer avec leurs analogues sous le paradigme du p-coalescent. Une grande littérature existe sur les propriétés de ces estimateurs. Par exemple, Watterson (1975) dérive la loi de probabilité du nombre K de sites polymorphes parmi un échantillon de n gènes sous le modèle de Wright-Fisher.

5.3.2 *P*-coalescent

La taille moyenne de la généalogie sous le p-coalescent est plus petite que celle dérivée pour le processus de coalescence ; à cause des liens de parenté sur le pedigree, certains gènes peuvent coalescer en un temps négligeable. Le taux de mutation a donc tendance à être estimé plus élevé lorsque les méthodes de la section 5.3.1 sont utilisées.

Par exemple, en conditionnant d'abord sur la généalogie $\mathbf{G}$, la moyenne K du nombre de sites polymorphes d'un échantillon de n gènes sous le p-coalescent s'exprime

$$E[K] = E[E[K|\mathbf{G}]] = E\left[\frac{\theta T^*_{\mathrm{TOT}}}{2}\right]$$

avec T^*_{TOT}, la taille de la généalogie mesurée en unité de N reproductions.

Par l'équation (4.6), lorsque $N \to \infty$, $E[K]$ peut s'écrire comme une espérance

sur $\mathbf{S}$, l'ensemble des indicateurs de méiose :

$$E[K] = \theta E\left[\sum_{k=1}^{|\mathbf{J}(\mathbf{S})|-1} \frac{1}{k}\right]. \tag{5.9}$$

L'estimateur de θ basé sur K sous le p-coalescent est

$$\tilde{\theta}^* = \frac{K}{E\left[\sum_{k=1}^{|\mathbf{J}(\mathbf{S})|-1} \frac{1}{k}\right]}.$$

L'espérance au dénominateur est complexe à calculer pour les cas $n > 2$, mais peut être approchée par une méthode de Monte-Carlo à l'aide de l'algorithme 3.1 de simulation de généalogies sur des pedigrees. Une telle procédure est illustrée dans les simulations de l'appendice A.

Pour un échantillon de taille n, la cardinalité $|\mathbf{J}(\mathbf{S})|$ de l'état IBD est toujours inférieure ou égale à n. Ainsi, le taux de mutation a toujours un estimé plus grand lorsque les relations de parenté sur le pedigree sont prises en compte : $\tilde{\theta}^* \geqslant \tilde{\theta}$.

Pour l'estimateur basé sur les différences de paires, l'espérance marginale de k_{ab} pour une paire de gènes est le cas particulier $n = 2$ de l'équation (5.9) :

$$E[k_{ab}] = \theta P(\mathbf{J}(\mathbf{S}) = 2) = \theta[1 - f_{ab}].$$

L'espérance de la moyenne $\bar{k}$ des différence de paires sous le p-coalescent est donc

$$E[\bar{k}] = \frac{1}{\binom{n}{2}} \sum_{\{a,b\} \subset \gamma_n} E\left[k_{ab}\right] = \frac{\theta}{\binom{n}{2}} \sum_{\{a,b\} \subset \gamma_n} [1 - f_{ab}].$$

Alors, l'estimateur de θ basé sur $\bar{k}$ prend la forme

$$\hat{\theta}^* = \frac{\binom{n}{2}\bar{k}}{\sum_{\{a,b\} \subset \gamma_n}[1 - f_{ab}]}.$$

Il est toujours plus grand ou égal à $\hat{\theta}$, l'estimateur analogue basé exclusivement sur le processus de coalescence, car

$$\sum_{\{a,b\} \subset \gamma_n} [1 - f_{ab}] \leqslant \binom{n}{2}.$$

L'algorithme 3.2 permet de calculer les valeurs de f_{ab} pour toutes les paires $\{a, b\} \subset \gamma_n$.

Les estimateurs $\tilde{\theta}^*$ et $\hat{\theta}^*$ décrivent comment sont affectées les estimations basées sur le nombre de sites polymorphes dans un modèle qui tient compte des relations familiales entre les membres de l'échantillon. La différence est plus manifeste lorsque les relations familiales sont fortes; par exemple, lorsque les probabilités f_{ab} de coalescence sur le pedigree sont élevées.

CONCLUSION

L'objectif du projet de recherche qui a conduit à ce mémoire était de développer un nouveau cadre théorique permettant d'intégrer les principes de la génétique familiale dans un modèle de génétique des populations, par exemple le modèle de Wright-Fisher.

Premièrement, une représentation formelle du phénomène d'hérédité est décrite au chapitre 1 en termes de graphes dirigés. Les notions de base qui y sont énoncées ont permis de présenter les chapitres subséquents dans un même langage mathématique.

Ensuite, le modèle de Wright-Fisher a été présenté au chapitre 2. Ce modèle de population a été choisi pour sa simplicité et parce qu'il a été étudié en profondeur et depuis longtemps par plusieurs mathématiciens. Le processus de coalescence est aussi étudié au chapitre 2 ; il permet de penser le modèle de Wright-Fisher en termes de généalogies évoluant du présent vers le passé.

Au chapitre 3, les principes fondamentaux de la génétique familiale sont présentés avec une approche originale inspirée du processus de coalescence. En effet, la construction de généalogies sur un pedigree est la question d'intérêt dans ce chapitre.

Tout cela mène vers le cœur du mémoire : le chapitre 4 où le modèle hybride, le p-coalescent, est présenté. L'effet des relations familiales sur la taille d'une généalogie y est décrit en détails en considérant différentes échelles de temps pour la génétique familiale et le modèle de Wright-Fisher.

Finalement, le chapitre 5 présente une approche originale pour construire un modèle de mutation à partir de graphes. L'objectif est de démontrer en quoi le modèle est équivalent à ce qui se retrouve dans la littérature concernant le modèle de Wright-Fisher et d'illustrer certaines particularités lorsqu'il est appliqué à un modèle tenant compte des relations familiales.

Le développement de cette nouvelle approche, s'il est d'intérêt, est à un stade encore très jeune. Un phénomène lié à la génétique des organismes diploïdes reste à être développé : les recombinaisons. L'approche de Hudson (1983) basée sur le processus de coalescence pourrait possiblement être généralisée au p-coalescent.

Ensuite, une théorie statistique pour les estimateurs de maximum de vraisemblance sous le p-coalescent constituerait une suite logique. Il est présentement difficile, mais intriguant, de voir comment adapter une démarche comme celle de Griffiths et Marjoram (1996) par exemple.

Pour finir, le p-coalescent a été développé en ayant toujours à l'esprit des possibles applications en génétique médicale, par exemple la cartographie génétique. Cependant, l'absence de précédent a conduit à développer en premier le formalisme du modèle. Ce modèle a le potentiel d'être adapté pour être appliqué en recherche médicale dans une certaine forme étant donné que les données génomiques provenant d'individus reliés par des pedigrees sont abondantes.

APPENDICE A

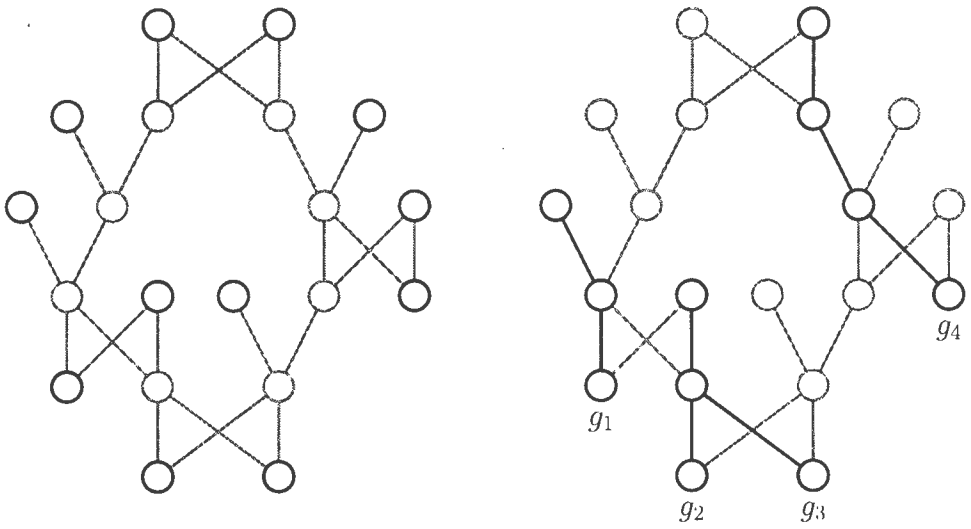
EXPÉRIENCES SUR LE CONCEPT DE GÉNÉALOGIE EN GÉNÉTIQUE FAMILIALE

L'objectif de cet appendice est d'illustrer certaines notions développées au chapitre 3 sur la génétique familiale et de mettre en évidence l'utilité de l'algorithme 3.1. Cet algorithme simule des réalisations aléatoires de généalogies pour échantillon de gènes provenant d'un pedigree.

Une série d'expériences est présentée afin de développer une certaine intuition pour le concept de généalogie en génétique familiale (section 3.2). Le pedigree qui servira d'exemple est introduit à la figure A.1 avec une réalisation possible de la généalogie.

D'abord, l'avantage de l'algorithme 3.1 est qu'il génère des généalogies sans nécessairement simuler l'entièreté de l'ensemble $\mathbf{S}$ des indicateurs de méiose. Le nombre total d'indicateurs de méiose sur un pedigree est $|\mathcal{D}|$: un pour chacun des non-fondateurs. La nature de l'algorithme 3.1 fait en sorte que l'ensemble, noté $\mathbf{S}^{(r)}$, des indicateurs de méiose qui sont simulés lors d'une réalisation est aléatoire. Un paramètre qui sera évalué à chacune des expériences est le nombre moyen $E[|\mathbf{S}^{(r)}|]$ d'indicateurs simulés.

Ensuite, la cardinalité moyenne $E[|\mathbf{J}(\mathbf{S})|]$ de l'état IBD ou le nombre de fondateurs



(a) (b)

Figure A.1 : (a) Un pedigree composé de $|\mathcal{I}| = 20$ gènes : $|\mathcal{F}| = 8$ fondateurs (en bleu) et $|\mathcal{D}| = 12$ non-fondateurs. (b) Une réalisation possible (en rouge) d'une généalogie pour un échantillon de $n = 4$ gènes. L'état IBD sur cet exemple est la partition $\mathbf{J}(\mathbf{S}) = \{\{g_1\}, \{g_2, g_3\}, \{g_4\}\}$; $|\mathbf{J}(\mathbf{S})| = 3$.

qui sont des ancêtres de l'échantillon sur la généalogie sera comparé avec le nombre n de gènes dans l'échantillon. L'état IBD (définition 3.4) est une partition de l'échantillon ; sa cardinalité est toujours inférieure ou égale au nombre de gènes dans l'échantillon.

Finalement, un concept en lien avec le p-coalescent (chapitre 4) est étudié, soit la taille moyenne $E[T^*_{\text{TOT}}]$ de la généalogie pour l'échantillon considéré, mesurée en unité de N reproductions.

Si les relations familiales sont ignorées, la généalogie d'un échantillon de n gènes est modélisée par le processus de coalescence. La taille moyenne $E[T_{\text{TOT}}]$ de la généalogie, mesurée en unité de N reproductions, sous le processus de coalescence

s'exprime comme deux fois une somme harmonique partielle (équation (2.6)) :

$$E[T_{\mathrm{TOT}}] = 2\sum_{k=1}^{n-1}\frac{1}{k}.$$

Pour le p-coalescent, la taille moyenne $E[T^*_{\mathrm{TOT}}]$ de la généalogie est donnée par l'espérance suivante (équation (4.6)) sur l'ensemble $\mathbf{S}$ des indicateurs de méiose :

$$E[T^*_{\mathrm{TOT}}] = E\left[2\sum_{k=1}^{|\mathbf{J}(\mathbf{S})|-1}\frac{1}{k}\right].$$

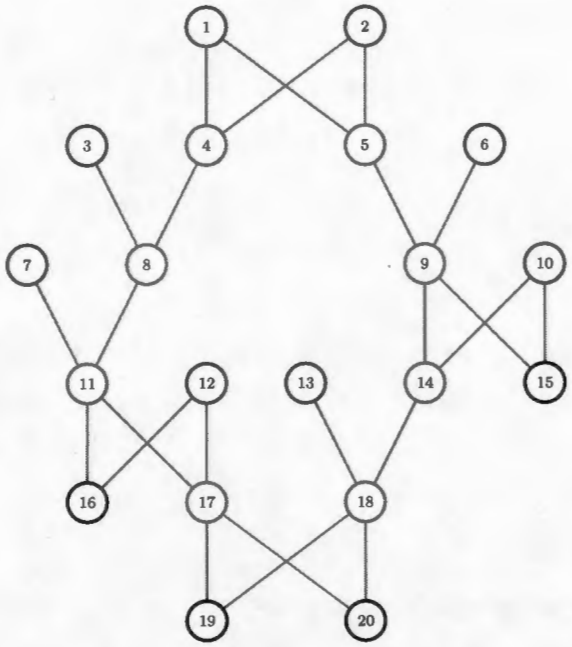
Puisque $P(|\mathbf{J}(\mathbf{S})| \leqslant n) = 1$, la taille moyenne de la généalogie sous le p-coalescent est toujours inférieure ou égale à celle dérivée sous le processus de coalescence :

$$E[T^*_{\mathrm{TOT}}] \leqslant E[T_{\mathrm{TOT}}].$$

Pour chacune des six expériences présentées, un total de 100 000 généalogies sont simulées à l'aide de l'algorithme 3.1 et la méthode de Monte-Carlo est appliquée pour obtenir des estimés des espérances considérées.

Expérience 1 : L'échantillon considéré est

$$\gamma_4 = \{15, 16, 19, 20\}.$$



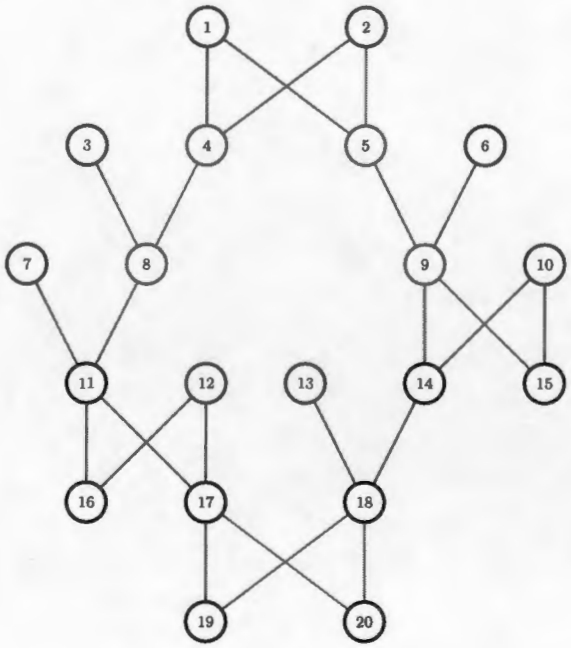
Le tableau des résultats montre que ces 4 gènes représentent approximativement 3 lignées parmi les fondateurs. De plus, environ le tiers des indicateurs de méiose ne sont pas simulés en moyenne.

Tableau A.1 : Résultats de l'expérience 1

$\lvert\mathcal{D}\rvert$	$=$ 12	$>$	$E[\lvert\mathbf{S}^{(r)}\rvert]$	$\approx$	7.96
n	$=$ 4	$>$	$E[\lvert\mathbf{J}(\mathbf{S})\rvert]$	$\approx$	2.91
$E[T_{\text{TOT}}]$	$=$ 3.67	$>$	$E[T^*_{\text{TOT}}]$	$\approx$	2.85

Expérience 2 : L'échantillon considéré est

$$\gamma_8 = \{11, 14, 15, 16, 17, 18, 19, 20\}.$$



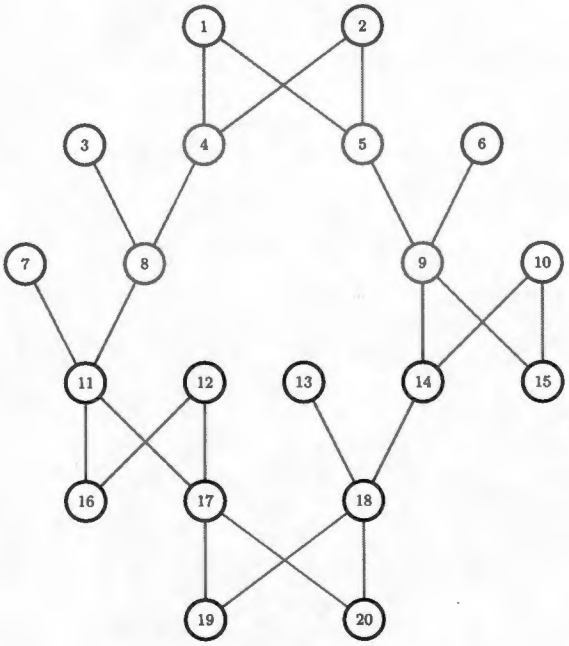
Par rapport à l'expérience précédente, 4 gènes ont été ajoutés à l'échantillon ; alors, le nombre moyen d'indicateurs de méiose simulés a augmenté. Il est à noter que les 8 gènes ont des liens de parenté forts : ils représentent en moyenne moins de 4 lignées parmi les fondateurs.

Tableau A.2 : Résultats de l'expérience 2

$\lvert\mathcal{D}\rvert$	$=$ 12	$>$	$E[\lvert\mathbf{S}^{(r)}\rvert]$	$\approx$	9.88
n	$=$ 8	$>$	$E[\lvert\mathbf{J}(\mathbf{S})\rvert]$	$\approx$	3.70
$E[T_{\text{TOT}}]$	$=$ 5.19	$>$	$E[T^*_{\text{TOT}}]$	$\approx$	3.41

Expérience 3 : L'échantillon considéré est

$$\gamma_{10} = \{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}.$$



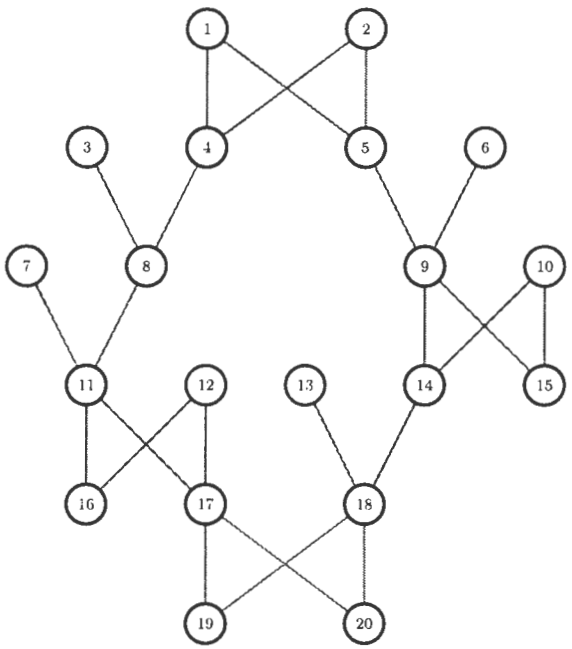
Par rapport à l'expérience précédente, les gènes 12 et 13 ont été ajoutés à l'échantillon. Puisque ces gènes sont des fondateurs, le nombre moyen d'indicateurs de méiose simulés est resté le même. Cependant, la cardinalité moyenne de l'état IBD a augmenté considérablement car les gènes 12 et 13 sont leur propre «ancêtre» parmi les fondateurs.

Tableau A.3 : Résultats de l'expérience 3

$\lvert\mathcal{D}\rvert$	$=$ 12	$>$	$E[\lvert\mathbf{S}^{(r)}\rvert]$	$\approx$	9.88
n	$=$ 10	$>$	$E[\lvert\mathbf{J}(\mathbf{S})\rvert]$	$\approx$	4.45
$E[T_{\text{TOT}}]$	$=$ 5.66	$>$	$E[T^*_{\text{TOT}}]$	$\approx$	3.89

Expérience 4 : L'échantillon considéré est

$$\gamma_{12} = \mathcal{D} = \{4, 5, 8, 9, 11, 14, 15, 16, 17, 18, 19, 20\}.$$



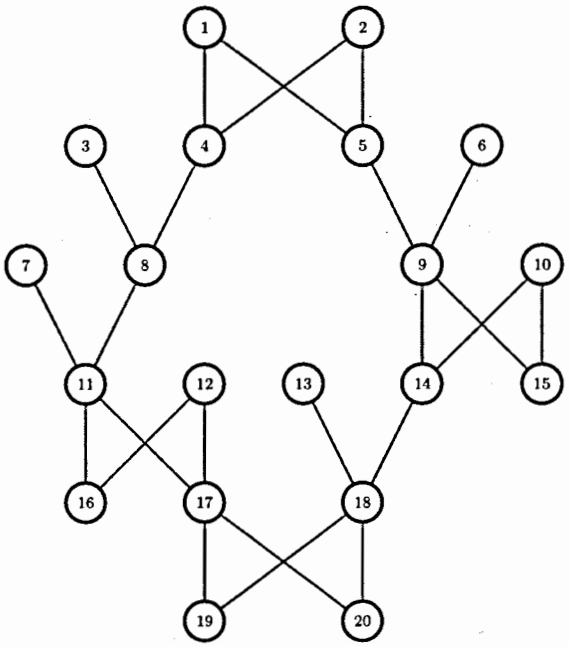
L'échantillon est composé de l'ensemble $\mathcal{D}$ de tous les non-fondateurs sur le pedigree. Sous cette configuration, tous les indicateurs de méiose ont besoin d'être simulés à chaque fois : chacun des non-fondateurs doit choisir son parent sur la généalogie.

Tableau A.4 : Résultats de l'expérience 4

$\lvert\mathcal{D}\rvert$	$=$ 12	$=$	$E[\lvert\mathbf{S}^{(r)}\rvert]$	$=$	12
n	$=$ 12	$>$	$E[\lvert\mathbf{J}(\mathbf{S})\rvert]$	$\approx$	5.00
$E[T_{\mathrm{TOT}}]$	$=$ 6.04	$\searrow$	$E[T^*_{\mathrm{TOT}}]$	$\approx$	4.07

Expérience 5 : L'échantillon considéré est

$$\gamma_{20} = \mathcal{I} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}.$$



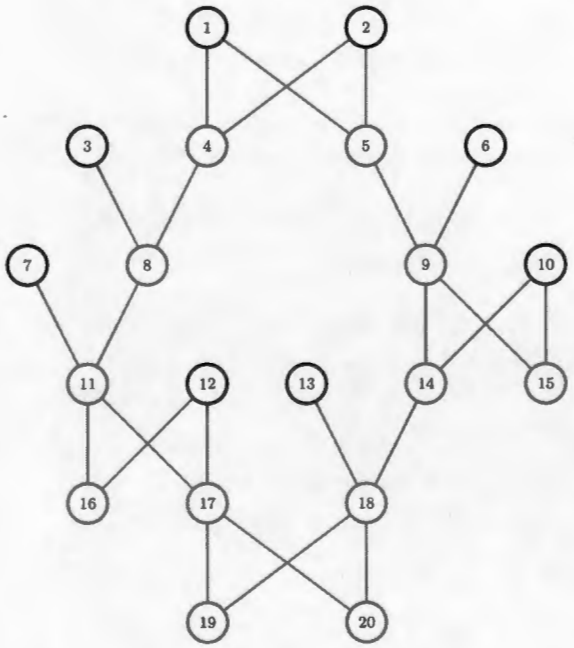
L'échantillon est composé de l'ensemble $\mathcal{I}$ de tous les gènes sur le pedigree. L'ensemble des indicateurs de méiose doit encore une fois être simulé au complet puisque tous les non-fondateurs font partie de l'échantillon. Comme tous les fondateurs sont également dans l'échantillon, la cardinalité de l'état IBD reste fixe à 8. Par conséquent, la taille moyenne de la généalogie sous le p-coalescent correspond exactement à celle d'un échantillon de 8 gènes sous le processus de coalescence.

Tableau A.5 : Résultats de l'expérience 5

$\lvert\mathcal{D}\rvert$	$=$	12	$=$	$E[\lvert\mathbf{S}^{(r)}\rvert]$	$=$ 12
n	$=$	20	$>$	$E[\lvert\mathbf{J}(\mathbf{S})\rvert]$	$=$ 8
$E[T_{\text{TOT}}]$	$=$	7.10	$>$	$E[T^*_{\text{TOT}}]$	$=$ 5.19

Expérience 6 : L'échantillon considéré est

$$\gamma_8 = \mathcal{F} = \{1, 2, 3, 6, 7, 10, 12, 13\}.$$



Cette fois, l'échantillon est constitué de l'ensemble $\mathcal{F}$ des fondateurs. Dans ce cas, aucun indicateur de méiose n'a besoin d'être simulé. De plus, la cardinalité de l'état IBD est égale au nombre de gènes dans l'échantillon et les tailles moyennes des généalogies sont égales sous les deux modèles ; en n'échantillonnant que des fondateurs, c'est le processus de coalescence exclusivement qui construit la généalogie.

Tableau A.6 : Résultats de l'expérience 6

$\lvert\mathcal{D}\rvert$	$= 12$	$>$	$E[\lvert\mathbf{S}^{(r)}\rvert]$	$= 0$
n	$= 8$	$=$	$E[\lvert\mathbf{J}(\mathbf{S})\rvert]$	$= 8$
$E[T_{\text{TOT}}]$	$= 5.19$	$=$	$E[T^*_{\text{TOT}}]$	$= 5.19$

BIBLIOGRAPHIE

Balding, D. J., Bishop, M. et Cannings, C. (2008). *Handbook of statistical genetics.* John Wiley & Sons.

Bang-Jensen, J. et Gutin, G. Z. (2008). *Digraphs : theory, algorithms and applications.* Springer Science & Business Media.

Bondy, J. A. et Murty, U. S. R. (1976). *Graph theory with applications.* Citeseer.

Campbell, N., Reece, J. et Taylor, M. (2009). *Biology : Concepts & Connections.* Pearson/Benjamin Cummings.

Chang, J. T. (1999). Recent common ancestors of all present-day individuals. *Advances in Applied Probability, 31*(4), 1002–1026.

Daley, D. et Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes.* Springer-Verlag.

Darwin, C. (1883). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* Appleton.

Dawkins, R. (1976). *The Selfish Gene.* Oxford University Press, USA.

Descary, M.-H. (2012). *DMAP : une nouvelle méthode de cartographie génétique fine adaptée à des modèles génétiques complexes.* Université du Québec à Montréal.

Durrett, R. (2012). *Essentials of Stochastic Processes.* Springer Texts in Statistics. Springer New York.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology, 3*(1), 87–112.

Ewens, W. J. (1974). A note on the sampling theory for infinite alleles and infinite sites models. *Theoretical population biology, 6*(2), 143–148.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection* (1 éd.). Clarendon Press.

Forest, M. (2010). *Cartographie genetique fine simultanee de deux genes.* Université du Québec à Montréal.

Gallager, R. (2011). *Discrete Stochastic Processes.* MIT OpenCourseWare.

Griffiths, R. C. et Marjoram, P. (1996). Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology, 3*(4), 479–502.

Grimmett, G. et Stirzaker, D. (2001). *Probability and random processes.* Oxford university press.

Hein, J., Schierup, M. et Wiuf, C. (2004). *Gene genealogies, variation and evolution : a primer in coalescent theory.* Oxford University Press, USA.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology, 23*(2), 183–201.

Johnson, N. L., Kemp, A. W. et Kotz, S. (2005). *Univariate discrete distributions* (3 éd.). John Wiley & Sons.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic processes and their applications, 13*(3), 235–248.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of applied probability, 19*(A), 27–43.

Kingman, J. F. C. (1992). *Poisson Processes.* Oxford Studies in Probability. Clarendon Press.

Koller, D. et Friedman, N. (2009). *Probabilistic graphical models : principles and techniques.* MIT press.

Lange, K. (2003). *Mathematical and statistical methods for genetic analysis.* Springer Science & Business Media.

Larribe, F., Lessard, S. et Schork, N. J. (2002). Gene mapping via the ancestral recombination graph. *Theoretical population biology, 62*(2), 215–229.

Last, G. et Penrose, M. (2017). *Lectures on the Poisson process.* Cambridge University Press.

Lauritzen, S. L. et Sheehan, N. A. (2003). Graphical models for genetic analyses. *Statistical Science, 18*(4), 489–514.

Malécot, G. (1948). *Les mathématiques de l'hérédité.* Paris : Maison et Cie.

Mendel, G. (1865). Experiments in plant hybridization. *Verhandlungen des naturforschenden Vereins Brünn.*

Möhle, M. (1998). Coalescent results for two-sex population models. *Advances in Applied Probability, 30*(2), 513–520.

Ross, S. (1997). *Simulation.* Academic Press.

Ross, S. M. (1996). *Stochastic processes.* Wiley New York.

Speed, D. et Balding, D. J. (2015). Relatedness in the post-genomic era : is it still useful ? *Nature Reviews Genetics, 16*(1), 33.

Tajima, F. (1983). Evolutionary relationship of dna sequences in finite populations. *Genetics, 105*(2), 437–460.

Tavaré, S. (2004). Ancestral inference in population genetics. In *Lectures on probability theory and statistics* 1–188. Springer.

Thompson, E. A. (2000). *Statistical inference from genetic data on pedigrees.* IMS.

Thompson, E. A. (2013). Identity by descent : variation in meiosis, across genomes, and in populations. *Genetics, 194*(2), 301–326.

Wakeley, J. (2009). *Coalescent theory.* Roberts & Company.

Wakeley, J., King, L., Low, B. S. et Ramachandran, S. (2012). Gene genealogies within a fixed pedigree, and the robustness of kingmans coalescent. *Genetics, 190*(4), 1433–1445.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology, 7*(2), 256–276.

Wright, S. (1931). Evolution in mendelian populations. *Genetics, 16*(2), 97–159.