

Université de Montréal

**APPLICATIONS DU PROCESSUS
ANCESTRAL AVEC RECOMBINAISON ET
CONVERSION EN GÉNÉTIQUE
STATISTIQUE**

par

LAMIAE SAIDI

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en STATISTIQUE

janvier 2014

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

**APPLICATIONS DU PROCESSUS
ANCESTRAL AVEC RECOMBINAISON ET
CONVERSION EN GÉNÉTIQUE
STATISTIQUE**

présenté par

LAMIAE SAIDI

a été évalué par un jury composé des personnes suivantes :

Murua, Alejandro

(président-rapporteur)

Lessard, Sabin

(directeur de recherche)

Labbe, Aurélie - Université McGill

(membre du jury)

Mémoire accepté le:

20 décembre 2013

RÉSUMÉ

Le processus ancestral est appliqué pour étudier la variabilité génétique et la mesure de déséquilibre de liaison de séquences d'ADN, et faire de l'inférence statistique sur les divers facteurs responsables de cette variabilité. En tenant compte, en premier lieu, des facteurs de dérive génétique, de mutation, et de recombinaison, les calculs exacts de la mesure de déséquilibre de liaison de deux loci sont retrouvés. De plus, une approximation du processus exact, SMC (sequentially Markov chain), est utilisée pour trouver la mesure d'association à deux loci, et une formule de covariance pour calculer cette mesure est corrigée. En intégrant le facteur de conversion dans le modèle de Moran, on trouve l'espérance des mesures de polymorphisme exprimées par les espérances des mesures de variation intra-locus et inter-locus. Celles-ci sont calculées à l'aide de temps espérés dans les états ancestraux. De plus, l'espérance du déséquilibre de liaison est trouvée et il est montré qu'elle diminue quand le taux de recombinaison augmente. En utilisant ces résultats théoriques, on présente une méthode pour estimer les paramètres de mutation, de recombinaison, et de conversion.

Mots clés : coalescent ; conversion ; déséquilibre de liaison ; modèle de Moran ; processus ancestral ; recombinaison ; variabilité génétique.

ABSTRACT

The ancestral process is applied to investigate the amount of DNA variation and the amount of linkage disequilibrium; it is also applied to make statistical inference about the multiple factors responsible for this variation. Considering genetic drift, mutation, and recombination events, the exact solutions for linkage disequilibrium between two loci are obtained. Furthermore, the association measure between two loci is obtained by using an approximation of the exact process, SMC (sequentially Markov chain), and correcting a covariance formula. After introducing intrachromosomal gene conversion under the Moran model, the expected amounts of variation within and between two loci are obtained using expected times spent in the ancestral states. Furthermore, the expectation of linkage disequilibrium is obtained and it is shown to decrease as the recombination rate is increased. Using these theoretical results, a method for estimating the mutation, recombination and gene conversion parameters is presented.

Keywords : Ancestral process ; coalescent ; conversion ; genetic variability ; linkage disequilibrium ; Moran model ; recombination.

TABLE DES MATIÈRES

Résumé	iii
Abstract	iv
Liste des figures	viii
Liste des tableaux	xi
Remerciements	xii
Introduction	1
Chapitre 1. Outils préliminaires de génétique statistique et rappels de modèles mathématiques d'évolution	4
1.1. Notions de génétique.....	4
1.1.1. Hérité.....	4
1.1.2. Allèle, locus et génotype	5
1.2. Équilibre de Hardy-Weinberg	6
1.2.1. Fréquence génotypique et fréquence allélique	6
1.2.2. Équilibre de Hardy-Weinberg	6
1.2.3. Déséquilibre de liaison	8
1.3. Facteurs d'évolution	11
1.3.1. Mutation	11
Modèle à une infinité de sites.....	12
Mesure de polymorphisme.....	12
1.3.2. Recombinaison	13

1.3.3.	Conversion.....	13
1.3.4.	Dérive génétique.....	14
1.4.	Modèles de populations.....	16
1.4.1.	Modèle de Wright-Fisher.....	17
	Hétérozygotie et dérive génétique.....	18
1.4.2.	Modèle de Moran.....	20
	Hétérozygotie et dérive génétique.....	21
Chapitre 2.	Processus ancestral.....	23
2.1.	Le coalescent de Kingman.....	23
2.2.	Processus ancestral sous le modèle de Wright-Fisher.....	25
	Coalescence de deux lignées.....	25
	Coalescence de deux lignées parmi n	26
	Approximation en temps continu.....	27
	Survenance d'un événement de mutation.....	27
	Survenance d'un événement de recombinaison.....	28
	Survenance d'un événement de conversion.....	29
	Taux de changement du matériel ancestral.....	30
2.3.	Processus ancestral sous le modèle de Moran.....	30
	Processus ancestral exact.....	30
	Approximation du processus ancestral exact.....	32
Chapitre 3.	Mesure de liaison.....	35
3.1.	Covariances exactes de temps de coalescence à deux loci avec recombinaison.....	35
3.2.	Covariances avec l'approximation du processus de coalescence avec recombinaison de McVean et Cardin.....	41

Chapitre 4. Mesure de polymorphisme à deux loci avec recombinaison et conversion	51
4.1. Changements de fréquences sous le modèle de Moran à deux loci ..	52
4.2. Variabilité génétique intra-locus et inter-locus	61
4.3. Estimation de paramètres	63
4.4. Calculs pour le modèle à une infinité de sites	64
Conclusion	79
Bibliographie	81
Annexe A. Preuves détaillées de quelques résultats	A-i
A.1. Mesure de polymorphisme de séquences d'ADN - π	A-i
A.2. Mesure de liaison en termes de covariances de temps de coalescence	A-ii
A.3. Processus de diffusion	A-vii

LISTE DES FIGURES

1.1	Génotypes et haplotypes.	5
1.2	Cinq séquences d'ADN à 4 sites polymorphes.....	13
1.3	Modèle de réparation de cassures double-brin pour l'enjambement lors de la méiose, par jonctions doubles de Holliday (Double-strand break repair model for meiotic cross-over via double Holliday junctions).....	15
2.1	Matériel ancestral d'un échantillon de taille $n = 4$ dans une population de taille $2N = 6$ qui suit le modèle de Moran.	34
3.1	Diagramme de transitions d'état pour le processus de coalescence avec recombinaison. Les transitions possibles entre les 9 états ancestraux de deux séquences à deux loci sont indiquées par des flèches associées aux probabilités de transition.....	40
3.2	Les cinq types de séquences ancestrales du SMC.	42
3.3	Diagramme de transitions d'état pour le processus de coalescence avec recombinaison SMC. Les transitions possibles entre les 9 états ancestraux de deux séquences échantillonnées considérées à deux loci sont indiquées par des flèches associées aux probabilités de transition.	44
3.4	Diagramme de transitions d'état pour le processus de coalescence avec recombinaison SMC. Les transitions possibles entre les 20 états ancestraux de trois séquences échantillonnées, une considérée au locus (l), une considérée au locus (m) et une considérée au deux loci, sont indiquées par des flèches associées aux probabilités de transition.	46

- 3.5 Diagramme de transitions d'état pour le processus de coalescence avec recombinaison SMC. Les transitions possibles entre les 17 états ancestraux de quatre séquences échantillonnées dont deux considérées au locus (l) et deux autres considérées au locus (m) sont indiquées par des flèches associées aux probabilités de transition. 48
- 3.6 Décroissance du déséquilibre de liaison en fonction de la distance génétique approximée par σ_d^2 sous le processus ancestral exact ($\sigma_{d,exact}^2$), sous l'approximation SMC du processus ancestral calculée dans l'article de McVean et Cardin ($\sigma_{d,McVean}^2$), et sous l'approximation SMC corrigée ($\sigma_{d,SMC}^2$). 49
- 4.1 Séquences résultantes d'un événement de mutation ou de conversion. Les événements sont indiqués par des flèches associées aux probabilités. 54
- 4.2 Chromosome i , avec $i = 1, \dots, n$ à deux gènes avec L sites chacun tel que chaque site $m = 1, \dots, L$ peut avoir deux types d'allèles, A et a . . 65
- 4.3 Événement établissant le lien entre l'hétérozygotie et la mutation. . . . 68
- 4.4 Diagramme de transitions d'état pour le processus de coalescence avec recombinaison et conversion à partir de deux lignées, une de type A et une autre de type a . Les transitions possibles entre les 5 états sont indiquées par des flèches associées aux probabilités de transition. . . . 71
- 4.5 Diagramme de transitions d'état pour le processus de coalescence avec recombinaison et conversion à partir de quatre lignées, deux de types A et deux de types a . Les transitions possibles du système sont indiquées par des flèches associées aux probabilités de transition. 75
- 4.6 Diagramme de transitions d'état pour le processus de coalescence avec recombinaison et conversion à partir de trois lignées, deux du même type et un de l'autre type. Les transitions possibles du système sont indiquées par des flèches associées aux probabilités de transition. . . . 76

A.1	Événements de mutation qui définissent l'espérance conditionnelle (A.2.1).	A-iii
A.2	Les statistiques de la généalogie au locus (A) de deux séquences i et j prises dans un échantillon.	A-vi

LISTE DES TABLEAUX

1.1	Fréquences de génotypes dans une population de taille $2N=294$	6
1.2	Effets de la ségrégation des lois de Mendel.	7
3.1	Etats ancestraux et transitions d'état à partir de deux séquences à deux loci (l) et (m).	37
4.1	Effets de la ségrégation et de la recombinaison.	53
4.2	Effets de la mutation et de la conversion.	54
4.3	Etats ancestraux et transitions d'état pour deux lignées, une de type A et une autre de type a	69
4.4	Etats ancestraux et transitions d'état pour quatre lignées, deux de type A et deux de type a	77
4.5	Etats ancestraux et transitions d'état pour trois lignées dont exactement deux de même type A ou a	78

REMERCIEMENTS

Tout d'abord, je tiens à remercier mon directeur de recherche M. Sabin Lessard qui a été ma boussole tout au long de mon parcours et qui le sera sûrement après ma maîtrise.

Je remercie également Mme Aurélie Labbe pour sa générosité et l'excellence de sa pédagogie dans son cours d'introduction à la bio-statistique qui m'a permis d'établir des liens entre mon sujet de recherche et la mise en pratique de la théorie.

Je dédie ce mémoire à mon frère Yussef qui m'a accompagnée pendant mon séjour. Ta présence a sans aucun doute contribué à faire de cette expérience une des plus amusantes et des plus paisibles que j'ai vécues. "I love so much about the things that you choose to be."

INTRODUCTION

La génétique quantitative et la génétique des populations sont des domaines étroitement liés qui ont pour objectif d'étudier les variations phénotypiques des individus d'une population. La génétique des populations traite de fréquences d'allèles et de génotypes, alors que la génétique quantitative se concentre autour de la contribution de la variation du milieu et des génotypes à la variation du phénotype. Plus largement, la génétique quantitative aborde des problèmes liés à l'hérédité de caractères complexes et constitue une extension à plusieurs locus de la génétique des populations. Le projet de recherche de ce mémoire se situe en génétique des populations et couvre des méthodes d'inférence de paramètres génétiques à partir d'hypothèses récentes en biologie moléculaire, notamment l'évolution de séquences complètes d'ADN en prenant en compte les mécanismes de transmission génétique.

Le processus de coalescence est une approche développée par Kingman (1982a,b,c) pour étudier les facteurs d'évolution d'une population en remontant le temps. Le coalescent de Kingman (1982a,b,c) est la limite du processus ancestral d'une vaste classe de structures de populations qui inclut les modèles de Wright-Fisher et de Moran. Hudson (1983) a étendu le processus de coalescence pour tenir compte de la recombinaison intragénique. Les covariances exactes entre les temps de coalescence à deux loci à partir de séquences échantillonnées sont obtenues à partir des résultats de Griffiths (1981 ; voir aussi Pluzhnikov et Donnelly 1996). Dans le but d'étudier la variabilité génétique, on obtient ces covariances pour des paires de lignées à deux loci et un échantillon de deux, trois ou quatre séquences, en considérant l'extension du processus de coalescence en présence de recombinaison, mais en absence de conversion, telle que décrite par Hudson (1983).

L'analyse de Nagylaki et Petes (1982) a révélé que le processus de conversion génétique intrachromosomique produit une éventuelle homogénéité d'une séquence parmi une famille de gènes répétés en tandem. Dans le but d'estimer les paramètres de conversion, de mutation et de recombinaison dans les familles multigéniques, Innan (2002a,b) a calculé des espérances et des variances de mesures de variation inter-locus et intra-locus pour deux copies de gènes, par une approximation par un processus diffusion du modèle de Wright-Fisher.

Dans ce mémoire, nous rappelons d'abord quelques notions de génétique statistique (chapitre 1). Puis nous décrivons le coalescent de Kingman (1982a). Ensuite, nous incorporons les résultats qui découlent de l'extension de Hudson (1983) du processus de coalescence avec recombinaison en y ajoutant la conversion pour deux modèles de populations, les modèles de Wright-Fisher et de Moran (chapitre 2).

Par la suite, nous voyons de façon détaillée la mise en pratique du processus ancestral exact sans conversion, en calculant les covariances de temps de coalescence qui interviennent dans une mesure du déséquilibre de liaison (Pluzhnikov et Donnelly 1996). La première nouvelle contribution du mémoire consiste à présenter les détails des calculs de covariances de temps de coalescence sous l'approximation du processus ancestral SMC (sequentially Markov Chain) qui sont omis dans l'article de McVean et Cardin (2005), et à corriger un résultat intermédiaire de ce dernier. Le SMC est une simplification du graphe de recombinaison ancestral qui est obtenue en éliminant la possibilité d'événements de coalescence entre les séquences qui ne partagent pas de matériel ancestral commun. Ce qui permet d'obtenir un coalescent Markovien par rapport aux sites le long des séquences. Ce SMC s'avère une bonne approximation du processus exact qui facilite grandement les simulations (chapitre 3).

Après avoir déduit les équations de récurrence en présence de mutation, de recombinaison, et de conversion sous les hypothèses du modèle de Moran, nous calculons les espérances de mesures de variation inter-locus et intra-locus trouvées

par Innan (2002a) en utilisant une approximation par un processus de diffusion. Ce qui permet d'estimer les paramètres de facteurs d'évolution à partir des polymorphismes de données d'ADN. Finalement, la seconde et dernière contribution de ce mémoire est d'utiliser le processus ancestral exact avec recombinaison et conversion pour trouver les espérances de mesures de variation inter-locus et intra-locus trouvées par Innan (2002b) pour le modèle à une infinité de sites. Nous calculons donc la covariance de temps de coalescence au même site et à des sites différents, ce qui nous permet d'obtenir l'hétérozygotie moyenne et le déséquilibre de liaison. Ces calculs effectués à partir du processus ancestral permettent d'obtenir de façon plus directe que par des processus de diffusion des formules pour les mesures de variabilité utilisées pour estimer les taux de mutation, de recombinaison et de conversion (Innan 2002a,b).

Chapitre 1

OUTILS PRÉLIMINAIRES DE GÉNÉTIQUE STATISTIQUE ET RAPPELS DE MODÈLES MATHÉMATIQUES D'ÉVOLUTION

Le premier chapitre débute par une présentation des notions élémentaires de génétique statistique que nous utilisons dans ce mémoire. Ensuite, nous introduisons des mesures de fréquences de génotypes et d'association d'allèles à deux loci ; soit l'équilibre de Hardy-Weinberg et le déséquilibre de liaison. La dernière section du chapitre décrit deux modèles de populations qui décrivent différentes modalités de reproduction ; le modèle de Moran (Moran et Watterson 1959 ; Feldman 1966) et le modèle de Wright-Fisher (1930-1931). Cette section est précédée par une description des facteurs d'évolution de mutation, de conversion intrachromosomique et de recombinaison intragénique sur lesquels sont fondés ces modèles mathématiques de populations.

1.1. NOTIONS DE GÉNÉTIQUE

1.1.1. Hérité

Chaque cellule humaine possède 23 paires de chromosomes : pour chaque paire, un chromosome est hérité de la mère et l'autre chromosome du père.

La méiose est une division cellulaire qui aboutit à la production de cellules sexuelles, ou gamètes, nécessaires pour la reproduction. Les gamètes sont des cellules haploïdes (23 chromosomes chacune).

La fécondation est la combinaison de deux gamètes, chacun résultant d'une méiose (ovule et spermatozoïde), produisant une cellule diploïde (23 paires de chromosomes).

1.1.2. Allèle, locus et génotype

Un allèle est une version d'un gène ou d'une séquence d'ADN (acide désoxyribonucléique) situé à un emplacement précis sur un chromosome. Il peut se retrouver à différents loci sur un même chromosome et même sur des chromosomes différents (duplication et translocation).

Un locus est l'emplacement physique précis et invariable qu'occupe un gène ou une séquence d'ADN sur un chromosome. La liste ordonnée des loci d'un génome est la cartographie génétique le représentant.

Un génotype est une paire d'allèles à un locus donné situé sur les deux copies du gène d'un individu, dont l'un est d'origine paternelle et l'autre d'origine maternelle. On dit qu'un individu est homozygote si les deux allèles au locus considéré sont identiques et qu'il est hétérozygote si les deux allèles sont différents. Un haplotype est une séquence d'allèles sur un même chromosome.

Dans la figure 1.1 où un locus correspond à un site occupé par un seul nucléotide (A pour adénine, C pour cytosine, G pour guanine, T pour thymine), les génotypes aux 9 loci (dans le sens vertical) pour cet individu sont : AA , TT , GG , AA , CC , AA , GT , GG et CC . Les haplotypes correspondants à ces loci (dans le sens horizontal) pour cet individu sont : $ATGACAGGC$ et $ATGACATGC$.

Maternel	A	T	G	A	C	A	G	G	C
Paternel	A	T	G	A	C	A	T	G	C

FIGURE 1.1. Génotypes et haplotypes.

1.2. ÉQUILIBRE DE HARDY-WEINBERG

1.2.1. Fréquence génotypique et fréquence allélique

La fréquence allélique d'un marqueur génétique est la proportion de cet allèle parmi un groupe d'individus. Elle peut être calculée à partir des fréquences génotypiques. Dans une population diploïde dans laquelle chaque individu est porteur d'une paire de gènes à un locus dont les allèles sont A et a , on peut avoir par exemple les fréquences génotypiques du Tableau 1.1.

TABLEAU 1.1. Fréquences de génotypes dans une population de taille $2N=294$.

Génotype	Nombre	Fréquence
AA	224	$P_{AA} = \frac{224}{294}$
Aa	64	$P_{Aa} = \frac{64}{294}$
aa	6	$P_{aa} = \frac{6}{294}$
Total	294	$1 = \frac{294}{294}$

On calcule alors les fréquences des allèles A et a comme suit :

$$P_A = P_{AA} + \frac{1}{2}P_{Aa} = \frac{2 \times 224 + 64}{2 \times 294} = 0,871,$$

$$P_a = P_{aa} + \frac{1}{2}P_{Aa} = \frac{2 \times 6 + 64}{2 \times 294} = 0,129,$$

de telle sorte que

$$P_A + P_a = 0,871 + 0,129 = 1.$$

1.2.2. Équilibre de Hardy-Weinberg

L'équilibre de Hardy-Weinberg (HWE, pour Hardy-Weinberg equilibrium) possède la propriété suivante : il permet de calculer la fréquence d'un génotype à partir des fréquences alléliques. De plus, les fréquences relatives des génotypes restent les mêmes au cours du temps. Dans une population diploïde de taille infinie à générations séparées dans laquelle les accouplements sont aléatoires d'une

génération à la suivante, l'équilibre de Hardy-Weinberg est atteint dès la seconde génération. Nous allons vérifier cela.

Soit un locus à deux allèles, A et a , dans une telle population. Il y a 9 arrangements possibles pour un croisement ordonné des deux individus à la génération t dont les fréquences sont données par les produits des fréquences génotypiques. Les lois de Mendel donnent la distribution conditionnelle des génotypes pour chaque croisement (voir le tableau 1.2).

TABLEAU 1.2. Effets de la ségrégation des lois de Mendel.

Croisement ordonné	Fréquence à la génération t	Distribution conditionnelle des génotypes		
		AA	Aa ou aA	aa
$AA \times AA$	$P_{AA}(t) \times P_{AA}(t)$	1	0	0
$AA \times Aa$	$P_{AA}(t) \times P_{Aa}(t)$	$\frac{1}{2}$	$\frac{1}{2}$	0
$AA \times aa$	$P_{AA}(t) \times P_{aa}(t)$	0	1	0
$Aa \times AA$	$P_{Aa}(t) \times P_{AA}(t)$	$\frac{1}{2}$	$\frac{1}{2}$	0
$Aa \times Aa$	$P_{Aa}(t) \times P_{Aa}(t)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$Aa \times aa$	$P_{Aa}(t) \times P_{aa}(t)$	0	$\frac{1}{2}$	$\frac{1}{2}$
$aa \times AA$	$P_{aa}(t) \times P_{AA}(t)$	0	1	0
$aa \times Aa$	$P_{aa}(t) \times P_{Aa}(t)$	0	$\frac{1}{2}$	$\frac{1}{2}$
$aa \times aa$	$P_{aa}(t) \times P_{aa}(t)$	0	0	1

Ainsi les fréquences des génotypes à la génération suivante $t + 1$ étant donné les fréquences des génotypes à la génération t sont :

$$\begin{aligned}
 P_{AA}(t+1) &= 1 \times P_{AA}(t)^2 + 0,5 \times 2 \cdot P_{AA}(t) \cdot P_{Aa}(t) + 0,25 \times P_{Aa}(t)^2 \\
 &= [P_{AA}(t) + 0,5 \times P_{Aa}(t)]^2 \\
 &= P_A(t)^2, \\
 P_{aa}(t+1) &= [P_{aa}(t) + 0,5 \times P_{Aa}(t)]^2 \\
 &= P_a(t)^2,
 \end{aligned}$$

$$\begin{aligned}
 P_{Aa}(t+1) &= 2[P_{AA}(t) + 0,5 \times P_{Aa}(t)] \cdot [P_{aa}(t) + 0,5 \times P_{Aa}(t)] \\
 &= 2P_A(t)P_a(t).
 \end{aligned}$$

De plus, on a :

$$P_{Aa}(t+1)^2 = 4 \cdot P_{AA}(t+1) \cdot P_{aa}(t+1).$$

Si les croisements aléatoires continuent, le même résultat est retrouvé à la génération $t+2$ par rapport à la génération $t+1$, c'est-à-dire que les fréquences des génotypes sont en équilibre. On dit qu'une population est en équilibre de Hardy-Weinberg (HWE) si elle vérifie

$$P_{Aa}^2 = 4 \cdot P_{AA} \cdot P_{aa}.$$

Dans cette population les fréquences des génotypes et les produits des fréquences alléliques correspondantes sont égales :

$$\begin{aligned}
 P_{AA} &= p_A^2, \\
 P_{Aa} &= 2p_A \cdot p_a, \\
 P_{aa} &= p_a^2.
 \end{aligned}$$

Lorsque les propriétés ci-dessus ne sont pas vérifiées, le déséquilibre de Hardy-Weinberg (HWD) est le résultat de plusieurs facteurs, notamment la sélection, la mutation, la migration, la dérive génétique et la structure de la population.

1.2.3. Déséquilibre de liaison

Le déséquilibre de liaison (LD, pour linkage disequilibrium) mesure l'association non aléatoire d'allèles à deux loci ou plus qui ne sont pas nécessairement situés sur le même chromosome. Les allèles qui existent aujourd'hui sont le résultat d'événements de mutation ancestraux. Si deux loci sont assez éloignés l'un de l'autre, la recombinaison génère de nouveaux arrangements des allèles laissant ainsi les arrangements d'allèles à proximité refléter plus vraisemblablement des haplotypes ancestraux.

Pour calculer le déséquilibre de liaison à deux loci (A) et (B), avec les allèles A, a , et B, b , respectivement, on suppose d'abord que la population est en équilibre de Hardy-Weinberg à chaque locus. Alors les fréquences génotypiques et alléliques satisfont :

$$P_{AA} = p_A^2, P_{Aa} = 2p_A \cdot p_a, P_{aa} = p_a^2, P_{AA} + P_{Aa} + P_{aa} = 1, p_A + p_a = 1,$$

$$P_B = p_B^2, P_{Bb} = 2p_B \cdot p_b, P_{bb} = p_b^2, P_{BB} + P_{Bb} + P_{bb} = 1, p_B + p_b = 1.$$

Les fréquences des haplotypes peuvent généralement être exprimées sous la forme :

$$p_{AB} = p_A p_B + D_{AB},$$

$$p_{Ab} = p_A p_b + D_{Ab},$$

$$p_{aB} = p_a p_B + D_{aB},$$

$$p_{ab} = p_a p_b + D_{ab},$$

avec $p_{AB} + p_{Ab} + p_{aB} + p_{ab} = 1$. La déviation D_{AB} de la fréquence de l'haplotype AB par rapport au produit des fréquences des allèles A et B correspond à la covariance des fréquences de A et B (0 ou 1) sur un haplotype tiré au hasard dans la population. Il en est de même pour les autres haplotypes. Pour revenir aux fréquences des allèles, on observe que

$$\begin{aligned} p_A &= p_{AB} + p_{Ab} \\ &= (p_A p_B + D_{AB}) + (p_A p_b + D_{Ab}) \\ &= p_A + D_{AB} + D_{Ab}, \end{aligned}$$

ce qui implique que $D_{AB} = -D_{Ab}$. De façon analogue, on montre plus généralement que

$$D_{AB} = -D_{Ab} = -D_{aB} = D_{ab} = D.$$

On peut ainsi réécrire les fréquences des haplotypes comme suit :

$$p_{AB} = p_A p_B + D,$$

$$p_{Ab} = p_A p_b - D,$$

$$p_{aB} = p_a p_B - D,$$

$$p_{ab} = p_a p_b + D,$$

où

$$D = p_{AB} - p_A p_B.$$

Quand $D = 0$, la population est en équilibre de liaison. Le déséquilibre de liaison de deux loci bialléliques peut aussi être mesuré par la corrélation entre les fréquences des allèles A et B sur un haplotype tiré au hasard dans la population (Hill et Robertson 1968), c'est-à-dire

$$R = \frac{D}{\sqrt{p_A p_a p_B p_b}}.$$

Une mesure connexe développée pour contourner la complexité des expressions des moments de la distribution de R (Ohta & Kimura 1971) est donnée par

$$\sigma_d^2 = \frac{\mathbb{E}[D]^2}{\mathbb{E}[p_A p_a p_B p_b]}. \quad (1.2.1)$$

En remontant le temps à partir d'un échantillon de n séquences (haplotypes), le temps de coalescence à un locus est défini comme le temps jusqu'à la fusion de toutes les lignées ancestrales à ce locus en une première lignée ancestrale commune (voir la section 2.1). McVean (2002) a démontré (voir l'annexe A.2) que l'équation (1.2.1) peut être exprimée en termes de covariances des temps de coalescence aux deux loci pour différentes configurations de chromosomes :

$$\sigma_d^2 = \frac{C_{ij,ij} - 2C_{ij,ik} + C_{ij,kl}}{\mathbb{E}[t]^2 + C_{ij,kl}}, \quad (1.2.2)$$

où $C_{ij,kl}$ est la covariance entre le temps de coalescence au locus (A) entre deux séquences échantillonnées i et j et le temps de coalescence au locus (B) entre deux séquences échantillonnées k et l (voir les covariances des états 1, 2, et 3 de la section (2.2.1)). Ici $\mathbb{E}[t]$ est l'espérance du temps de coalescence d'une paire de

chromosomes à un locus. Sous le coalescent standard, $\mathbb{E}[t] = 1$. Ces covariances sont déduites en résolvant un système d'équations linéaires (Griffiths 1981 ; voir aussi Pluzhnikov et Donnelly 1996).

1.3. FACTEURS D'ÉVOLUTION

Dans cette section, nous présentons une brève description des facteurs d'évolution considérés dans les modèles mathématiques d'évolution des populations : la mutation, la recombinaison et la conversion.

1.3.1. Mutation

On peut voir le facteur de mutation comme un processus faisant le pont entre les arbres généalogiques et les données génétiques. En effet, étant donné la structure de l'arbre généalogique d'un ensemble de données, chaque mutation sur une branche de l'arbre divise l'échantillon en deux groupes, les séquences échantillonnées qui contiennent la mutation et celles qui ne la contiennent pas, d'où le polymorphisme observé.

Chaque ensemble de données génétiques a un modèle de mutation sous-jacent. On distingue deux types de modèles : les modèles basés sur les différents allèles, et les modèles de séquences de nucléotides qui identifient les sites où les allèles sont différents. La différence majeure entre les deux est que les premiers ne contiennent aucune information sur les liens ancestraux entre les allèles de l'échantillon alors que les seconds génèrent cette information au moyen des variations polymorphiques au sein de loci dans l'échantillon.

Dans ce mémoire, nous considérons le modèle à une infinité de sites (Kimura 1969 ; Watterson 1975). Notons que tous ces modèles peuvent être étudiés sous le coalescent.

La modélisation de la mutation sous le coalescent de Kingman suppose que toute variation est de sélection neutre. Par définition, sous un modèle neutre, la variation génétique n'affecte pas le succès de la reproduction des organismes et n'a

pas d'influence sur la structure des arbres généalogiques de l'échantillon. Ainsi, les processus généalogique et de mutation peuvent être séparés (Tavaré 1984), et tout modèle de mutation est applicable en prenant en compte les changements le long des branches de chaque généalogie possible.

Modèle à une infinité de sites

Le modèle à une infinité de sites assume que chaque mutation a lieu à un site non muté. Il suppose un grand nombre de positions possibles dont chacune a un taux de mutation très petit. Cette hypothèse est souvent appropriée pour les séquences d'ADN dont le taux de mutation par nucléotide est particulièrement bas, de l'ordre de 10^{-8} à 10^{-9} par génération dans le cas de plusieurs organismes (Drake *et al.* 1998). D'autant plus qu'au sein de plusieurs organismes les niveaux de polymorphisme sont assez bas pour négliger les mutations multiples. Par exemple, le taux de polymorphisme du génome humain, qui est le ratio de sites qui sont polymorphes lors de la comparaison d'une paire de séquences est de l'ordre de 10^{-3} (Cargill *et al.* 1999; Stephens *et al.* 2001). De plus si chaque site peut potentiellement muter, le fait que les génomes humains soient identiques à 99,9% implique que deux mutations ou plus au même site sont improbables. Ainsi les hypothèses du modèle à une infinité de sites sont acceptables pour les séquences d'ADN.

Mesure de polymorphisme

Une façon de mesurer les polymorphismes de séquences d'ADN est de calculer le nombre moyen de différences entre les paires de séquences dans l'échantillon, appelé π (Tajima 1983). Le choix de cette mesure est justifié par le fait que son espérance est liée au taux de mutation (voir la démonstration dans l'annexe (A.1)). Cette mesure est notamment une extension du concept d'hétérozygotie chez des individus d'une espèce diploïde (Tajima 1983). Elle est calculée en comparant chaque séquence à chacune des autres, en comptant le nombre de différences

entre elles, et en prenant la moyenne de ces dernières. Elle est donnée par

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^n k_{ij}, \quad (1.3.1)$$

où k_{ij} est le nombre de différences entre la séquence i et la séquence j dans un échantillon de n séquences. Par exemple, dans la figure 1.2 on a $k_{12} = 1$ et $k_{34} = 2$.

Séquence	1	. . .	A	. . .	G	. . .	C	. . .	G	. . .
Séquence	2	. . .	A	. . .	G	. . .	T	. . .	G	. . .
Séquence	3	. . .	A	. . .	A	. . .	T	. . .	T	. . .
Séquence	4	. . .	T	. . .	G	. . .	T	. . .	T	. . .
Séquence	5	. . .	T	. . .	G	. . .	T	. . .	T	. . .

FIGURE 1.2. Cinq séquences d'ADN à 4 sites polymorphes.

Source: Wakeley 2008 - Chapter 1.

1.3.2. Recombinaison

La théorie de Fisher-Muller (1930-1932) attribue l'avantage évolutif de la recombinaison à sa capacité d'incorporer en un seul individu des mutations potentiellement avantageuses ayant apparues séparément chez différents individus. Pendant la méiose, les différentes résolutions des jonctions de Holliday (Stahl 1994) aboutissent à deux types d'enjambement : la recombinaison et la conversion. Pour distinguer les deux types d'enjambement, on présente dans la prochaine sous-section (voir la figure 1.3) le modèle de Szostak de réparation de cassures double-brin par jonctions doubles de Holliday (double-strand break repair model via double Holliday junctions) lors de la méiose.

1.3.3. Conversion

L'homogénéité de la variation des séquences d'ADN dans les familles multigéniques est issue de plusieurs mécanismes, dont la conversion et l'enjambement inégal. Ce qui suit est une description du modèle de Szostak (voir la figure 1.3)

de réparation de cassures double-brin par jonctions doubles de Holliday (double-strand break repair model for meiotic cross-over via double Holliday junctions) qui explique les deux mécanismes.

Soient deux duplexes d'ADN qui s'engagent dans un enjambement. Les informations génétiques respectives des duplexes qui permettent de suivre les échanges sont c, d, e et C, D, E . Suite aux cassures et résections des deux brins du premier duplex (1), on perd de l'information génétique du premier duplex. Une invasion de brin s'en suit (2), et forme une boucle sur le brin envahi, créant ainsi des hétéroduplexes. (3) et (4a, b) Les sections d'ADN perdues ont été réparées (encadrement par une ligne entre-coupée). (5) Pour la résolution, il y a plusieurs possibilités de coupure et ligation aux points ①, ②, ③ et ④. Si on a une coupure et ligation (② et ④) ou (① et ③), on a de la conversion sans recombinaison. On illustre ici celle de ② et ④ (5b) qui résulte alors dans les duplexes ($c c'$) et ($e e'$) gardant leurs positions initiales, résultant ainsi dans des duplexes non-recombinants. De façon plus essentielle, suite à la réparation des mésappariements qui répare d en D , ceci résulte dans une conversion puisque D a été transféré à l'autre brin d'ADN.

Si par contre la résolution est différente des deux côtés, alors on a une coupure et ligation du type (① et ④) ou (② et ③). On illustre ici celle de ② et ③ (5a) qui résulte dans l'échange de positions du duplex ($e e'$) avec le duplex ($E E'$) pendant que ($c c'$) et ($C C'$) gardent leurs positions initiales, résultant ainsi dans des duplexes recombinants. Toutefois, de façon plus essentielle, suite à la réparation des mésappariements qui répare d en D , ceci résulte dans une conversion puisque D une fois encore a été transféré à l'autre brin.

Ainsi pendant la recombinaison ou la non recombinaison l'information D est transférée à l'autre brin.

1.3.4. Dérive génétique

Les modèles considérés dans ce mémoire sont sous l'hypothèse d'une population de taille finie et constante et dont la reproduction est un processus aléatoire.

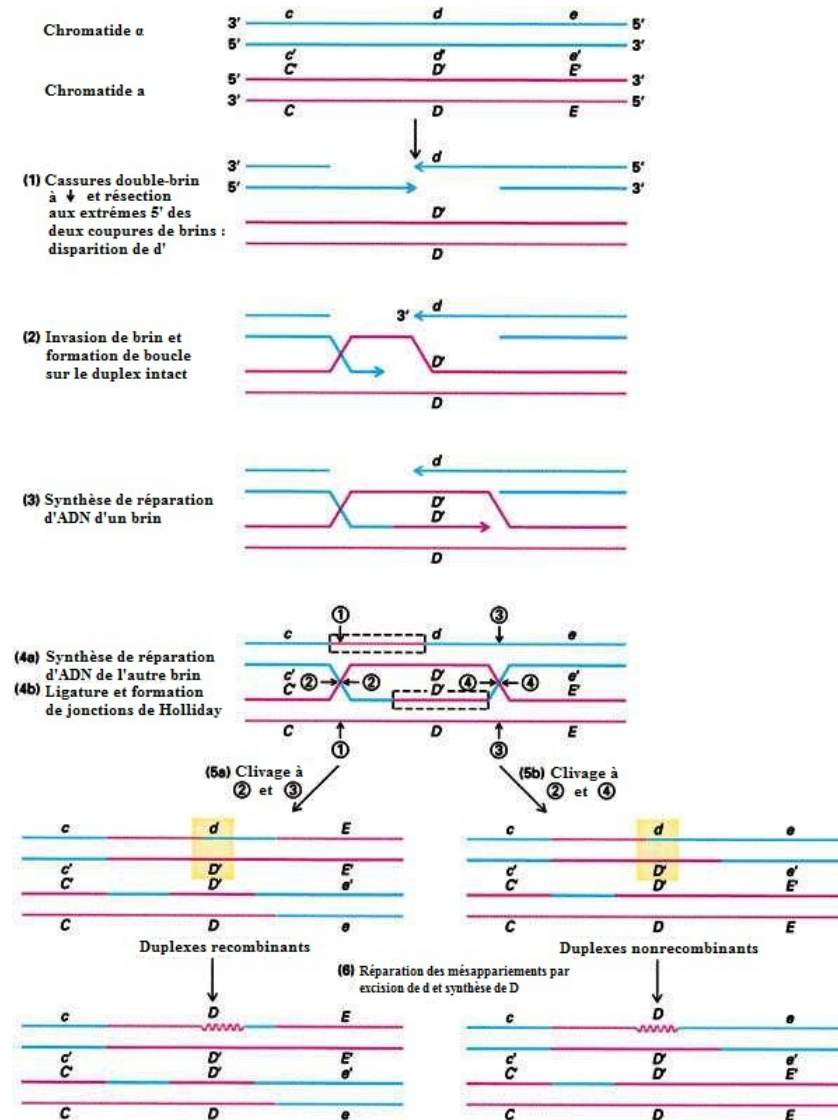


FIGURE 1.3. Modèle de réparation de cassures double-brin pour l'enjambement lors de la méiose, par jonctions doubles de Holliday (Double-strand break repair model for meiotic cross-over via double Holliday junctions).

Source: Lecture 13, 2005 : Recombination and Gene Conversion - Bruce Walsh, University of Arizona

Ainsi certains individus ne laissent pas de descendants à la génération suivante.

Cette perte aléatoire de lignées génétiques dans le temps est appelée la dérive génétique. Cette dernière regardée en remontant le temps est le coalescent (Chapitre 2).

Une des mesures de dérive génétique présentée dans ce mémoire, l'hétérozygotie, est discutée davantage dans la section suivante de présentation des modèles considérés, soient les modèles de Wright-Fisher et de Moran.

1.4. MODÈLES DE POPULATIONS

L'analyse théorique de la génétique des populations est fondée sur notre capacité à construire des modèles qui prennent en compte des caractéristiques biologiques essentielles tout en étant assez simples pour des analyses mathématiques. Les deux modèles les plus utilisés en génétique des populations sont le modèle de Wright-Fisher et le modèle de Moran. Aucun des deux n'a été développé conformément aux connaissances biologiques connues d'un organisme particulier. Les deux modèles appartiennent toutefois à une grande classe de modèles qui décrivent différentes modalités de reproduction et se fondent sur des hypothèses biologiques des populations.

Plus important encore, tous ces modèles aboutissent au coalescent de Kingman sous certaines conditions limites. Le modèle de Wright-Fisher représente un cas de générations sans chevauchement et le modèle de Moran représente un cas idéal de générations se chevauchant. La réalité des populations se trouvent entre ses deux extrêmes. Le coalescent de Kingman est une approximation du processus ancestral pour un échantillon sous le modèle de Moran ou de Wright-Fisher quand la population est grande. Notons que certaines de ses caractéristiques sont exactes sous le modèle de Moran. Nous achevons ce chapitre par une description des deux modèles dans le sens conventionnel d'écoulement du temps avant d'aborder le processus ancestral dans le chapitre suivant.

1.4.1. Modèle de Wright-Fisher

Dans le modèle introduit par Fisher (1930) et Wright (1931), la taille de la population est supposée finie et constante au cours du temps. On considère un organisme haploïde, dont la population consiste de $2N$ copies du génome.

Dans cette section, on présente le modèle de Wright-Fisher sans recombinaison ni conversion, de telle sorte que chaque individu a un seul parent. Les générations ne se chevauchent pas, c'est-à-dire que tous les individus d'une même génération meurent après avoir produit les individus de la génération suivante. De plus dans le modèle neutre il est supposé qu'aucun individu particulier n'est favorisé par la sélection. Chacun des $2N$ individus de la génération suivante est produit comme s'il choisissait son parent uniformément parmi les $2N$ parents de la génération précédente et indépendamment des autres.

Regardé dans le sens conventionnel d'écoulement du temps, un individu i produit un nombre aléatoire k_i de descendants à la génération suivante qui suit une loi $\text{Bin}(2N, \frac{1}{2N})$. Autrement dit, chacun des $2N$ descendants a une probabilité égale à $\frac{1}{2N}$ que l'individu i soit son parent indépendamment des autres. Cependant dans ce sens on n'a pas indépendance des nombres k_i de descendants de première génération des parents $i = 1, \dots, 2N$ à une génération donnée. En effet la connaissance du nombre de descendants d'un individu particulier change la loi du nombre de descendants d'un autre puisque $\sum_{i=1}^{2N} k_i = 2N$.

Cependant, en remontant le temps, les parents d'un échantillon aléatoire d'individus à une génération donnée sont choisis uniformément parmi tous les individus de la génération précédente et indépendamment les uns des autres.

Supposons que le gène d'intérêt à un locus se présente sous la forme d'un de deux allèles, A ou a . On note Z_t le nombre d'individus porteurs de l'allèle A dans la population à la génération t . D'après les hypothèses ci-dessus, la suite $\{Z_t, t \geq 0\}$ est une chaîne de Markov à temps discret à valeurs dans $S = \{0, 1, \dots, 2N\}$ décrite par la propriété suivante : la loi de Z_t sachant Z_{t-1} est binomiale de paramètres

$2N$ et $\frac{Z_{t-1}}{2N}$, notée $\text{Bin}(2N, \frac{Z_{t-1}}{2N})$. En d'autres termes, la matrice de transition $P = (p_{i,j})$ de la chaîne est donnée par

$$p_{i,j} = \mathbb{P}\left(Z_t = j \mid Z_{t-1} = i\right) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}, \quad (1.4.1)$$

pour $0 \leq i, j \leq 2N$. Il s'ensuit que :

$$\mathbb{E}\left[Z_t \mid Z_{t-1} = i\right] = i, \quad (1.4.2)$$

$$\text{Var}\left[Z_t \mid Z_{t-1} = i\right] = 2N \frac{i}{2N} \left(1 - \frac{i}{2N}\right). \quad (1.4.3)$$

Hétérozygotie et dérive génétique

L'hétérozygotie dans une population haploïde est la probabilité que deux individus choisis aléatoirement sans remise dans la population totale de taille $2N$ soient de types différents au même locus. Dans une population diploïde, ceci correspond à la probabilité qu'un individu soit hétérozygote à ce locus.

Soit $p_t = \frac{Z_t}{2N}$ la fréquence de l'allèle A à un locus donné dans une population haploïde de taille $2N$ à la génération t avec a l'allèle alternatif au locus en question. La fréquence p_t varie dans le temps. Selon que A est choisi plus souvent ou moins souvent p_t aura tendance à augmenter ou diminuer. L'effet de ceci est défini comme la *dérive génétique*.

L'hétérozygotie à la génération t est

$$H_t = 2p_t(1 - p_t) = \frac{2Z_t(2N - Z_t)}{2N(2N - 1)}.$$

Proposition 1.4.1. *Sous le modèle de Wright-Fisher, l'hétérozygotie moyenne décroît de façon exponentielle à un taux de $\frac{1}{2N}$ par génération.*

DÉMONSTRATION. En conditionnant sur la fréquence de A à la génération $t - 1$, on obtient que

$$\begin{aligned}\mathbb{E}\left[H_t \mid p_{t-1}\right] &= \mathbb{E}\left[2p_t(1 - p_t) \mid p_{t-1}\right] \\ &= 2\mathbb{E}\left[p_t - p_t^2 \mid p_{t-1}\right] \\ &= 2\left(\mathbb{E}\left[p_t \mid p_{t-1}\right] - \text{Var}\left[p_t \mid p_{t-1}\right] - \mathbb{E}\left[p_t \mid p_{t-1}\right]^2\right).\end{aligned}$$

En utilisant les relations (1.4.2) et (1.4.3), on a les expressions suivantes :

$$\begin{aligned}\mathbb{E}\left[p_t \mid p_{t-1}\right] &= p_{t-1}, \\ \text{Var}\left[p_t \mid p_{t-1}\right] &= \frac{p_{t-1}(1 - p_{t-1})}{2N}.\end{aligned}$$

On trouve alors que

$$\begin{aligned}\mathbb{E}\left[H_t \mid p_{t-1}\right] &= 2\left(p_{t-1} - \frac{p_{t-1}(1 - p_{t-1})}{2N} - p_{t-1}^2\right) \\ &= 2p_{t-1}(1 - p_{t-1})\left(1 - \frac{1}{2N}\right) \\ &= H_{t-1}\left(1 - \frac{1}{2N}\right).\end{aligned}$$

Cela nous permet de déduire par récurrence que :

$$\begin{aligned}\mathbb{E}\left[H_t \mid p_0\right] &= \mathbb{E}\left[\mathbb{E}\left[H_t \mid p_{t-1}\right] \mid p_0\right] = H_0\left(1 - \frac{1}{2N}\right)^t \\ &\approx H_0 e^{-\frac{t}{2N}}.\end{aligned}\tag{1.4.4}$$

□

Cette approximation est valide pour $2N$ assez grand. La décroissance de l'hétérozygotie moyenne est une mesure de la dérive génétique. On dit que sous le modèle de Wright-Fisher la dérive génétique se produit à un taux de $\frac{1}{2N}$ par génération.

1.4.2. Modèle de Moran

L'une des hypothèses du modèle de Wright-Fisher est que les générations sont séparées de telle sorte que les individus d'une génération donnée meurent tous après avoir généré la génération suivante. Le modèle de Moran (Moran et Watterson 1959 ; Feldman 1966) est le modèle le plus simple avec chevauchement de générations, car un seul individu est remplacé à la fois. Comme le modèle de Wright-Fisher, le modèle de Moran est un modèle stochastique avec dérive génétique. Dans ce modèle également la taille de la population est supposée finie et constante au cours du temps, égale à $2N$.

Le temps est discret et à chaque instant, deux individus sont choisis au hasard avec remplacement parmi les $2N$ individus de la population, et ils produisent un descendant. L'auto-fécondation se produit donc avec probabilité $\frac{1}{2N}$. Ensuite un individu est choisi au hasard parmi les mêmes $2N$ individus et celui-ci est remplacé par le descendant. Un instant correspond donc à un événement de naissance-mort. Supposons encore que le gène d'intérêt à un locus se présente sous la forme d'un de deux allèles, A ou a . Soit Z_t le nombre d'individus porteurs de l'allèle A dans la population à l'instant t (le t -ième événement de naissance-mort). D'après les hypothèses ci-dessus, la suite $\{Z_t, t \geq 0\}$ est une chaîne de Markov à temps discret à valeurs dans $S = \{0, 1, \dots, 2N\}$ dont la matrice de transition $P = (p_{i,j})$ est donnée par

$$p_{i,j} = \mathbb{P}\left(Z_t = j \mid Z_{t-1} = i\right) = \begin{cases} \frac{i}{2N} \left(1 - \frac{i}{2N}\right), & \text{si } j = i - 1, \\ 1 - 2\frac{i}{2N} \left(1 - \frac{i}{2N}\right), & \text{si } j = i, \\ \frac{i}{2N} \left(1 - \frac{i}{2N}\right), & \text{si } j = i + 1, \\ 0, & \text{autrement,} \end{cases}$$

pour $0 \leq i, j \leq 2N$. Il s'ensuit que

$$\mathbb{E}\left[Z_t \mid Z_{t-1} = i\right] = i, \quad (1.4.5)$$

$$\text{Var}\left[Z_t \mid Z_{t-1} = i\right] = 2\frac{i}{2N}\left(1 - \frac{i}{2N}\right). \quad (1.4.6)$$

Hétérozygotie et dérive génétique

On a le résultat suivant.

Proposition 1.4.2. *Sous le modèle de Moran, l'hétérozygotie moyenne décroît de façon exponentielle à un taux de $\frac{1}{2N^2}$ par événement de naissance-mort.*

DÉMONSTRATION. En introduisant la fréquence de A à tout instant $t \geq 0$, donnée par $p_t = \frac{Z_t}{2N}$, on a comme précédemment

$$\begin{aligned} \mathbb{E}\left[H_t \mid p_{t-1}\right] &= \mathbb{E}\left[2p_t(1-p_t) \mid p_{t-1}\right] \\ &= 2\left(\mathbb{E}\left[p_t \mid p_{t-1}\right] - \text{Var}\left[p_t \mid p_{t-1}\right] - \mathbb{E}\left[p_t \mid p_{t-1}\right]^2\right). \end{aligned}$$

En utilisant (1.4.5) et (1.4.6), on a les expressions suivantes pour les espérance et variance conditionnelles de p_t étant donné p_{t-1} :

$$\begin{aligned} \mathbb{E}\left[p_t \mid p_{t-1}\right] &= p_{t-1}, \\ \text{Var}\left[p_t \mid p_{t-1}\right] &= 2\frac{p_{t-1}(1-p_{t-1})}{4N^2}. \end{aligned}$$

On trouve alors que

$$\begin{aligned} \mathbb{E}\left[H_t \mid p_{t-1}\right] &= 2\left(p_{t-1} - \frac{p_{t-1}(1-p_{t-1})}{2N^2} - p_{t-1}^2\right) \\ &= 2p_{t-1}(1-p_{t-1})\left(1 - \frac{1}{2N^2}\right) \\ &= H_{t-1}\left(1 - \frac{1}{2N^2}\right). \end{aligned}$$

Cela permet de conclure par récurrence que

$$\begin{aligned}\mathbb{E}\left[H_t\middle|p_0\right] &= \mathbb{E}\left[\mathbb{E}\left[H_t\middle|p_{t-1}\right]\middle|p_0\right] = H_0\left(1 - \frac{1}{2N^2}\right)^t \\ &\approx H_0e^{-\frac{t}{2N^2}}.\end{aligned}\tag{1.4.7}$$

□

Cette approximation est valide pour $2N$ assez grand. On dit que sous le modèle de Moran la dérive génétique se produit à un taux de $\frac{1}{2N^2}$ par événement de naissance-mort. Pour comparer la dérive sous le modèle de Moran à celle sous le modèle de Wright-Fisher, on considère que $2N$ événements de naissance-mort sous le premier correspond à une génération sous le second. Le choix de cette échelle est approprié du point de vue d'un individu puisque la probabilité qu'un individu décède en un événement de naissance-mort est $1/2N$, et donc la durée de vie d'un individu en nombre d'événements de naissance-mort suit une loi géométrique de paramètre $1/2N$ d'espérance de $2N$ ce qui correspond à une génération (Chapitre 3 du livre de Wakeley, 2008). En posant $\tau = t/2N$, on trouve que

$$\mathbb{E}\left[H_t\middle|p_0\right] \approx H_0e^{-\frac{\tau}{N}}.\tag{1.4.8}$$

Une comparaison de (1.4.8) et (1.4.4) indique que le taux de dérive génétique est deux fois plus grand sous le modèle de Moran comparé au modèle de Wright-Fisher. Autrement dit, une taille de population égale à $2N$ dans le modèle de Moran correspond à une taille de population égale à N dans le modèle de Wright-Fisher (Ewens 2004, p.121).

Chapitre 2

PROCESSUS ANCESTRAL

Dans ce chapitre, nous présentons le coalescent standard de Kingman (1982a,b,c) à temps continu. Puis nous montrons comment il peut être utilisé pour chacun des modèles décrits dans le chapitre précédent : le modèle de Wright-Fisher et le modèle de Moran. Par la suite, nous considérons l'extension de Hudson (1983) du processus de coalescence avec recombinaison intragénique en y ajoutant la conversion. Cette partie regroupe tous les résultats, concernant la survenance et le temps jusqu'à la survenance d'un événement de mutation, de coalescence, de recombinaison ou de conversion, utilisés dans la suite du mémoire.

2.1. LE COALESCENT DE KINGMAN

Le processus de coalescence est une approche développée par Kingman pour étudier les facteurs d'évolution d'une population en remontant le temps. Le coalescent de Kingman (1982a,b,c) est la limite du processus ancestral d'une vaste classe de structures de population qui inclut les modèles de Wright-Fisher et de Moran. En remontant le temps, il retrace les lignées ancestrales, qui suivent les ancêtres génétiques à un locus d'un échantillon aléatoire de taille n dans une population ambiante de taille finie, disons $2N$, lorsque celle-ci devient de plus en plus grande, i.e., $2N \rightarrow \infty$. En numérotant les éléments de l'échantillon de 1 à n , le coalescent de Kingman est une chaîne de Markov à temps continu sur les partitions de l'ensemble $\{1, \dots, n\}$. Un sous-ensemble correspond à une lignée qui est ancestrale aux éléments du sous-ensemble seulement. Chaque paire de sous-ensembles fusionne au taux 1 indépendamment des autres. Cela correspond à un

événement de coalescence de deux lignées ancestrales en remontant le temps.

A chaque événement de coalescence deux lignées fusionnent en une lignée ancestrale commune et diminue ainsi d'une unité le nombre de lignées ancestrales jusqu'à la dernière. La lignée ancestrale restante suivant le dernier événement de coalescence est l'ancêtre commun le plus récent (most recent common ancestor, ou MRCA) de l'échantillon. La généalogie ainsi obtenue pour l'échantillon dans une population nous permet d'étudier la variabilité dans l'échantillon et d'obtenir de l'information sur la population. C'est une approche idéale pour faire de l'inférence statistique.

Le nombre de lignées dans le coalescent de Kingman est un processus de mort. Le taux avec lequel le nombre de lignées passe de n à $n - 1$ est le nombre de paires possibles avec n lignées, soit $\binom{n}{2} = \frac{n(n-1)}{2}$. C'est le paramètre de la loi exponentielle pour le temps passé avec n lignées. L'approximation limite par le coalescent de Kingman (1982a,b) pour des modèles de population de taille finie à temps discret est déduite en regardant, d'un instant à l'instant précédent, la probabilité que deux lignées aient un ancêtre en commun. Pour passer à temps continu, le processus considère l'accélération du passage du temps en même temps que la taille de la population, $2N$, grandit, ce qui permet d'obtenir une limite non triviale.

On cite deux résultats bien connus sur la fonction exponentielle et la loi exponentielle qui seront utilisés dans cette section.

Résultat 2.1.1. *Pour tout $x \in \mathbb{R}$, on a $\lim_{n \rightarrow +\infty} (1 + \frac{x}{n})^n = e^x$.*

Résultat 2.1.2. *Soient X et Y deux variables aléatoires indépendantes qui suivent une loi exponentielle de paramètres respectifs λ_1 et λ_2 . Alors $\min(X, Y)$ est une variable aléatoire qui suit une loi exponentielle de paramètre $\lambda_1 + \lambda_2$. En effet,*

pour tout nombre réel $t > 0$, on a

$$\begin{aligned} \mathbb{P}(\min(X + Y) \leq t) &= 1 - \mathbb{P}(\min(X + Y) > t) \\ &= 1 - \mathbb{P}(X > t) \cdot \mathbb{P}(Y > t) \\ &= 1 - \exp(-(\lambda_1 + \lambda_2)t). \end{aligned}$$

De plus, la probabilité que la variable X soit plus petite que Y est donnée par

$$\begin{aligned} \mathbb{P}(X < Y) &= \int_0^\infty \int_0^y f_X(x) f_Y(y) dx dy \\ &= \int_0^\infty \int_0^y \lambda_1 \exp(-\lambda_1 x) \exp(-\lambda_2 x) dx dy = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \end{aligned}$$

2.2. PROCESSUS ANCESTRAL SOUS LE MODÈLE DE WRIGHT-FISHER

Coalescence de deux lignées

Choisissons aléatoirement deux individus dans une population haploïde de taille $2N$ qui évolue sous le modèle neutre de Wright-Fisher, sans recombinaison ni conversion. Alors on a :

- La probabilité qu'ils aient le même parent est $\frac{2N}{(2N)^2} = \frac{1}{2N}$.
- La probabilité qu'ils n'aient pas le même parent est donc $\mathbb{P}(2) = 1 - \frac{1}{2N}$.
- La probabilité que leurs grands-parents soient différents, c'est-à-dire que leurs parents soient différents et que les parents de ceux-ci soient différents est $\mathbb{P}(2)^2 = (1 - \frac{1}{2N})^2$.
- La probabilité qu'après k générations, les deux lignées soient toujours distinctes est $\mathbb{P}(2)^k = (1 - \frac{1}{2N})^k$, pour tout entier $k \geq 1$.

Cela signifie que le temps en nombre de générations jusqu'à la coalescence des deux lignées, représenté par τ_2 , est une variable aléatoire de loi géométrique de paramètre $\frac{1}{2N}$, notée $\text{Géo}(\frac{1}{2N})$.

Coalescence de deux lignées parmi n

On considère un échantillon aléatoire de taille n à une génération donnée dans une population de taille $2N$ qui évolue sous le modèle de Wright-Fisher. On définit $\mathbb{P}(n)$: la probabilité que les n individus aient des parents tous distincts à la génération précédente. Cet événement se réalise si le parent du premier individu de l'échantillon est choisi parmi $2N$ individus, le parent du second individu est choisi parmi les $2N - 1$ autres individus et ainsi de suite :

$$\begin{aligned} \mathbb{P}(n) &= \frac{2N}{2N} \left(\frac{2N-1}{2N} \right) \left(\frac{2N-2}{2N} \right) \cdots \left(\frac{2N-n+1}{2N} \right) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{2N} \right) \\ &= 1 - \left(\frac{1+2+\cdots+(n-1)}{2N} \right) + o\left(\frac{1}{N}\right) \\ &= 1 - \frac{n(n-1)}{4N} + o\left(\frac{1}{N}\right) = 1 - \frac{\binom{n}{2}}{2N} + o\left(\frac{1}{N}\right). \end{aligned}$$

Quand la population est grande, la probabilité d'avoir plus de deux lignées qui coalescent est négligeable par rapport à la probabilité qu'exactement deux lignées coalescent. A chaque génération un événement de coalescence se produit dans l'échantillon avec probabilité

$$1 - \mathbb{P}(n) = \frac{\binom{n}{2}}{2N} + o\left(\frac{1}{N}\right).$$

Ainsi, τ_n le temps en nombre de générations jusqu'à la prochaine coalescence dans l'échantillon suit une loi géométrique de paramètre $1 - \mathbb{P}(n)$, notée $\text{Géo}(1 - \mathbb{P}(n))$, c'est-à-dire que

$$\mathbb{P}(\tau_n > k) = \left(1 - \frac{\binom{n}{2}}{2N} + o\left(\frac{1}{N}\right) \right)^k,$$

pour tout entier $k \geq 1$.

Approximation en temps continu

On considère T_n le temps jusqu'à une première coalescence en nombre de $2N$ générations à partir d'un échantillon de taille n . On a

$$\begin{aligned} \mathbb{P}(T_n > y) &= \mathbb{P}(\tau_n > \lfloor 2Ny \rfloor) \\ &= \left(1 - \frac{\binom{n}{2}}{2N} + o\left(\frac{1}{N}\right)\right)^{\lfloor 2Ny \rfloor} \longrightarrow e^{-\binom{n}{2}y}, \end{aligned}$$

lorsque $N \rightarrow \infty$, pour tout nombre réel $y > 0$, où $\lfloor 2Ny \rfloor$ dénote la partie entière de $2Ny$. Cela signifie que T_n suit une loi exponentielle de paramètre $\binom{n}{2}$, notée $\text{Exp}\left(\frac{n(n-1)}{2}\right)$. A la fin de ce temps, on se retrouve avec $n-1$ lignées ancestrales et on reste dans cet état un temps T_{n-1} qui suit asymptotiquement une loi exponentielle de paramètre $\binom{n-1}{2}$, indépendant de T_n , et ainsi de suite. Cela correspond asymptotiquement au processus de coalescence de Kingman.

Survenance d'un événement de mutation

Sous le modèle de Wright-Fisher, on suppose qu'un événement de mutation se produit chez tout descendant en une génération avec probabilité $\mu = \frac{\theta}{4N}$ et ne se produit pas avec probabilité $1 - \mu$, indépendamment des autres.

Il résulte de l'hypothèse ci-dessus que X_{mut} le nombre de mutations sur une lignée ancestrale en $2N$ générations est de loi binomiale de paramètres $2N$ et μ , notée $\text{Bin}(2N, \mu)$. Lorsque $N \rightarrow \infty$, la loi de X_{mut} est une loi de Poisson de paramètre $\frac{\theta}{2}$, notée $\text{Poisson}\left(\frac{\theta}{2}\right)$.

En effet, puisque X_{mut} le nombre de mutations sur une lignée ancestrale en $2N$ générations suit une loi $\text{Bin}(2N, \mu)$, on a

$$\begin{aligned} \mathbb{P}(X_{mut} = k) &= \binom{2N}{k} \mu^k (1 - \mu)^{2N-k} \\ &= \frac{2N(2N-1)\cdots(2N-k+1)}{k(k-1)\cdots 1 \times 2^k} \times \frac{\theta^k}{(2N)^k} \times \frac{\left(1 - \frac{\theta}{4N}\right)^{2N}}{\left(1 - \frac{\theta}{4N}\right)^k}, \end{aligned}$$

d'où

$$\lim_{N \rightarrow +\infty} \mathbb{P}(X_{mut} = k) = \frac{\theta^k}{k! 2^k} e^{-\frac{\theta}{2}},$$

pour tout entier $k \geq 0$.

On considère maintenant le temps avant qu'un événement de mutation ne se produise sur une lignée ancestrale. Soit τ_{mut} , ce temps en nombre de générations et T_{mut} , ce temps correspondant en nombre de $2N$ générations. Alors τ_{mut} est de loi Géo (μ), et on a donc

$$\begin{aligned} \mathbb{P}(T_{mut} > t) &= \mathbb{P}(\tau_{mut} > \lfloor 2Nt \rfloor) = (1 - \mu)^{\lfloor 2Nt \rfloor} \\ &= \left(1 - \frac{\theta}{4N}\right)^{\lfloor 2Nt \rfloor}, \end{aligned}$$

d'où

$$\lim_{N \rightarrow +\infty} \mathbb{P}(T_{mut} > t) = e^{-\frac{\theta}{2}t},$$

pour tout nombre réel $t > 0$. La variable aléatoire T_{mut} suit donc asymptotiquement une loi $\text{Exp}\left(\frac{\theta}{2}\right)$.

En supposant, l'indépendance (voir le résultat 2.1.2), le temps avant qu'un événement de mutation et un seul ne se produise sur une lignée ancestrale parmi n suit asymptotiquement une loi $\text{Exp}\left(\frac{n\theta}{2}\right)$.

Survenance d'un événement de recombinaison

Sous le modèle de Wright-Fisher avec recombinaison, on suppose qu'un événement de recombinaison de deux parents se produit avec probabilité $\rho = \frac{R}{4N}$ pour donner un descendant recombinant, sinon le descendant est une copie d'un des parents choisi au hasard avec probabilité $1 - \rho$, et ce indépendamment des autres descendants.

On s'intéresse au temps avant qu'un événement de recombinaison n'affecte une lignée ancestrale. De façon analogue à l'arrivée d'un événement de mutation, soit τ_{rec} , le temps avant l'arrivée d'un événement de recombinaison sur une lignée ancestrale en nombre de générations et T_{rec} , ce temps en nombre de $2N$

génération. Alors τ_{rec} suit une loi Géo (ρ), et on a

$$\begin{aligned}\mathbb{P}(T_{rec} > t) &= \mathbb{P}(\tau_{rec} > \lfloor 2Nt \rfloor) = (1 - \rho)^{\lfloor 2Nt \rfloor} \\ &= \left(1 - \frac{R}{4N}\right)^{\lfloor 2Nt \rfloor},\end{aligned}$$

d'où

$$\lim_{N \rightarrow +\infty} \mathbb{P}(T_{rec} > t) = e^{-\frac{R}{2}t},$$

pour tout nombre réel $t > 0$. La variable aléatoire T_{rec} suit asymptotiquement une loi $\text{Exp}\left(\frac{R}{2}\right)$.

Ainsi, en supposant l'indépendance et en utilisant le résultat (2.1.2), le temps avant la première recombinaison affectant les lignées ancestrales d'un échantillon de taille n suit asymptotiquement une loi $\text{Exp}\left(\frac{nR}{2}\right)$.

Survenance d'un événement de conversion

Sous le modèle de Wright-Fisher avec conversion à deux loci, on suppose qu'un événement de conversion intrachromosomique affecte une séquence produite avec probabilité $c = \frac{C}{4N}$ par locus.

De façon analogue à ce qui a été fait précédemment, τ_{conv} , le temps avant un événement de conversion chromosomique à un des deux loci sur une séquence ancestrale en nombre de générations et T_{conv} ce temps correspondant en nombre de $2N$ générations, satisfont

$$\begin{aligned}\mathbb{P}(T_{conv} > t) &= \mathbb{P}(\tau_{conv} > \lfloor 2Nt \rfloor) = (1 - 2c)^{\lfloor 2Nt \rfloor} \\ &= \left(1 - \frac{2C}{4N}\right)^{\lfloor 2Nt \rfloor},\end{aligned}$$

d'où

$$\lim_{N \rightarrow +\infty} \mathbb{P}(T_{conv} > t) = e^{-Ct},$$

pour tout nombre réel $t > 0$. La variable T_{conv} suit asymptotiquement une loi $\text{Exp}(C)$.

Ainsi, en utilisant le résultat (2.1.2), et en supposant l'indépendance le temps avant la première conversion et une seule affectant le matériel ancestral d'un échantillon de taille n suit asymptotiquement une loi $\text{Exp}(nC)$.

Taux de changement du matériel ancestral

En supposant les hypothèses d'indépendance requises, le processus ancestral limite d'un échantillon pris dans une population qui suit le modèle de Wright-Fisher avec recombinaison, mutation et conversion à deux loci est caractérisé par les taux de changement suivants lorsque la taille est n : $\frac{n(n-1)}{2}$ comme taux de coalescence, $\frac{nR}{2}$ comme taux de recombinaison, et nC comme taux de conversion, pour un taux total de $\frac{n(n-1)}{2} + \frac{nR}{2} + nC$.

2.3. PROCESSUS ANCESTRAL SOUS LE MODÈLE DE MORAN

Processus ancestral exact

On considère un échantillon de taille n dans une population de taille $2N$ qui évolue sous un modèle de Moran à deux loci. A chaque instant, deux individus sont choisis au hasard avec remplacement et ils produisent un descendant qui remplace un individu choisi au hasard. On considère les événements suivants :

- Le descendant est un recombinant des deux parents avec probabilité $\frac{R}{2N}$;
- Le descendant subit une conversion avec probabilité $\frac{C}{N}$.

On imagine que les individus occupent des positions distinctes numérotées de 1 à $2N$ et que les individus échantillonnés occupent les positions de 1 à n . La figure 2.1 illustre le matériel ancestral d'un échantillon de taille $n = 4$ (désigné par les loci en forme de cercle) dans une population de taille $2N = 6$. Dans cet exemple, du présent au passé, l'échantillon connaît des événements de coalescence, de conversion, de mutation et de recombinaison jusqu'à la dernière lignée restante ancestrale aux deux loci de l'échantillon qu'on notera MRCA (most recent common ancestor).

Plus généralement, en remontant les lignées de l'échantillon un instant en arrière, il y a un événement de coalescence pure de i et j , pour $i, j = 1, \dots, n$ et $i \neq j$, si un des deux cas suivants se produit : le descendant est une copie exacte de j , ce qui a la probabilité $\frac{1}{2N} \left(1 - \frac{R}{2N}\right) \left(1 - \frac{C}{N}\right)$, et il remplace l'individu situé à la position occupée par l'individu i , ce qui a la probabilité $\frac{1}{2N}$, ou vice versa. Ceci aboutit à la probabilité de coalescence pure suivante :

$$\frac{1}{2N^2} \left(1 - \frac{R}{2N}\right) \left(1 - \frac{C}{N}\right) = \frac{1}{2N^2} \left(1 + O\left(\frac{1}{N}\right)\right) \leq \frac{1}{2N^2}.$$

Aussi, il y a un événement de recombinaison pure de i en un instant en arrière, pour $i = 1, \dots, n$, si le descendant produit est un recombinant sans conversion de $k, l = n + 1, \dots, 2N$ et $k \neq l$, ce qui se produit avec la probabilité $\frac{R}{2N} \left(1 - \frac{C}{N}\right) \frac{2N-n}{2N} \frac{2N-n-1}{2N}$, et il remplace l'individu situé à la position de i , ce qui arrive avec la probabilité $\frac{1}{2N}$. Cela mène à la probabilité de recombinaison pure suivante :

$$\frac{R(2N-n)(2N-n-1)}{(2N)^4} \left(1 - \frac{C}{N}\right) = \frac{R}{(2N)^2} \left(1 + O\left(\frac{1}{N}\right)\right) \leq \frac{R}{(2N)^2}.$$

Enfin, il y a un événement de conversion pure de i en un instant en arrière, pour $i = 1, \dots, n$, si le descendant est un produit de conversion sans recombinaison de k qui n'est pas dans l'échantillon, pour $k = n + 1, \dots, 2N$, ce qui se produit avec la probabilité $\frac{C}{N} \left(1 - \frac{R}{2N}\right) \frac{2N-n}{2N}$, et il remplace l'individu à la position de i , ce qui a la probabilité $\frac{1}{2N}$. Cela mène à la probabilité de conversion pure suivante :

$$\frac{C(2N-n)}{4N^3} \left(1 - \frac{R}{2N}\right) = \frac{C}{2N^2} \left(1 + O\left(\frac{1}{N}\right)\right) \leq \frac{C}{2N^2}.$$

De plus, notons que les probabilités d'événements pures en un instant en arrière pour un échantillon de taille n sont des fonctions d'ordre $\frac{1}{N^2}$. En revanche, les probabilités d'événements simultanés de coalescence, recombinaison ou conversion affectant les lignées de l'échantillon en un instant en arrière sont des fonctions d'ordre $\frac{1}{N^3}$.

Approximation du processus ancestral exact

Etant donné un échantillon de taille n , le nombre total d'événements de coalescence pure est $\frac{n(n-1)}{2}$, alors que le nombre est n pour les événements de recombinaison pure et pour les événements de conversion pure. Ainsi, la probabilité totale de changement en un instant en arrière est donnée par

$$p_n = \frac{\lambda_n}{2N^2} + O\left(\frac{1}{N^3}\right),$$

où

$$\lambda_n = \frac{n(n-1+R+2C)}{2}.$$

Cette dernière quantité représente le taux de changement total à la limite lorsque la taille de la population est grande avec une unité de temps de $2N^2$ événements de naissance-mort. La probabilité conditionnelle de chacun des événements purs de coalescence, de recombinaison et de conversion est alors

$$\mathbb{P}(Coa_n) = \frac{1}{\lambda_n} + O\left(\frac{1}{N}\right),$$

$$\mathbb{P}(Rec_n) = \frac{R}{2\lambda_n} + O\left(\frac{1}{N}\right),$$

et,

$$\mathbb{P}(Conv_n) = \frac{C}{\lambda_n} + O\left(\frac{1}{N}\right),$$

respectivement, alors que la probabilité conditionnelle des événements simultanés est

$$\mathbb{P}(Sim_n) = O\left(\frac{1}{N}\right).$$

Lorsque $N \rightarrow \infty$, la probabilité conditionnelle des événements simultanés est négligeable.

Soit τ_n le temps en unité de temps jusqu'au prochain changement d'un échantillon de taille n causé par un événement de coalescence, de recombinaison ou de conversion. La variable τ_n suit une loi géométrique de paramètre p_n . Maintenant,

soit T_n le temps correspondant en nombre de $2N^2$ événements de naissance-mort.

On a

$$\begin{aligned}\mathbb{P}(T_n > y) &= \mathbb{P}(\tau_n > \lfloor 2N^2 y \rfloor) \\ &= \left(1 - \frac{\lambda_n}{2N^2} - O\left(\frac{1}{N^3}\right)\right)^{\lfloor 2N^2 y \rfloor} \longrightarrow e^{-\lambda_n y},\end{aligned}$$

lorsque $N \rightarrow \infty$, pour tout nombre réel $y > 0$. La variable T_n converge en distribution vers une variable aléatoire exponentielle de paramètre λ_n . A la fin de ce temps, un événement de coalescence, de recombinaison, ou de conversion se produit avec la probabilité $\frac{n(n-1)}{2\lambda_n}$, $\frac{nR}{2\lambda_n}$ ou $\frac{nC}{\lambda_n}$, respectivement. Cela signifie que les taux de coalescence, de recombinaison et de conversion sont donnés par $\frac{n(n-1)}{2}$, $\frac{nR}{2}$, et nC , respectivement.

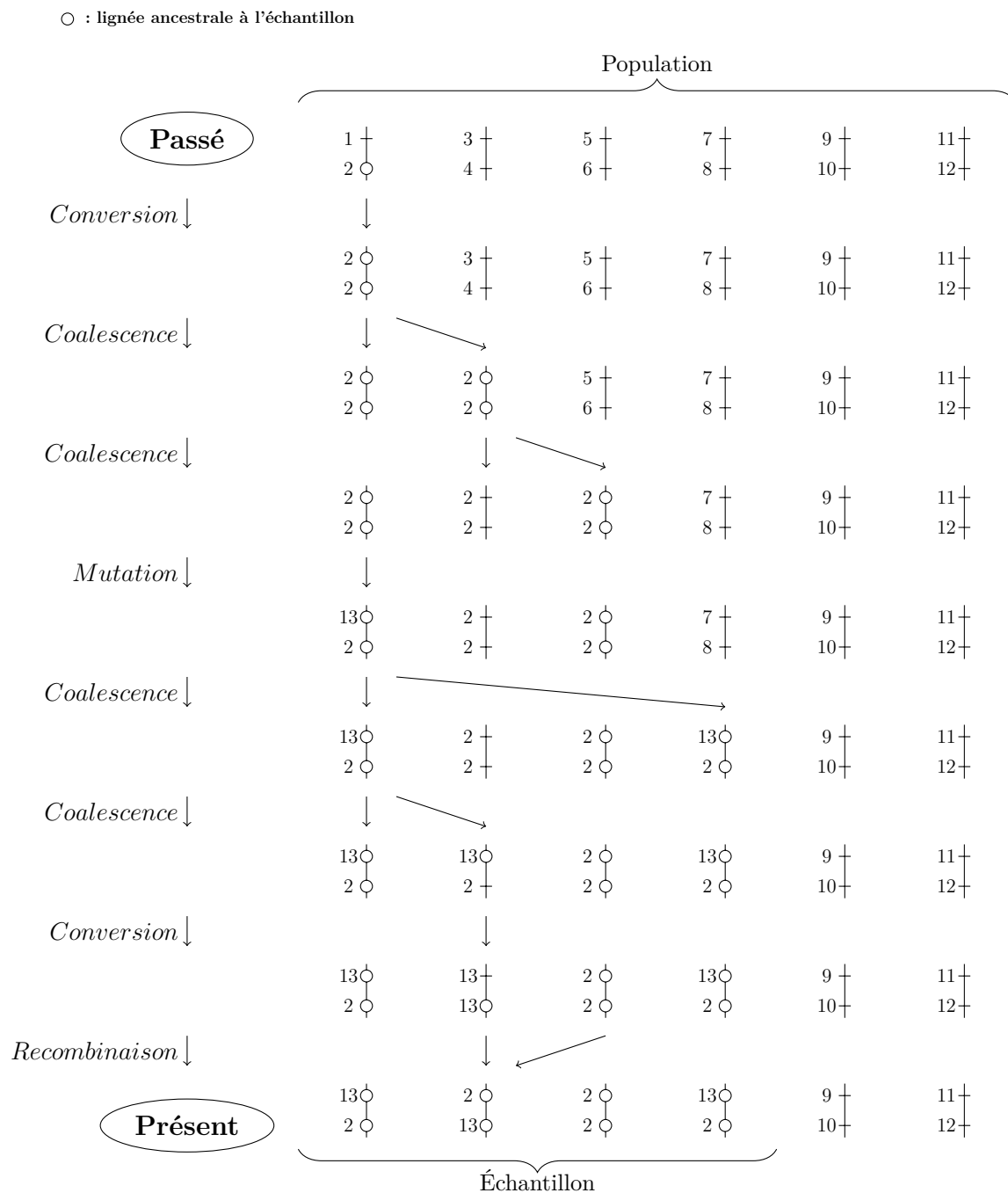


FIGURE 2.1. Matériel ancestral d'un échantillon de taille $n = 4$ dans une population de taille $2N = 6$ qui suit le modèle de Moran.

Chapitre 3

MESURE DE LIAISON

Ce chapitre a d'abord comme objectif de calculer les covariances dans l'expression de la mesure de déséquilibre de liaison à deux loci σ_d^2 du chapitre 1 à l'aide du processus de coalescence exact avec recombinaison. Ensuite, ces mêmes covariances sont calculées en utilisant l'approximation du processus de coalescence SMC (sequentially Markov chain) avec recombinaison proposée par McVean et Cardin (2005). Nous concluons par une comparaison des mesures ainsi obtenues, étant donné que le résultat pour la troisième covariance diffère de celui trouvé dans l'article original.

3.1. COVARIANCES EXACTES DE TEMPS DE COALESCENCE À DEUX LOCI AVEC RECOMBINAISON

Griffiths (1981) a considéré le modèle à deux loci et à partir de ces résultats les covariances entre les temps de coalescence à deux loci pour des séquences échantillonnées ont été déduites (Pluzhnikov et Donnelly 1996). Dans ce chapitre, dans le but d'étudier la variabilité génétique (c'est-à-dire l'équation (1.2.2)), on revoit ces covariances pour des paires de lignées à deux loci et un échantillon de deux, trois ou quatre séquences. On considère l'extension du processus de coalescence en présence de recombinaison, mais en absence de conversion, telle que décrite par Hudson (1983,1990). On rappelle que le processus ancestral forme une chaîne de Markov, et on considère les états accessibles et les probabilités de transition à partir de deux séquences à deux loci (voir le tableau 3.1 et la figure 3.1 tirée de Simonsen et Churchill (1997)).

Pour deux loci (l) et (m), soient $n^{(l)}$, $n^{(m)}$ et $n^{(lm)}$ les nombres de séquences ancestrales à un échantillon au locus (l) seulement, au locus (m) seulement et aux deux loci (l) et (m), respectivement. Le vecteur $\underline{n} = (n^{(l)}, n^{(m)}, n^{(lm)})$ représente l'état du matériel ancestral à un moment donné dans le passé. Le nombre total de séquences ancestrales est alors donné par $n = n^{(l)} + n^{(m)} + n^{(lm)}$. Ici on se restreint au cas où $1 \leq n^{(l)} + n^{(lm)} \leq 2$ et $1 \leq n^{(m)} + n^{(lm)} \leq 2$.

En remontant le temps, chaque paire de séquences ancestrales coalesce au taux 1 et chaque séquence ancestrale recombine au taux $\frac{R}{2}$, indépendamment les unes des autres. Tout événement de coalescence change l'état du matériel ancestral alors que ceci n'est le cas pour un événement de recombinaison que s'il affecte une séquence ancestrale aux deux loci. Le taux de changement de l'état \underline{n} à l'état \underline{n}' est donné par

$$\lambda_{\underline{n}, \underline{n}'} = \begin{cases} \frac{n^{(l)}(n^{(l)}-1)}{2} + n^{(l)}n^{(lm)} & \text{si } \underline{n}' = (n^{(l)} - 1, n^{(m)}, n^{(lm)}), \\ \frac{n^{(m)}(n^{(m)}-1)}{2} + n^{(m)}n^{(lm)} & \text{si } \underline{n}' = (n^{(l)}, n^{(m)} - 1, n^{(lm)}), \\ \frac{n^{(lm)}(n^{(lm)}-1)}{2} & \text{si } \underline{n}' = (n^{(l)}, n^{(m)}, n^{(lm)} - 1), \\ \frac{R}{2}n^{(lm)} & \text{si } \underline{n}' = (n^{(l)} + 1, n^{(m)} + 1, n^{(lm)} - 1), \\ n^{(l)}n^{(m)} & \text{si } \underline{n}' = (n^{(l)} - 1, n^{(m)} - 1, n^{(lm)} + 1). \end{cases}$$



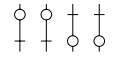
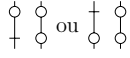





Le taux de changement total à partir de l'état \underline{n} est donc

$$\lambda_{\underline{n}} = \frac{n(n-1)}{2} + n^{(lm)}\frac{R}{2}.$$

La probabilité de transition de l'état \underline{n} à l'état \underline{n}' étant donné un changement de l'état \underline{n} (voir le tableau 3.1 et la figure 3.1) est alors donnée par

$$p_{\underline{n}, \underline{n}'} = \frac{\lambda_{\underline{n}, \underline{n}'}}{\lambda_{\underline{n}}}.$$

TABLEAU 3.1. Etats ancestraux et transitions d'état à partir de deux séquences à deux loci (l) et (m).

Etat	\underline{n}	Evénements possibles	Prochain état
	O : lignée ancestrale		
1 (Début)	(0,0,2) 	Rec Coa	2 9
2	(1,1,1) 	Coa Rec Coa Coa	1 3 4 5
3	(2,2,0) 	Coa Coa Coa	2 6 7
4	(1,0,1) 	Rec Coa	6 9
5	(0,1,1) 	Rec Coa	7 9
6	(2,1,0) 	Coa Coa	4 8
7	(1,2,0) 	Coa Coa	5 8
8	(1,1,0) 	Coa	9
9 (Fin)	(0,0,1) 	Etat absorbant	-

Le temps jusqu'au premier changement d'état, $T_{\underline{n}}$, est de loi exponentielle de paramètre $\lambda_{\underline{n}}$. Il s'ensuit que

$$\mathbb{E}[T_{\underline{n}}] = \frac{1}{\lambda_{\underline{n}}},$$

$$\mathbb{E}[T_{\underline{n}}^2] = \frac{2}{\lambda_{\underline{n}}^2}.$$

On définit : $T_{\underline{n}}^{(l)}$ le temps jusqu'au MRCA (most recent common ancestor) au locus (l) à partir de l'état \underline{n} , soit le temps jusqu'au dernier événement de coalescence menant à la dernière lignée ancestrale au locus (l) de l'échantillon ; et $T_{\underline{\gamma}}^{(l)}$ le temps jusqu'au MRCA au locus (l) à partir de l'état $\underline{\gamma}$. Ici $\underline{\gamma}$ représente un vecteur aléatoire indépendant de $T_{\underline{n}}$ qui prend la valeur \underline{n}' avec probabilité $p_{\underline{n},\underline{n}'}$.

On a

$$T_{\underline{n}}^{(l)} = T_{\underline{n}} + T_{\underline{\gamma}}^{(l)}, \quad (3.1.1)$$

d'où

$$\mathbb{E}[T_{\underline{\gamma}}^{(l)}] = \mathbb{E}[T_{\underline{n}}^{(l)}] - \mathbb{E}[T_{\underline{n}}] = 1 - \frac{1}{\lambda_{\underline{n}}},$$

si $n^{(l)} + n^{(lm)} = 2$. Dans le cas où $n^{(l)} + n^{(lm)} = 1$, le MRCA au locus (l) est trouvé, et ainsi on a $T_{\underline{n}}^{(l)} = 0$. De façon analogue, on a :

$$T_{\underline{n}}^{(m)} = T_{\underline{n}} + T_{\underline{\gamma}}^{(m)}, \quad (3.1.2)$$

d'où

$$\mathbb{E}[T_{\underline{\gamma}}^{(m)}] = 1 - \frac{1}{\lambda_{\underline{n}}},$$

si $n^{(m)} + n^{(lm)} = 2$, et $T_{\underline{n}}^{(m)} = 0$ si $n^{(m)} + n^{(lm)} = 1$ puisque le MRCA au locus (m) est alors trouvé.

En multipliant (3.1.1) et (3.1.2), puis en prenant l'espérance on trouve

$$T_{\underline{n}}^{(l)} \times T_{\underline{n}}^{(m)} = T_{\underline{\gamma}}^{(l)} \times T_{\underline{\gamma}}^{(m)} + T_{\underline{n}}^2 + T_{\underline{n}}T_{\underline{\gamma}}^{(l)} + T_{\underline{n}}T_{\underline{\gamma}}^{(m)}. \quad (3.1.3)$$

On a donc

$$\begin{aligned} \mathbb{E}[T_{\underline{n}}^{(l)} \times T_{\underline{n}}^{(m)}] &= \mathbb{E}[T_{\underline{\gamma}}^{(l)} \times T_{\underline{\gamma}}^{(m)} + T_{\underline{n}}^2 + T_{\underline{n}}T_{\underline{\gamma}}^{(l)} + T_{\underline{n}}T_{\underline{\gamma}}^{(m)}] \\ &= \mathbb{E}[T_{\underline{\gamma}}^{(l)} \times T_{\underline{\gamma}}^{(m)}] + \mathbb{E}[T_{\underline{n}}^2] + \mathbb{E}[T_{\underline{n}}]\mathbb{E}[T_{\underline{\gamma}}^{(l)}] + \mathbb{E}[T_{\underline{n}}]\mathbb{E}[T_{\underline{\gamma}}^{(m)}] \\ &= \mathbb{E}[T_{\underline{\gamma}}^{(l)} \times T_{\underline{\gamma}}^{(m)}] + \frac{2}{\lambda_{\underline{n}}^2} + \frac{1}{\lambda_{\underline{n}}} - \frac{1}{\lambda_{\underline{n}}^2} + \frac{1}{\lambda_{\underline{n}}} - \frac{1}{\lambda_{\underline{n}}^2} \\ &= \mathbb{E}[T_{\underline{\gamma}}^{(l)} \times T_{\underline{\gamma}}^{(m)}] + \frac{1}{2\lambda_{\underline{n}}}, \end{aligned}$$

c'est-à-dire

$$\mathbb{E}[T_{\underline{n}}^{(l)} \times T_{\underline{n}}^{(m)}] = \sum_{\underline{n}'} p_{\underline{n}, \underline{n}'} \mathbb{E}[T_{\underline{n}'}^{(l)} \times T_{\underline{n}'}^{(m)}] + \frac{1}{2\lambda_{\underline{n}}}, \quad (3.1.4)$$

lorsque $n^{(l)} + n^{(lm)} = n^{(m)} + n^{(lm)} = 2$, où $p_{\underline{n}, \underline{n}'}$ est la probabilité de transition de l'état \underline{n} à \underline{n}' (voir la figure 3.1, dans laquelle les états sont numérotés de 1 à 9 non soulignés).

L'équation générale (3.1.4) permet de calculer l'espérance du produit des temps jusqu'au MRCA aux deux loci pour chaque état et donc leur covariance donnée par

$$\begin{aligned} \mathbb{C}_{\underline{n}} &= \mathbb{E}[T_{\underline{n}}^{(l)} \times T_{\underline{n}}^{(m)}] - \mathbb{E}[T_{\underline{n}}^{(l)}] \times \mathbb{E}[T_{\underline{n}}^{(m)}] \\ &= \mathbb{E}[T_{\underline{n}}^{(l)} \times T_{\underline{n}}^{(m)}] - 1. \end{aligned} \quad (3.1.5)$$

En développant le système d'équations ci-dessus, on trouve que

$$\begin{aligned} \mathbb{E}[T_{\underline{1}}^{(l)} \times T_{\underline{1}}^{(m)}] &= \frac{R}{1+R} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] + \frac{1}{1+R} \mathbb{E}[T_{\underline{9}}^{(l)} \times T_{\underline{9}}^{(m)}] + \frac{2}{1+R} \\ &= \frac{R}{1+R} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] + \frac{2}{1+R}, \end{aligned}$$

$$\begin{aligned} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] &= \frac{R}{6+R} \mathbb{E}[T_{\underline{3}}^{(l)} \times T_{\underline{3}}^{(m)}] + \frac{2}{6+R} \mathbb{E}[T_{\underline{4}}^{(l)} \times T_{\underline{4}}^{(m)}] \\ &+ \frac{2}{6+R} \mathbb{E}[T_{\underline{5}}^{(l)} \times T_{\underline{5}}^{(m)}] + \frac{2}{6+R} \mathbb{E}[T_{\underline{1}}^{(l)} \times T_{\underline{1}}^{(m)}] + \frac{2}{6+R} \\ &= \frac{R}{6+R} \mathbb{E}[T_{\underline{3}}^{(l)} \times T_{\underline{3}}^{(m)}] + \frac{2}{6+R} \mathbb{E}[T_{\underline{1}}^{(l)} \times T_{\underline{1}}^{(m)}] + \frac{2}{6+R}, \end{aligned}$$

$$\begin{aligned} \mathbb{E}[T_{\underline{3}}^{(l)} \times T_{\underline{3}}^{(m)}] &= \frac{2}{3} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] + \frac{1}{6} \mathbb{E}[T_{\underline{6}}^{(l)} \times T_{\underline{6}}^{(m)}] + \frac{1}{6} \mathbb{E}[T_{\underline{7}}^{(l)} \times T_{\underline{7}}^{(m)}] + \frac{1}{3} \\ &= \frac{2}{3} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] + \frac{1}{3}. \end{aligned}$$

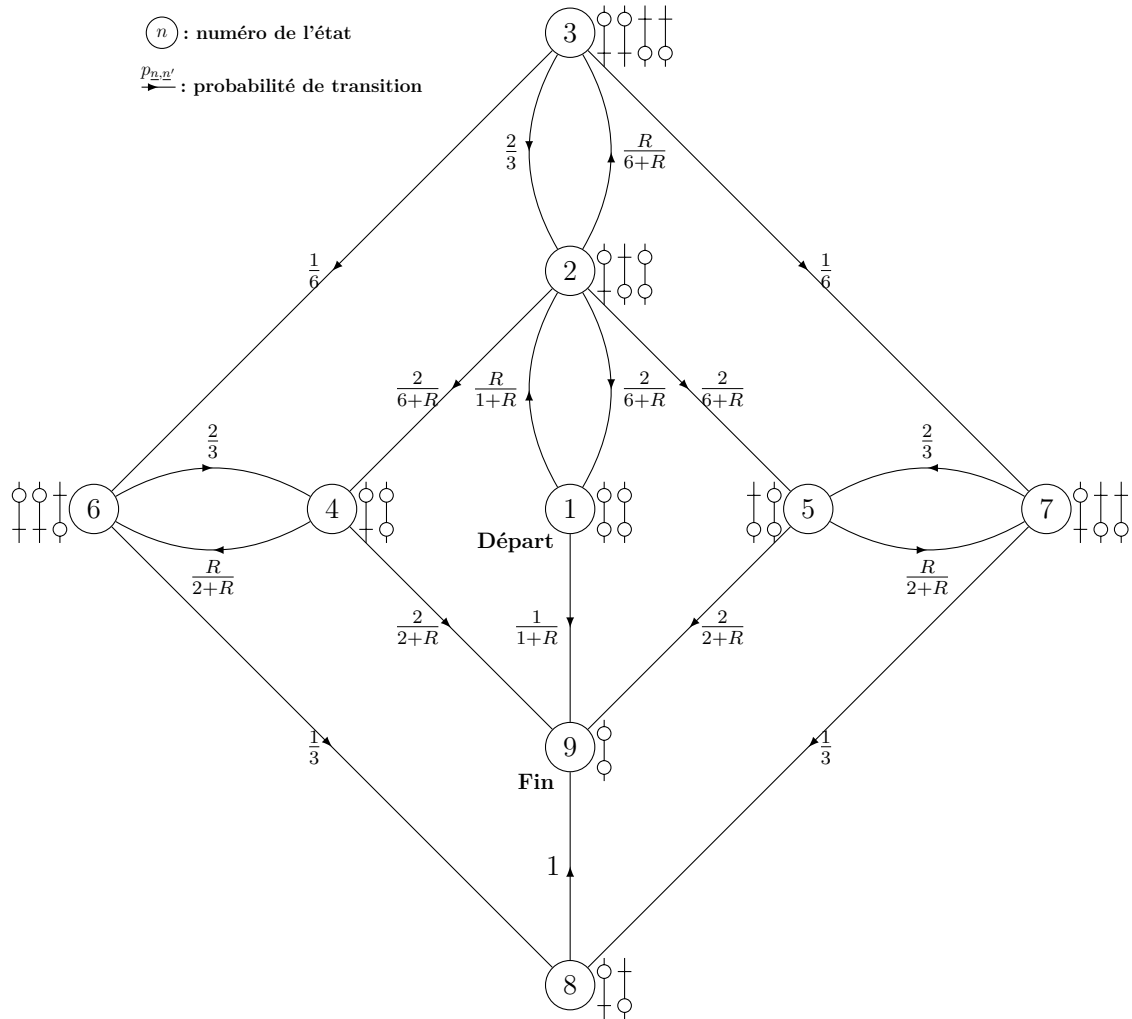


FIGURE 3.1. Diagramme de transitions d'état pour le processus de coalescence avec recombinaison. Les transitions possibles entre les 9 états ancestraux de deux séquences à deux loci sont indiquées par des flèches associées aux probabilités de transition.

Source: Simonsen et Churchill 1997.

Finalement par (3.1.5), on obtient que

$$\begin{aligned}
 (1 + R)\mathcal{C}_1 - R\mathcal{C}_2 &= 1, \\
 -2\mathcal{C}_1 + (6 + R)\mathcal{C}_2 - R\mathcal{C}_3 &= 0, \\
 -2\mathcal{C}_2 + 3\mathcal{C}_3 &= 0,
 \end{aligned}$$

dont la solution est :

$$\begin{aligned} \mathbb{C}_1 &= \frac{18 + R}{R^2 + 13R + 18}, \\ \mathbb{C}_2 &= \frac{6}{R^2 + 13R + 18}, \\ \mathbb{C}_3 &= \frac{4}{R^2 + 13R + 18}. \end{aligned}$$

3.2. COVARIANCES AVEC L'APPROXIMATION DU PROCESSUS DE COALESCENCE AVEC RECOMBINAISON DE McVEAN ET CARDIN

L'approximation du processus de coalescence avec recombinaison proposé par McVean et Cardin (2005), SMC (sequentially Markov chain), suggère une simple modification du processus ancestral tel que considéré par Hudson (1983). Si deux lignées ancestrales n'ont pas de matériel ancestral en commun, on considère qu'elles ne coalescent pas. Ceci a pour effet indirect de réduire le nombre d'événements de recombinaison. De plus, le SMC assume que toutes les recombinaisons se produisent à un point unique entre les deux loci considérés.

Ici cependant, l'espace d'états $\underline{n} = (n^{(l)}, n^{(m)}, n^{(lm)})$ du processus exact doit être augmenté, pour tenir compte du matériel ancestral de séquences échantillonnées non seulement aux loci considérés mais aux deux loci. On a cinq types de séquences : des séquences considérées au locus (l) ancestrales seulement au locus (l), ou ancestrales aux loci (l) et (m) ; des séquences considérées au locus (m) ancestrales seulement au locus (m) ou ancestrales aux loci (l) et (m) ; et finalement des séquences qui sont considérées aux loci (l) et (m) où elles sont ancestrales. Les nombres de ces types sont dénotés par $n^{(l)}$, $n_*^{(l)}$, $n^{(m)}$, $n_*^{(m)}$, $n^{(lm)}$, respectivement. Le vecteur $\underline{n} = (n^{(l)}, n_*^{(l)}, n^{(m)}, n_*^{(m)}, n^{(lm)})$ représente l'état du matériel ancestral (voir la figure 3.2), et le nombre total de séquences ancestrales est alors donné par $n = n^{(l)} + n_*^{(l)} + n^{(m)} + n_*^{(m)} + n^{(lm)}$.

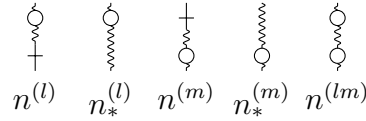


FIGURE 3.2. Les cinq types de séquences ancestrales du SMC.

Le taux de changement de l'état \underline{n} à l'état \underline{n}' est donné par

$$\lambda_{\underline{n}, \underline{n}'}^* = \begin{cases} \frac{n^{(l)}(n^{(l)}-1)}{2} + n^{(l)}(n_*^{(l)} + n^{(lm)}) & \text{si } \underline{n}' = (n^{(l)} - 1, n_*^{(l)}, n^{(m)}, n_*^{(m)}, n^{(lm)}), \\ \frac{n^{(m)}(n^{(m)}-1)}{2} + n^{(m)}(n_*^{(m)} + n^{(lm)}) & \text{si } \underline{n}' = (n^{(l)}, n_*^{(l)}, n^{(m)} - 1, n_*^{(m)}, n^{(lm)}), \\ \frac{n_*^{(l)}(n_*^{(l)}-1)}{2} + n_*^{(l)} n^{(lm)} & \text{si } \underline{n}' = (n^{(l)}, n_*^{(l)} - 1, n^{(m)}, n_*^{(m)}, n^{(lm)}), \\ \frac{n_*^{(m)}(n_*^{(m)}-1)}{2} + n_*^{(m)} n^{(lm)} & \text{si } \underline{n}' = (n^{(l)}, n_*^{(l)}, n^{(m)}, n_*^{(m)} - 1, n^{(lm)}), \\ \frac{n^{(lm)}(n^{(lm)}-1)}{2} & \text{si } \underline{n}' = (n^{(l)}, n_*^{(l)}, n^{(m)}, n_*^{(m)}, n^{(lm)} - 1), \\ \frac{R}{2} n_*^{(l)} & \text{si } \underline{n}' = (n^{(l)} + 1, n_*^{(l)} - 1, n^{(m)}, n_*^{(m)}, n^{(lm)}), \\ \frac{R}{2} n_*^{(m)} & \text{si } \underline{n}' = (n^{(l)}, n_*^{(l)}, n^{(m)} + 1, n_*^{(m)} - 1, n^{(lm)}), \\ \frac{R}{2} n^{(lm)} & \text{si } \underline{n}' = (n^{(l)} + 1, n_*^{(l)}, n^{(m)} + 1, n_*^{(m)}, n^{(lm)} - 1), \\ n^{(l)} n_*^{(m)} & \text{si } \underline{n}' = (n^{(l)} - 1, n_*^{(l)}, n^{(m)}, n_*^{(m)} - 1, n^{(lm)} + 1), \\ n_*^{(l)} n^{(m)} & \text{si } \underline{n}' = (n^{(l)}, n_*^{(l)} - 1, n^{(m)} - 1, n_*^{(m)}, n^{(lm)} + 1), \\ n_*^{(l)} n_*^{(m)} & \text{si } \underline{n}' = (n^{(l)}, n_*^{(l)} - 1, n^{(m)}, n_*^{(m)} - 1, n^{(lm)} + 1), \end{cases}$$

pour un taux de changement total

$$\lambda_{\underline{n}}^* = \frac{n(n-1) - 2n^{(l)}n^{(m)}}{2} + (n_*^{(l)} + n_*^{(m)} + n^{(lm)}) \frac{R}{2}.$$

La probabilité de transition de l'état \underline{n} à l'état \underline{n}' (voir les figures 3.3, 3.4 et 3.5) est donnée par

$$p_{\underline{n}, \underline{n}'}^* = \frac{\lambda_{\underline{n}, \underline{n}'}^*}{\lambda_{\underline{n}}^*}.$$

L'équation générale (3.1.4) s'applique toujours avec le taux de changement total et les probabilités de transition ci-dessus, à savoir

$$\mathbb{E}[T_{\underline{n}}^{(l)} \times T_{\underline{n}}^{(m)}] = \begin{cases} \sum_{\underline{n}'} p_{\underline{n}, \underline{n}'}^* \mathbb{E}[T_{\underline{n}'}^{(l)} \times T_{\underline{n}'}^{(m)}] + \frac{2}{\lambda_{\underline{n}}^*}, & \text{lorsque } n^{(l)} + n_*^{(l)} + n^{(lm)} = 2 \\ & \text{et } n^{(m)} + n_*^{(m)} + n^{(lm)} = 2, \\ 0 & \text{lorsque } n^{(l)} + n_*^{(l)} + n^{(lm)} = 1 \\ & \text{ou } n^{(m)} + n_*^{(m)} + n^{(lm)} = 1. \end{cases}$$

Pour trouver la première covariance, on considère les transitions possibles et les probabilités associées pour les 9 états ancestraux de deux séquences considérées aux deux loci (l) et (m), c'est-à-dire à partir de l'état $\underline{n} = (0, 0, 0, 0, 2)$. On trouve les espérances ci-dessous pour les différents états (voir la figure 3.3 pour la notation des états de 1 à 9 non soulignées et les probabilités de transition).

Pour l'état $\underline{3}$, on a $\lambda_{\underline{3}} = 2$ et

$$\begin{aligned} \mathbb{E}[T_{\underline{3}}^{(l)} \times T_{\underline{3}}^{(m)}] &= \frac{1}{2} \mathbb{E}[T_{\underline{6}}^{(l)} \times T_{\underline{6}}^{(m)}] + \frac{1}{2} \mathbb{E}[T_{\underline{7}}^{(l)} \times T_{\underline{7}}^{(m)}] + 1 \\ &= 1. \end{aligned}$$

Pour l'état $\underline{2}$, on a $\lambda_{\underline{2}} = \frac{4+R}{2}$ et

$$\begin{aligned} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] &= \frac{R}{4+R} \mathbb{E}[T_{\underline{3}}^{(l)} \times T_{\underline{3}}^{(m)}] + \frac{2}{4+R} \mathbb{E}[T_{\underline{4}}^{(l)} \times T_{\underline{4}}^{(m)}] \\ &+ \frac{2}{4+R} \mathbb{E}[T_{\underline{5}}^{(l)} \times T_{\underline{5}}^{(m)}] + \frac{4}{4+R} \\ &= \frac{R}{4+R} \mathbb{E}[T_{\underline{3}}^{(l)} \times T_{\underline{3}}^{(m)}] + \frac{4}{4+R} \\ &= 1. \end{aligned}$$

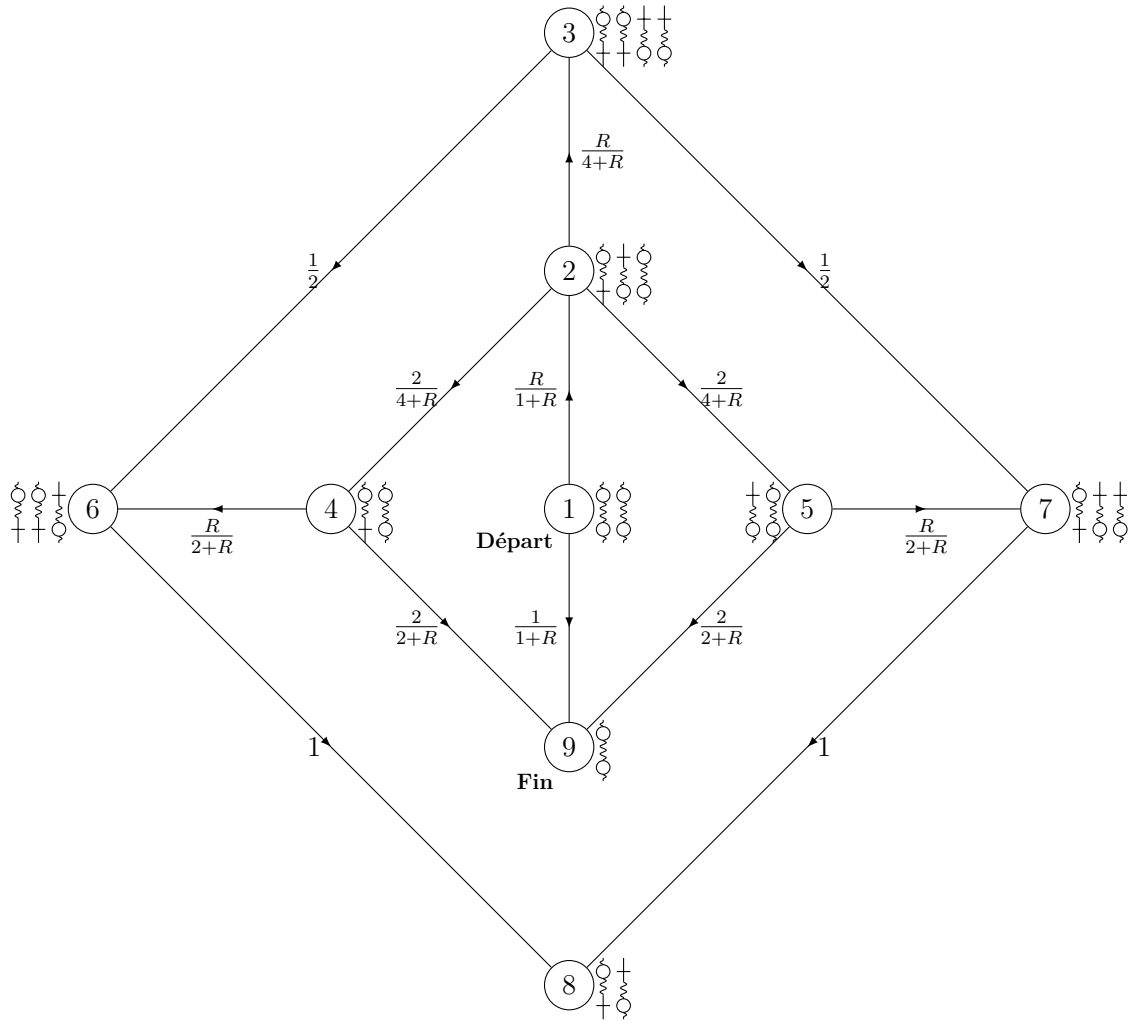


FIGURE 3.3. Diagramme de transitions d'état pour le processus de coalescence avec recombinaison SMC. Les transitions possibles entre les 9 états ancestraux de deux séquences échantillonnées considérées à deux loci sont indiquées par des flèches associées aux probabilités de transition.

Finalement, pour l'état $\underline{1}$, on a $\lambda_{\underline{1}} = 1 + R$ et

$$\begin{aligned}
 \mathbb{E}[T_{\underline{1}}^{(l)} \times T_{\underline{1}}^{(m)}] &= \frac{R}{1+R} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] + \frac{1}{1+R} \mathbb{E}[T_{\underline{9}}^{(l)} \times T_{\underline{9}}^{(m)}] + \frac{2}{1+R} \\
 &= \frac{R}{1+R} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] + \frac{2}{1+R} \\
 &= \frac{2+R}{1+R}.
 \end{aligned}$$

D'après l'équation (3.1.5), on a alors

$$\mathbb{C}^*_{\underline{1}} = \mathbb{E}[T_{\underline{1}}^{(l)} \times T_{\underline{1}}^{(m)}] - 1 = \frac{1}{1+R}.$$

Maintenant pour la deuxième covariance, dont l'état initial est constitué de trois séquences échantillonnées, une considérée à chacun des loci et une considérée aux deux loci, c'est-à-dire $\underline{n} = (0, 1, 0, 1, 1)$, la figure 3.4 illustre les transitions possibles et les probabilités associées pour les 20 états ancestraux.

On trouve les espérances successives suivantes :

$$\begin{aligned} \mathbb{E}[T_{\underline{3}'}^{(l)} \times T_{\underline{3}'}^{(m)}] &= \mathbb{E}[T_{\underline{3}''}^{(l)} \times T_{\underline{3}''}^{(m)}] \\ &= \frac{4}{8+R} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] + \frac{R}{8+R} \mathbb{E}[T_{\underline{3}}^{(l)} \times T_{\underline{3}}^{(m)}] + \frac{4}{8+R}, \\ &= 1, \end{aligned}$$

puis

$$\begin{aligned} \mathbb{E}[T_{\underline{2}'}^{(l)} \times T_{\underline{2}'}^{(m)}] &= \mathbb{E}[T_{\underline{2}''}^{(l)} \times T_{\underline{2}''}^{(m)}] \\ &= \frac{1}{3+R} \mathbb{E}[T_{\underline{1}}^{(l)} \times T_{\underline{1}}^{(m)}] + \frac{R}{2(3+R)} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] \\ &\quad + \frac{R}{2(3+R)} \mathbb{E}[T_{\underline{3}'}^{(l)} \times T_{\underline{3}'}^{(m)}] + \frac{2}{3+R} \\ &= \frac{(R+2)^2}{(R+1)(R+3)}, \end{aligned}$$

puis

$$\begin{aligned} \mathbb{E}[T_{\underline{3}^{(3)}}^{(l)} \times T_{\underline{3}^{(3)}}^{(m)}] &= \frac{2}{5+R} \mathbb{E}[T_{\underline{2}'}^{(l)} \times T_{\underline{2}'}^{(m)}] + \frac{1}{5+R} \mathbb{E}[T_{\underline{2}}^{(l)} \times T_{\underline{2}}^{(m)}] \\ &\quad + \frac{R}{5+R} \mathbb{E}[T_{\underline{3}'}^{(l)} \times T_{\underline{3}'}^{(m)}] + \frac{2}{5+R} \\ &= \frac{R^3 + 9R^2 + 23R + 17}{(R+1)(R+3)(R+5)}, \end{aligned}$$

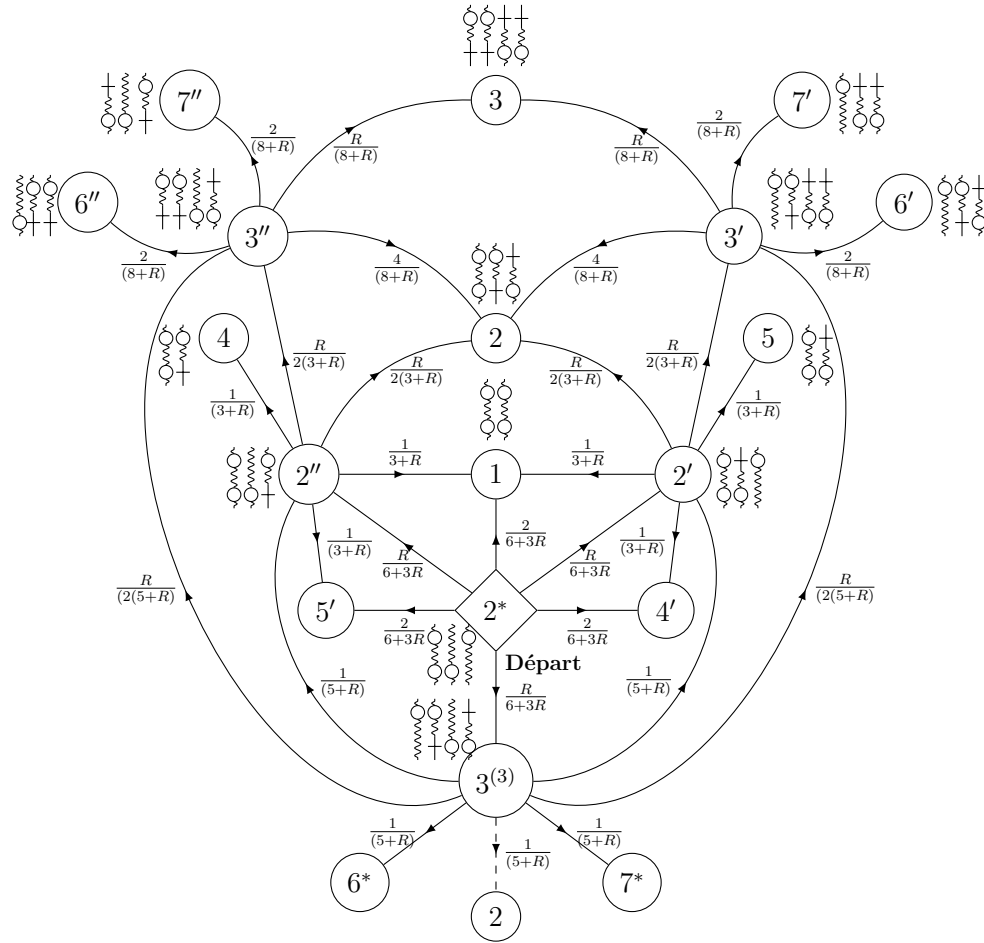


FIGURE 3.4. Diagramme de transitions d'état pour le processus de coalescence avec recombinaison SMC. Les transitions possibles entre les 20 états ancestraux de trois séquences échantillonnées, une considérée au locus (l), une considérée au locus (m) et une considérée au deux loci, sont indiquées par des flèches associées aux probabilités de transition.

et finalement

$$\begin{aligned}
 \mathbb{E}[T_{2^*}^{(l)} \times T_{2^*}^{(m)}] &= \frac{2}{3(2+R)} \mathbb{E}[T_{1}^{(l)} \times T_{1}^{(m)}] + \frac{R}{3(2+R)} \mathbb{E}[T_{2''}^{(l)} \times T_{2''}^{(m)}] \\
 &+ \frac{R}{3(2+R)} \mathbb{E}[T_{2'}^{(l)} \times T_{2'}^{(m)}] + \frac{R}{3(2+R)} \mathbb{E}[T_{3^{(3)}}^{(l)} \times T_{3^{(3)}}^{(m)}] \\
 &+ \frac{4}{3(2+R)}.
 \end{aligned}$$

Ainsi, on obtient que

$$\begin{aligned} \mathbb{C}_{\underline{2}}^* &= \mathbb{E}\left[T_{\underline{2}^*}^{(l)} \times T_{\underline{2}^*}^{(m)}\right] - 1 \\ &= \frac{30 + 4R(7 + R)}{3(1 + R)(2 + R)(3 + R)(5 + R)}. \end{aligned}$$

Finalement, pour la troisième covariance, les transitions possibles et les probabilités associées pour les 17 états ancestraux de deux séquences échantillonnées considérées à un locus et deux autres considérées à l'autre locus sont illustrées à la figure 3.5.

On trouve les espérances suivantes :

$$\begin{aligned} \mathbb{E}\left[T_{\underline{3}^{(7)}}^{(l)} \times T_{\underline{3}^{(7)}}^{(m)}\right] &= \mathbb{E}\left[T_{\underline{3}^{(8)}}^{(l)} \times T_{\underline{3}^{(8)}}^{(m)}\right] \\ &= \frac{4}{6 + R} \mathbb{E}\left[T_{\underline{2}'}^{(l)} \times T_{\underline{2}'}^{(m)}\right] + \frac{R}{6 + R} \mathbb{E}\left[T_{\underline{3}'}^{(l)} \times T_{\underline{3}'}^{(m)}\right] \\ &\quad + \frac{2}{6 + R} \\ &= \frac{(2 + R)(11 + 8R + R^2)}{(1 + R)(3 + R)(6 + R)}, \end{aligned}$$

puis

$$\begin{aligned} \mathbb{E}\left[T_{\underline{3}^{(5)}}^{(l)} \times T_{\underline{3}^{(5)}}^{(m)}\right] &= \mathbb{E}\left[T_{\underline{3}^{(6)}}^{(l)} \times T_{\underline{3}^{(6)}}^{(m)}\right] \\ &= \frac{R}{3(4 + R)} \mathbb{E}\left[T_{\underline{3}^{(7)}}^{(l)} \times T_{\underline{3}^{(7)}}^{(m)}\right] + \frac{4}{3(4 + R)} \mathbb{E}\left[T_{\underline{2}'}^{(l)} \times T_{\underline{2}'}^{(m)}\right] \\ &\quad + \frac{2R}{3(4 + R)} \mathbb{E}\left[T_{\underline{3}^{(3)}}^{(l)} \times T_{\underline{3}^{(3)}}^{(m)}\right] + \frac{4}{3(4 + R)} \mathbb{E}\left[T_{\underline{2}^*}^{(l)} \times T_{\underline{2}^*}^{(m)}\right] \\ &\quad + \frac{4}{3(4 + R)} \\ &= \frac{7920 + 17556R + 15160R^2 + 6667R^3 + 1575R^4 + 189R^5 + 9R^6}{9(1 + R)(2 + R)(3 + R)(4 + R)(5 + R)(6 + R)}, \end{aligned}$$

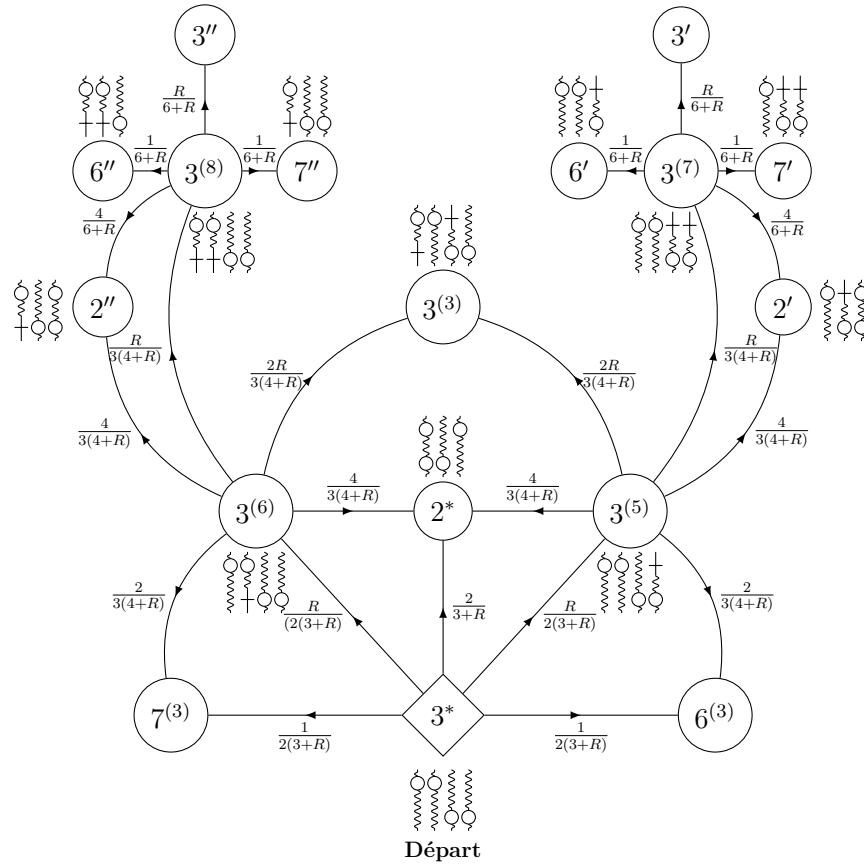


FIGURE 3.5. Diagramme de transitions d'état pour le processus de coalescence avec recombinaison SMC. Les transitions possibles entre les 17 états ancestraux de quatre séquences échantillonnées dont deux considérées au locus (l) et deux autres considérées au locus (m) sont indiquées par des flèches associées aux probabilités de transition.

et finalement

$$\begin{aligned}
 \mathbb{E}[T_{\underline{3}^*}^{(l)} \times T_{\underline{3}^*}^{(m)}] &= \frac{R}{3+R} \mathbb{E}[T_{\underline{3}^{(5)}}^{(l)} \times T_{\underline{3}^{(5)}}^{(m)}] + \frac{2}{3+R} \mathbb{E}[T_{\underline{2}^*}^{(l)} \times T_{\underline{2}^*}^{(m)}] \\
 &+ \frac{1}{3+R} \\
 &= \frac{7920 + 17820R + 15340R^2 + 6691R^3 + 1575R^4 + 189R^5 + 9R^6}{9(1+R)(2+R)(3+R)(4+R)(5+R)(6+R)}.
 \end{aligned}$$

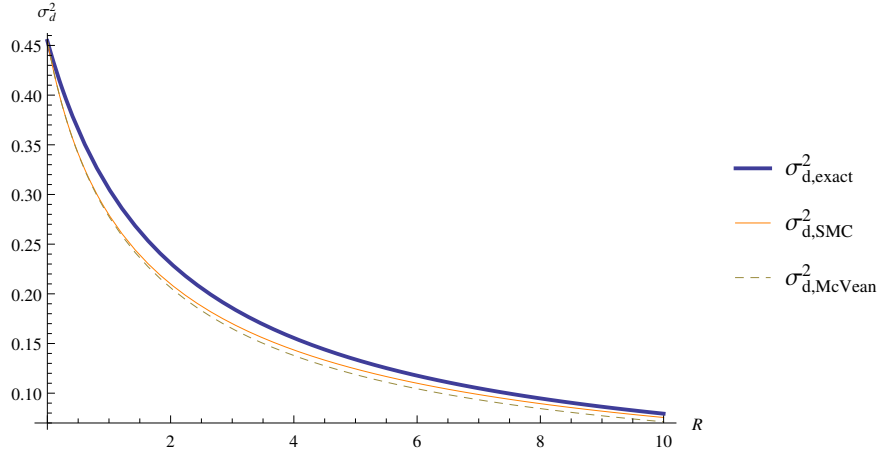


FIGURE 3.6. Décroissance du déséquilibre de liaison en fonction de la distance génétique approximée par σ_d^2 sous le processus ancestral exact ($\sigma_{d,exact}^2$), sous l'approximation SMC du processus ancestral calculée dans l'article de McVean et Cardin ($\sigma_{d,McVean}^2$), et sous l'approximation SMC corrigée ($\sigma_{d,SMC}^2$).

Ainsi, on trouve que

$$\begin{aligned} \mathbb{C}_{\underline{3}}^* &= \mathbb{E}\left[T_{\underline{3}^*}^{(l)} \times T_{\underline{3}^*}^{(m)}\right] - 1 \\ &= \frac{4(19R^3 + 181R^2 + 486R + 360)}{9(1+R)(2+R)(3+R)(4+R)(5+R)(6+R)}. \end{aligned}$$

Ce dernier résultat est différent du résultat trouvé dans l'article de McVean et Cardin (2005), soit :

$$\mathbb{C}_{\underline{3},McVean}^* = \frac{2(21600 + R(37080 + R(21690 + R(5017 + R(165 - R(91 + 9R))))))}{9(1+R)(2+R)(3+R)(4+R)((5+R)^2)((6+R)^2)}.$$

Les valeurs résultantes de σ_d^2 , approximant la mesure de déséquilibre de liaison, sous les deux modèles sont illustrées à la figure 3.6. La rectification de la

troisième covariance ne change pas significativement les conclusions. En éliminant les événements de coalescence entre deux lignées ancestrales qui n'ont pas de matériel ancestral en commun, le déséquilibre de liaison diminue légèrement comparé au processus de coalescence standard.

Chapitre 4

MESURE DE POLYMORPHISME À DEUX LOCI AVEC RECOMBINAISON ET CONVERSION

Dans le but d'estimer les paramètres de conversion, de mutation et de recombinaison dans les familles multigéniques, Innan (2002a,b) a calculé des espérances et des variances de mesures de variation inter-locus et intra-locus pour deux copies de gènes, par une approximation par un processus de diffusion (voir l'annexe A.3) sous le modèle de reproduction de Wright-Fisher. L'article original (Innan 2002a) ne mentionne pas les étapes pour trouver les changements de fréquences sous le modèle de Wright Fisher. Le modèle de reproduction considéré et décrit en détail dans ce chapitre est le modèle de Moran pour une population de taille $2N$ avec une unité de temps de $2N^2$ événements de naissance-mort. On imagine que les $2N$ séquences à deux loci représentant les individus de la population sont distribués sur $2N$ emplacements distincts. Chaque emplacement est occupé par une seule séquence. A chaque événement de naissance-mort, deux individus sont choisis au hasard avec remplacement et ils produisent une séquence qui remplace la séquence située à un emplacement choisi au hasard. On considère les événements suivants :

- Une recombinaison des deux parents se produit avec probabilité $\rho = \frac{R}{2N}$ pour donner un descendant recombinant, sinon le descendant est une copie d'un des parents avec probabilités $1 - \rho$.

- Une mutation se produit sur la séquence produite avec probabilité $\mu = \frac{\theta}{2N}$ par locus.
- Une conversion intrachromosomique se produit sur la séquence produite avec probabilité $c = \frac{C}{2N}$ par locus.

Dans le modèle neutre, on ne considère pas le facteur de sélection. Pour R , θ , C fixes et N grand, on considère que la probabilité de deux événements ou plus simultanés est négligeable (μ^2 , μc , $\mu \rho c$). Dans ce cas l'ordre dans lequel les facteurs d'évolution précédents sont considérés lors de la reproduction n'importe pas.

4.1. CHANGEMENTS DE FRÉQUENCES SOUS LE MODÈLE DE MORAN À DEUX LOCI

Dans le but d'estimer les paramètres de conversion, de mutation et de recombinaison dans les familles multigéniques, Innan (2002a) a calculé des espérances et des variances de mesures de variation inter-locus et intra-locus pour deux copies de gènes, par une approximation par un processus de diffusion du modèle de Wright-Fisher. Le modèle de Moran à deux loci avec deux allèles mène à la même approximation par une diffusion comme il sera montré ci-dessous. Dans l'application d'Innan à la famille multigénique *Amy* de la *Drosophila melanogaster*, il a été montré que le taux de conversion est 60 – 165 fois plus élevé que le taux de mutation aux mêmes loci.

Innan considère donc des séquences à deux loci de types suivants : AA , Aa , aA , aa , avec les allèles A et a à chacun des deux loci. Les fréquences des haplotypes numérotés de 1 à 4, respectivement, sont :

$$\begin{array}{cccc}
 A \circ & A \circ & a \circ & a \circ \\
 | & | & | & | \\
 A \circ & a \circ & A \circ & a \circ \\
 x_1 & x_2 & x_3 & x_4
 \end{array}$$

avec $x_1 + x_2 + x_3 + x_4 = 1$ dans une population de $2N$ haplotypes. Les fréquences des haplotypes correspondent aux probabilités de choisir les haplotypes respectifs dans un tirage au hasard. Le croisement de deux haplotypes tirés au hasard peut

être de 16 types ordonnés différents dont les probabilités sont données au tableau (4.1).

TABLEAU 4.1. Effets de la ségrégation et de la recombinaison.

Croisement femelle \times mâle	Probabilité	Distribution conditionnelle des haplotypes après ségrégation et recombinaison			
		AA	Aa	aA	aa
$AA \times AA$	$x_1 \times x_1$	1	0	0	0
$AA \times Aa$	$x_1 \times x_2$	$\frac{1}{2}$	$\frac{1}{2}$	0	0
$AA \times aA$	$x_1 \times x_3$	$\frac{1}{2}$	0	$\frac{1}{2}$	0
$AA \times aa$	$x_1 \times x_4$	$\frac{1}{2}(1 - \rho)$	$\frac{1}{2}\rho$	$\frac{1}{2}\rho$	$\frac{1}{2}(1 - \rho)$
$Aa \times AA$	$x_2 \times x_1$	$\frac{1}{2}$	$\frac{1}{2}$	0	0
$Aa \times Aa$	$x_2 \times x_2$	0	1	0	0
$Aa \times aA$	$x_2 \times x_3$	$\frac{1}{2}\rho$	$\frac{1}{2}(1 - \rho)$	$\frac{1}{2}(1 - \rho)$	$\frac{1}{2}\rho$
$Aa \times aa$	$x_2 \times x_4$	0	$\frac{1}{2}$	0	$\frac{1}{2}$
$aA \times AA$	$x_3 \times x_1$	$\frac{1}{2}$	0	$\frac{1}{2}$	0
$aA \times Aa$	$x_3 \times x_2$	$\frac{1}{2}\rho$	$\frac{1}{2}(1 - \rho)$	$\frac{1}{2}(1 - \rho)$	$\frac{1}{2}\rho$
$aA \times aA$	$x_3 \times x_3$	0	0	1	0
$aA \times aa$	$x_3 \times x_4$	0	0	$\frac{1}{2}$	$\frac{1}{2}$
$aa \times AA$	$x_4 \times x_1$	$\frac{1}{2}(1 - \rho)$	$\frac{1}{2}\rho$	$\frac{1}{2}\rho$	$\frac{1}{2}(1 - \rho)$
$aa \times Aa$	$x_4 \times x_2$	0	$\frac{1}{2}$	0	$\frac{1}{2}$
$aa \times aA$	$x_4 \times x_3$	0	0	$\frac{1}{2}$	$\frac{1}{2}$
$aa \times aa$	$x_4 \times x_4$	0	0	0	1

Soit p_i pour $i = 1, \dots, 4$ la probabilité qu'un nouvel individu produit soit de type i . On a

$$\begin{aligned}
 p_1 &= x_1^2 + x_1(x_2 + x_3) + \rho x_2 x_3 + (1 - \rho)x_1 x_4 \\
 &= x_1 - x_1 x_4 + \rho x_2 x_3 + (1 - \rho)x_1 x_4 = x_1 - \rho D,
 \end{aligned}$$

où $D = x_1x_4 - x_2x_3$ représente le déséquilibre de liaison. De même, on a

$$p_2 = x_2 + \rho D,$$

$$p_3 = x_3 + \rho D,$$

$$p_4 = x_4 - \rho D.$$

On considère par la suite que le descendant subit une mutation ou une conversion. La figure 4.1 illustre les événements de mutation et de conversion possibles. Les séquences résultantes des événements (1) et (2) sont seulement possibles suite à un événement de mutation à un locus, avec probabilité μ . Alors que les séquences résultantes des événements (3) et (4) sont possible suite à un événement de mutation ou un événement de conversion à un locus, avec probabilité $\mu + c$, sachant que les probabilités d'événements simultanés sont négligeables.



FIGURE 4.1. Séquences résultantes d'un événement de mutation ou de conversion. Les événements sont indiqués par des flèches associées aux probabilités.

TABLEAU 4.2. Effets de la mutation et de la conversion.

Haplotype	Probabilité	Distribution conditionnelle des haplotypes après mutation et conversion			
		AA	Aa	aA	aa
AA	p_1	$1 - 2\mu$	μ	μ	0
Aa	p_2	$\mu + c$	$1 - 2(\mu + c)$	0	$\mu + c$
aA	p_3	$\mu + c$	0	$1 - 2(\mu + c)$	$\mu + c$
aa	p_4	0	μ	μ	$1 - 2\mu$

Suite au facteur de mutation et conversion, soit p'_i pour $i = 1, \dots, 4$ la probabilité qu'un haplotype produit au hasard soit de type i . D'après le tableau (4.2), on a

$$\begin{aligned} p'_1 &= (1 - 2\mu)p_1 + (\mu + c)(p_2 + p_3), \\ p'_2 &= \mu(p_1 + p_4) + (1 - 2(\mu + c))p_2, \\ p'_3 &= \mu(p_1 + p_4) + (1 - 2(\mu + c))p_3, \\ p'_4 &= (1 - 2\mu)p_4 + (\mu + c)(p_2 + p_3). \end{aligned}$$

Ainsi, on trouve que

$$\begin{aligned} p'_1 &= (1 - 2\mu)x_1 + (\mu + c)(x_2 + x_3) - (1 - 4\mu - 2c)\rho D, \\ p'_2 &= \mu(x_1 + x_4) + (1 - 2(\mu + c))x_2 + (1 - 4\mu - 2c)\rho D, \\ p'_3 &= \mu(x_1 + x_4) + (1 - 2(\mu + c))x_3 + (1 - 4\mu - 2c)\rho D, \\ p'_4 &= (1 - 2\mu)x_4 + (\mu + c)(x_2 + x_3) - (1 - 4\mu - 2c)\rho D. \end{aligned}$$

Toutefois, étant donné qu'on considère que la probabilité de deux événements simultanés ou plus est négligeable lorsque N est grand, on trouve les approximations suivantes :

$$\begin{aligned} p'_1 &= (1 - 2\mu)x_1 + (\mu + c)(x_2 + x_3) - \rho D, \\ p'_2 &= \mu(x_1 + x_4) + (1 - 2(\mu + c))x_2 + \rho D, \\ p'_3 &= \mu(x_1 + x_4) + (1 - 2(\mu + c))x_3 + \rho D, \\ p'_4 &= (1 - 2\mu)x_4 + (\mu + c)(x_2 + x_3) - \rho D. \end{aligned}$$

À cette étape, un individu de la population est choisi au hasard et il est remplacé par l'haplotype produit au hasard. Soient, X'_i pour $i = 1, \dots, 4$, les fréquences des haplotypes après le remplacement (à l'instant $\frac{1}{2N^2}$ plus tard avec une unité de temps de $2N^2$ événements de naissance-mort). Alors, on a

$$X'_i = \begin{cases} x_i + \frac{1}{2N} & \text{avec probabilité } p'_i(1 - x_i), \\ x_i & \text{avec probabilité } 1 - (x_i + p'_i - 2x_i p'_i), \\ x_i - \frac{1}{2N} & \text{avec probabilité } (1 - p'_i)x_i, \end{cases}$$

pour $i = 1, \dots, 4$.

Tel qu'utilisé par Innan sous le modèle de reproduction de Wright-Fisher, on utilise le processus de diffusion sous ce modèle, pour calculer les espérances des moments des fréquences d'allèles et par la suite estimer les paramètres de mutation, de conversion et de recombinaison. Le générateur infinitésimal L du processus de diffusion correspondant $\{\mathbf{X}_t, t \geq 0\}$ de moyennes infinitésimales $a_i(\mathbf{x})$ et de variances infinitésimales $b_{i,j}(\mathbf{x})$ pour $i, j = 1, 2, 3$, et $\mathbf{x} = (x_1, x_2, x_3, x_4)$ (voir l'annexe A.3) est obtenu en faisant tendre N vers l'infini, avec une unité de temps de $2N^2$ événements de naissance-mort. Il est en fait donné en fonction de trois fréquences puisque la quatrième fréquence est donnée par $x_1 + x_2 + x_3 + x_4 = 1$. Le générateur est donc de forme

$$L(g(\mathbf{x})) = \sum_{i=1}^3 a_i(\mathbf{x}) \frac{dg}{dx_i}(\mathbf{x}) + \frac{1}{2} \sum_{i,j=1}^3 b_{i,j}(\mathbf{x}) \frac{d^2g}{dx_i dx_j}(\mathbf{x}). \quad (4.1.1)$$

En équilibre, la fonction $g(\mathbf{x})$ satisfait l'équation

$$\mathbb{E} \left[L(g) \right] = 0. \quad (4.1.2)$$

Pour obtenir les moyennes infinitésimales $a_i(\mathbf{x})$ (voir la définition dans l'annexe (A.3)) on définit $\Delta X_i = X'_i - X_i$ pour $i, j = 1, 2, 3$ et on calcule

$$\begin{aligned} \mathbb{E} \left[\Delta X_1 \middle| \mathbf{X} = \mathbf{x} \right] &= \frac{1}{2N^2} (-2\mu N x_1 + (\mu N + cN)(x_2 + x_3) - \rho N D), \\ \mathbb{E} \left[\Delta X_2 \middle| \mathbf{X} = \mathbf{x} \right] &= \frac{1}{2N^2} (-2(\mu N + cN)x_2 + \mu N(x_1 + x_4) + \rho N D), \\ \mathbb{E} \left[\Delta X_3 \middle| \mathbf{X} = \mathbf{x} \right] &= \frac{1}{2N^2} (-2(\mu N + cN)x_3 + \mu N(x_1 + x_4) + \rho N D), \\ \mathbb{E} \left[\Delta X_4 \middle| \mathbf{X} = \mathbf{x} \right] &= \frac{1}{2N^2} (-2\mu N x_4 + (\mu N + cN)(x_2 + x_3) - \rho N D). \end{aligned}$$

En remplaçant les probabilités $\rho = \frac{R}{2N}$, $\mu = \frac{\theta}{2N}$, et $c = \frac{C}{2N}$ on trouve que :

$$\begin{aligned} a_1(x) &= \frac{1}{2}(-2\theta x_1 + (\theta + C)(x_2 + x_3) - R.D), \\ a_2(x) &= \frac{1}{2}(-2(\theta + C)x_2 + \theta(x_1 + x_4) + R.D), \\ a_3(x) &= \frac{1}{2}(-2(\theta + C)x_3 + \theta(x_1 + x_4) + R.D). \end{aligned}$$

Ici, on a utilisé le fait qu'un événement de naissance-mort correspond à un intervalle de temps de longueur $h = \frac{1}{2N^2}$ avec l'unité de temps du processus de diffusion.

Pour obtenir les variances infinitésimales $b_{i,j}(\mathbf{x})$, pour $i, j = 1, 2, 3$, on calcule

$$\begin{aligned} \mathbb{E}[\Delta X_i \cdot \Delta X_j | \mathbf{X} = \mathbf{x}] &= \mathbb{E}[(X'_i - X_i)(X'_j - X_j) | \mathbf{X} = \mathbf{x}] \\ &= \mathbb{E}[X'_i \cdot X'_j - X_i \cdot X'_j - X'_i \cdot X_j + X_i \cdot X_j | \mathbf{X} = \mathbf{x}] \\ &= \mathbb{E}[X'_i \cdot X'_j | \mathbf{X} = \mathbf{x}] - x_i \mathbb{E}[X'_j | \mathbf{X} = \mathbf{x}] - x_j \mathbb{E}[X'_i | \mathbf{X} = \mathbf{x}] \\ &\quad + x_i \cdot x_j. \end{aligned}$$

On aura besoin de la distribution conjointe de (X'_i, X'_j) étant donné que $\mathbf{X} = \mathbf{x}$ pour $i, j = 1, 2, 3$, et $i \neq j$. Sous cette condition on a que

$$(X'_i, X'_j) = \begin{cases} (x_i + \frac{1}{2N}, x_j) & q_1 = p'_i(1 - x_i - x_j), \\ (x_i - \frac{1}{2N}, x_j) & q_2 = (1 - p'_i - p'_j)x_i, \\ (x_i, x_j + \frac{1}{2N}) & q_3 = p'_j(1 - x_i - x_j), \\ (x_i, x_j - \frac{1}{2N}) & q_4 = (1 - p'_i - p'_j)x_j, \\ (x_i + \frac{1}{2N}, x_j - \frac{1}{2N}) & q_5 = p'_i x_j, \\ (x_i - \frac{1}{2N}, x_j + \frac{1}{2N}) & q_6 = p'_j x_i, \\ (x_i, x_j) & q_7 = 1 - q_1 - q_2 - q_3 - q_4 - q_5 - q_6. \end{cases}$$

Ainsi, on trouve que

$$\begin{aligned} \mathbb{E}\left[\Delta X_1 \cdot \Delta X_2 \mid \mathbf{X} = \mathbf{x}\right] &= \\ &= \frac{\mu x_1^2 + x_2((\mu + c)(x_2 + x_3) - \rho D) + x_1(-2(2\mu + c - 1)x_2 + \mu x_4 + \rho D)}{2} \cdot \frac{1}{2N^2}, \\ \mathbb{E}\left[\Delta X_1 \cdot \Delta X_3 \mid \mathbf{X} = \mathbf{x}\right] &= \\ &= \frac{\mu x_1^2 + x_3((\mu + c)(x_2 + x_3) - \rho D) + x_1(-2(2\mu + c - 1)x_3 + \mu x_4 + \rho D)}{2} \cdot \frac{1}{2N^2}, \\ \mathbb{E}\left[\Delta X_2 \cdot \Delta X_3 \mid \mathbf{X} = \mathbf{x}\right] &= \\ &= \frac{x_3(\mu(x_1 + x_4) + \rho D) + x_2(\mu(x_1 + x_4) - 2(2\mu + 2c - 1)x_3 + \rho D)}{2} \cdot \frac{1}{2N^2}. \end{aligned}$$

En prenant $\rho = \frac{R}{2N}$, $\mu = \frac{\theta}{2N}$, et $c = \frac{C}{2N}$ on trouve que :

$$\begin{aligned} b_{1,2}(\mathbf{x}) &= b_{2,1}(\mathbf{x}) = -x_1x_2, \\ b_{1,3}(\mathbf{x}) &= b_{3,1}(\mathbf{x}) = -x_1x_3, \\ b_{2,3}(\mathbf{x}) &= b_{3,2}(\mathbf{x}) = -x_2x_3. \end{aligned}$$

De façon analogue, on trouve que :

$$\begin{aligned} \mathbb{E}\left[(\Delta X_1)^2 \mid \mathbf{X} = \mathbf{x}\right] &= \\ &= \frac{-2x_1^2(1 - 2\mu) - 2x_1((\mu + c)(x_2 + x_3) - \rho D - 1) + ((\mu + c)(x_2 + x_3) - \rho D)}{2} \cdot \frac{1}{2N^2}, \\ \mathbb{E}\left[(\Delta X_2)^2 \mid \mathbf{X} = \mathbf{x}\right] &= \\ &= \frac{-2x_2^2(1 - 2\mu - 2c) - 2x_2(\mu(x_1 + x_4) + \mu + c + \rho D - 1) + \rho D}{2} \cdot \frac{1}{2N^2}, \\ \mathbb{E}\left[(\Delta X_3)^2 \mid \mathbf{X} = \mathbf{x}\right] &= \\ &= \frac{-2x_3^2(1 - 2\mu - 2c) - 2x_3(\mu(x_1 + x_4) + \mu + c + \rho D - 1) + \rho D}{2} \cdot \frac{1}{2N^2}. \end{aligned}$$

Il s'ensuit que

$$\begin{aligned} b_{1,1}(\mathbf{x}) &= x_1(1 - x_1), \\ b_{2,2}(\mathbf{x}) &= x_2(1 - x_2), \\ b_{3,3}(\mathbf{x}) &= x_3(1 - x_3). \end{aligned}$$

Conséquemment, on obtient que

$$\begin{aligned} L(g(\mathbf{x})) &= \frac{x_1(1 - x_1)}{2} \frac{d^2g}{dx_1^2}(\mathbf{x}) + \frac{x_2(1 - x_2)}{2} \frac{d^2g}{dx_2^2}(\mathbf{x}) + \frac{x_3(1 - x_3)}{2} \frac{d^2g}{dx_3^2}(\mathbf{x}) \\ &\quad - x_1x_2 \frac{d^2g}{dx_1dx_2}(\mathbf{x}) - x_1x_3 \frac{d^2g}{dx_1dx_3}(\mathbf{x}) - x_2x_3 \frac{d^2g}{dx_2dx_3}(\mathbf{x}) \\ &\quad + \frac{1}{2}[-2\theta x_1 + (\theta + C)(x_2 + x_3) - R.D] \frac{dg}{dx_1}(\mathbf{x}) \\ &\quad + \frac{1}{2}[-2(\theta + C)x_2 + \theta(x_1 + x_4) + R.D] \frac{dg}{dx_2}(\mathbf{x}) \\ &\quad + \frac{1}{2}[-2(\theta + C)x_3 + \theta(x_1 + x_4) + R.D] \frac{dg}{dx_3}(\mathbf{x}). \end{aligned} \tag{4.1.3}$$

Les variables x_1 , x_2 , et x_3 de l'équation 4.1.3 sont remplacées par p , q , et D , où :

$p = x_1 + x_2$ est la fréquence des séquences avec A au locus (l) :



alors que $q = x_1 + x_3$ est la fréquence des séquences avec A au locus (m) :



et $D = x_1x_4 - x_2x_3 = x_1 - pq$. L'équation 4.1.3 devient :

$$L(g) = \frac{L'(g)}{2}, \tag{4.1.4}$$

où

$$\begin{aligned}
L'(g) &= p(1-p)\frac{d^2g}{dp^2} + q(1-q)\frac{d^2g}{dq^2} \\
&+ [pq(1-p)(1-q) + D(1-2p)(1-2q) - D^2]\frac{d^2g}{dD^2} \\
&+ 2D\frac{d^2g}{dpdq} + 2D(1-2p)\frac{d^2g}{dq dD} + 2D(1-2q)\frac{d^2g}{dq dD} \\
&+ [\theta(1-2p) - C(p-q)]\frac{dg}{dp} + [\theta(1-2q) + C(p-q)]\frac{dg}{dq} \\
&+ [Cp(1-p) + Cq(1-q) - (2 + 4\theta + 2C + R)D]\frac{dg}{dD}. \quad (4.1.5)
\end{aligned}$$

En assignant à g les valeurs p , q , p^2 , q^2 , pq , et D dans les équations (4.1.2) et (4.1.5), on obtient les équations de l'article original (Innan 2002a) :

$$\begin{aligned}
\mathbb{E}[L'(p)] &= \mathbb{E}[\theta(1-2p) - C(p-q)] = 0, \\
\mathbb{E}[L'(q)] &= \mathbb{E}[\theta(1-2q) + C(p-q)] = 0,
\end{aligned}$$

d'où

$$\mathbb{E}[p] = \mathbb{E}[q] = \frac{1}{2};$$

puis

$$\begin{aligned}
\mathbb{E}[L'(p^2)] &= \mathbb{E}[2p(1-p) + 2p[\theta(1-2p) - C(p-q)]] \\
&= 1 + \theta - 2(1 + 2\theta + C)E[p^2] + 2CE[pq] = 0, \quad (4.1.6)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[L'(q^2)] &= \mathbb{E}[2q(1-q) + 2q[\theta(1-2q) + C(p-q)]] \\
&= 1 + \theta - 2(1 + 2\theta + C)E[q^2] + 2CE[pq] = 0,
\end{aligned}$$

d'où

$$E[p^2] = E[q^2], \quad (4.1.7)$$

Et finalement

$$\begin{aligned}\mathbb{E}\left[L'(pq)\right] &= 2\mathbb{E}[D] + \theta + C\mathbb{E}[p^2] + C\mathbb{E}[q^2] - (4\theta + 2C)\mathbb{E}[pq] \quad (4.1.8) \\ &= 0,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[L'(D)\right] &= C - C\mathbb{E}[p^2] - C\mathbb{E}[q^2] - (2 + 4\theta + 2C + R)\mathbb{E}[D] \quad (4.1.9) \\ &= 0.\end{aligned}$$

Le système des quatre équations (4.1.6), (4.1.7), (4.1.8) et (4.1.9) donne les expressions en fonction de R , θ et C , suivantes :

$$\begin{aligned}\mathbb{E}[p^2] &= \mathbb{E}[q^2] = \frac{\lambda}{\omega}, \\ \mathbb{E}[pq] &= -\frac{1+\theta}{2C} + \frac{(1+\alpha)\lambda}{C\omega}, \\ \mathbb{E}[D] &= \frac{C}{\beta}\left(1 - \frac{2\lambda}{\omega}\right),\end{aligned}$$

où

$$\begin{aligned}\alpha &= 2\theta + C, \\ \beta &= 2 + 2\alpha + R, \\ \lambda &= 4C^2 + \beta[2\theta C + 2\alpha(1 + \theta)], \\ \omega &= 8C^2 + 4\beta[\alpha(1 + \alpha) - C^2].\end{aligned}$$

4.2. VARIABILITÉ GÉNÉTIQUE INTRA-LOCUS ET INTER-LOCUS

On rappelle que l'hétérozygotie intra-locus est la probabilité que deux séquences choisies aléatoirement sans remise dans la population totale de taille $2N$ soient de types différents au même locus. Ainsi, pour l'hétérozygotie au locus (l), $h_{w,(l)}$, on choisit deux séquences dans l'ordre au hasard et on définit

$$I_i = \begin{cases} 1 & \text{si la } i^{\text{ème}} \text{ séquence choisie est de type } \begin{matrix} A \circ \\ \dagger \end{matrix}, \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, on a

$$I_1 = \begin{cases} 1 & \text{si la première séquence choisie est de type } \begin{matrix} A \circ \\ | \\ \dagger \end{matrix}, \\ 0 & \text{sinon,} \end{cases}$$

$$I_2 = \begin{cases} 1 & \text{si la seconde séquence choisie est de type } \begin{matrix} A \circ \\ | \\ \dagger \end{matrix}, \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, on a

$$I_1(1-I_2) = \begin{cases} 1 & \text{si les deux séquences choisies dans cet ordre sont de types } \begin{matrix} A \circ & a \circ \\ | & | \\ \dagger & \dagger \end{matrix}, \\ 0 & \text{sinon.} \end{cases}$$

On a alors

$$\begin{aligned} \mathbb{E}(h_{w,(l)}) &= \mathbb{E}(2I_1(1-I_2)) = \mathbb{E}\left(\mathbb{E}(2I_1(1-I_2)) \middle| p\right) \\ &= \mathbb{E}(2p(1-p)). \end{aligned}$$

De façon analogue pour l'espérance de l'hétérozygotie au locus (m) , on a

$$\mathbb{E}(h_{w,(m)}) = \mathbb{E}(2q(1-q)) = \mathbb{E}(h_{w,(l)}).$$

Ainsi l'espérance de la mesure de variabilité intra-locus est donnée par la moyenne de $h_{w,(l)}$ et $h_{w,(m)}$, soit

$$\mathbb{E}(h_w) = \mathbb{E}(h_{w,(l)}) = \mathbb{E}(h_{w,(m)}).$$

Maintenant, on définit la mesure de variabilité génétique inter-locus, h_b comme étant la probabilité que deux allèles choisis aléatoirement à deux loci différents soient de types différents. Son espérance est donnée par

$$\mathbb{E}(h_b) = \mathbb{E}\left(\begin{matrix} A \circ & \dagger \\ | & | \\ \dagger & a \circ \end{matrix} \text{ ou } \begin{matrix} a \circ & \dagger \\ | & | \\ \dagger & A \circ \end{matrix}\right) = \mathbb{E}(p(1-q) + (1-p)q) = 1 - 2\mathbb{E}(pq).$$

Puisque les espérances des mesures de variabilité inter-locus et intra-locus et le déséquilibre de liaison sont des fonctions de θ , R et C , alors il est possible d'estimer ces paramètres à partir des polymorphismes de données d'ADN.

4.3. ESTIMATION DE PARAMÈTRES

Pour estimer les mesures de variabilité et le déséquilibre de liaison, on considère un seul site sur une paire de gènes de la région *Amy*, le gène proximal et le gène distal, avec deux nucléotides possibles T et C , de façon à ce qu'il y ait quatre haplotypes possibles, TT , TC , CT et CC . Les fréquences des haplotypes sont respectivement, n_1 , n_2 , n_3 , et n_4 , avec $n = n_1 + n_2 + n_3 + n_4$.

Soient les estimés de l'hétérozygotie intra-gène donnés par

$$h_{w,p} = \frac{(n_1 + n_2)(n_3 + n_4)}{\binom{n}{2}},$$

$$h_{w,d} = \frac{(n_1 + n_3)(n_2 + n_4)}{\binom{n}{2}},$$

et par conséquent

$$h_w = \frac{h_{w,p} + h_{w,d}}{2}.$$

La mesure de variabilité inter-gène est estimée par

$$h_b = \frac{(n_1 + n_2)(n_2 + n_4) + (n_1 + n_3)(n_3 + n_4) - n_2 - n_3}{n(n-1)},$$

et l'estimé du déséquilibre de liaison donné par Nei et Roychoudhury (1974) est

$$\delta = \frac{n_1 n_4 - n_2 n_3}{n(n-1)}.$$

Ainsi, les mesures h_w , h_b et δ peuvent être calculées à chaque site des deux gènes (parmi 37 sites synonymes, voir Inomata *et al.* 1995) pour obtenir les moyennes. Les moyennes de $h_{w,p}$ et $h_{w,d}$ correspondent aux mesures de polymorphisme de chaque gène (voir l'équation 1.3.1), π_{wp} et π_{wd} , les nombres moyens de différences

entre paires pour le gène proximal et le gène distal, dont la moyenne π_w est le nombre moyen de différences entre paires intra-gène. De façon analogue, π_b , le nombre moyen de différences entre paires inter-gène, est la moyenne de h_b parmi les sites considérés. Enfin, la moyenne de δ correspond au déséquilibre de liaison moyen entre deux gènes, représenté par d . Innan (2002a) obtient ainsi des estimés de $\mathbb{E}(h_w)$, $\mathbb{E}(h_b)$ et $\mathbb{E}(D)$, et par ricochet des estimés de θ , C et R . Cette estimation assume que les trois paramètres sont constants le long de la région d'ADN. Les estimés obtenus sont donc des moyennes pour tous les sites considérés.

4.4. CALCULS POUR LE MODÈLE À UNE INFINITÉ DE SITES

Dans la publication suivante d'Innan (2002b), les résultats théoriques précédents du modèle à deux loci avec mutation récurrente sont étendus à un modèle de mutation à une infinité de sites. La particularité du modèle à une infinité de sites est que le modèle assume que le taux de mutation est si petit à chaque site qu'il n'y a pas de mutations multiples, en d'autres mots la probabilité d'une seule mutation à un site est égale à la probabilité d'au moins une mutation au même site.

Comme dans la section précédente les mesures h_w , h_b , et D sont considérées dans une famille multigénique *Amy* avec deux gènes dupliqués (loci) *I* et *II*, chacun constitué de L sites, avec $L \rightarrow \infty$. On suppose que n chromosomes sont sélectionnés aléatoirement d'une population et que les deux gènes sont sélectionnés sur chaque chromosome, voir la figure 4.2.

Les mesures de variabilité intra-gène pour le gène *I* et pour le gène *II* sont respectivement

$$\pi_{w1} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=1+1}^n k_{11}(ij), \quad (4.4.1)$$

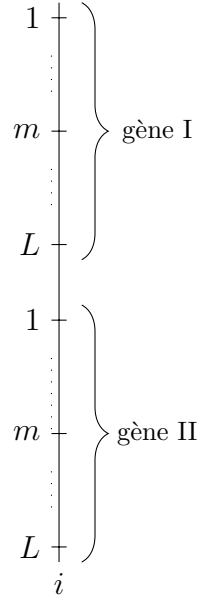


FIGURE 4.2. Chromosome i , avec $i = 1, \dots, n$ à deux gènes avec L sites chacun tel que chaque site $m = 1, \dots, L$ peut avoir deux types d'allèles, A et a .

$$\pi_{w2} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=1+1}^n k_{22}(ij), \quad (4.4.2)$$

où $k_{11}(ij)$ et $k_{22}(ij)$ sont respectivement le nombre de sites différents entre les chromosomes i et j au niveau du gène I et au niveau du gène II , respectivement.

Soit $k_{12}(ij)$ le nombre de sites différents entre le gène I du chromosome i et le gène II du chromosome j . La mesure de variabilité inter-gène est

$$\pi_b = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k_{12}(ij). \quad (4.4.3)$$

On définit, D_{sum} le déséquilibre de liaison parmi les L sites. On a $D_{sum} = \sum_{m=1}^L D_m$, où D_m est le déséquilibre de liaison au site m , pour $m = 1, \dots, L$, donné par

$$D_m = \frac{n_{AA}n_{aa} - n_{Aa}n_{aA}}{n(n-1)}. \quad (4.4.4)$$

Ici n_{xy} représente le nombre de chromosomes avec $x = A$ ou a et $y = A$ ou a , respectivement, au site m des gènes I et II .

Ce modèle à deux gènes et à une infinité de sites contient L modèles à deux loci pour les deux gènes dupliqués. Par conséquent, les espérances des quatre mesures de variations (4.4.1), (4.4.2), (4.4.3) et (4.4.4) sont données par

$$\begin{aligned}\frac{\mathbb{E}[\pi_{w1}]}{L} &= \frac{\mathbb{E}[\pi_{w2}]}{L} = \mathbb{E}[h_w], \\ \frac{\mathbb{E}[\pi_b]}{L} &= \mathbb{E}[h_b], \\ \frac{\mathbb{E}[D_{sum}]}{L} &= \mathbb{E}[D].\end{aligned}$$

L'objectif de cette section est de trouver les expressions de $\mathbb{E}[h_w]$, $\mathbb{E}[h_b]$ et $\mathbb{E}[D]$ lorsque $L \rightarrow \infty$ et $L\theta = \Theta$ en utilisant le processus ancestral limite, plutôt qu'une approximation par une diffusion sous le modèle de Wright-Fisher comme dans Innan (2002b).

Pour ce faire, on considère le matériel ancestral d'un échantillon de deux séquences avec A ou a à chaque locus de chaque séquence qui définit une lignée échantillonnable du même type. Une séquence ancestrale à un locus est dite de type A , a ou a, A selon qu'elle est ancestrale à ce locus à des lignées échantillonnables de type A seulement, à des lignées échantillonnables de type a seulement, ou à des lignées échantillonnables de type a et des lignées échantillonnables de type A .

On introduit :

- $n_A^{(l)}$, le nombre de séquences ancestrales de type A au locus (l) et non ancestrales au locus (m) ;
- $n_A^{(m)}$, le nombre de séquences ancestrales de type A au locus (m) et non ancestrales au locus (l) ;
- $n_{AA}^{(lm)}$, le nombre de séquences ancestrales de type A aux deux loci (l) et (m) ;
- $n_a^{(l)}$, le nombre de séquences ancestrales de type a au locus (l) et non ancestrales au locus (m) ;

- $n_a^{(m)}$, le nombre de séquences ancestrales de type a au locus (m) et non ancestrales au locus (l) ;
- $n_{aa}^{(lm)}$, le nombre de séquences ancestrales de type a aux deux loci (l) et (m) ;
- $n_{Aa}^{(lm)}$, le nombre de séquences ancestrales de type A au locus (l) et de type a au locus (m) ;
- $n_{aA}^{(lm)}$, le nombre de séquences ancestrales de type a au locus (l) et de type A au locus (m) ;
- $n_{a,A}$, le nombre de séquences ancestrales de type a, A à au moins un locus.

L'état du matériel ancestral à un moment donné dans le passé jusqu'à la coalescence d'une lignée de type A avec une lignée de type a est représenté par le vecteur $\underline{n} = (n_A^{(l)}, n_A^{(m)}, n_{AA}^{(lm)}, n_a^{(l)}, n_a^{(m)}, n_{aa}^{(lm)}, n_{Aa}^{(lm)}, n_{aA}^{(lm)})$. On définit les variables :

$$n_A = n_A^{(l)} + n_A^{(m)} + n_{AA}^{(lm)} + n_{Aa}^{(lm)} + n_{aA}^{(lm)},$$

$$n^{(l)} = n_A^{(l)} + n_a^{(l)},$$

$$n^{(m)} = n_A^{(m)} + n_a^{(m)},$$

$$n^{(lm)} = n_{AA}^{(lm)} + n_{aa}^{(lm)} + n_{Aa}^{(lm)} + n_{aA}^{(lm)},$$

$$n = n^{(l)} + n^{(m)} + n^{(lm)}.$$

En remontant le temps à partir de l'état $\underline{n} \neq (0, 0, 0, 0, 0, 0, 0, 0)$, chaque paire de séquences ancestrales coalesce au taux 1, alors que chaque séquence ancestrale subit une recombinaison aux taux $R/2$ et une conversion au taux C . On considère seulement les événements de recombinaison qui modifient l'état du matériel ancestral, donc ceux qui affectent une séquence ancestrale aux deux loci, alors qu'on considère tous les événements de conversion qu'ils changent ou non l'état du matériel ancestral. Le taux total d'arrivée d'un événement est donc

$$\lambda_{\underline{n}} = \frac{n(n-1)}{2} + n^{(lm)} \frac{R}{2} + nC,$$

et le temps $T_{\underline{n}}$ avant l'arrivée d'un tel événement satisfait

$$\begin{aligned}\mathbb{E}[T_{\underline{n}}] &= \frac{1}{\lambda_{\underline{n}}}, \\ \mathbb{E}[T_{\underline{n}}^2] &= \frac{2}{\lambda_{\underline{n}}^2}.\end{aligned}$$

En introduisant le facteur de conversion au processus ancestral avec recombinaison, le MRCA n'est plus associé à chaque locus pris séparément, mais au deux loci. On définit $T_{\underline{n}}^A$, la longueur de la branche ancestrale à *toutes* les lignées échantillonnées de type A et non ancestrale aux lignées échantillonnées de type a à partir de l'état \underline{n} . On a alors

$$T_{\underline{n}}^A = \begin{cases} T_{\underline{n}} + T_{\underline{\gamma}}, & \text{si } n_A = 1, \\ T_{\underline{\gamma}} & \text{si } n_A > 1, \\ 0 & \text{si } n = 0. \end{cases} \quad (4.4.5)$$

Ici $\underline{\gamma} = \underline{\gamma}(\underline{n})$ désigne l'état du matériel ancestral juste après le temps $T_{\underline{n}}$. Cette variable aléatoire prend la valeur \underline{n}' avec probabilité $p_{\underline{n},\underline{n}'} = \frac{\lambda_{\underline{n},\underline{n}'}}{\lambda_{\underline{n}}}$, où $\lambda_{\underline{n},\underline{n}'}$ est le taux de transition de l'état \underline{n} à l'état \underline{n}' .

On note que la longueur $T_{\underline{n}}$ est nulle si dans l'état de départ il y a une seule lignée ancestrale aux lignées échantillonnées de type A et non ancestrale aux lignées de type a , où $n_A = 1$, sinon $T_{\underline{n}} = 1$. De plus $T_{\underline{n}}^A = 0$ si $\underline{n} = (0, 0, 0, 0, 0, 0, 0, 0)$, car il y a alors une branche qui est ancestrale aux deux types au même locus.

On considère que A est le type mutant. La mesure d'hétérozygotie (4.2.1) pour deux lignées échantillonnées au locus (l), est équivalente à l'espérance de la probabilité que la mutation se soit produite sur la branche ancestrale à la lignée échantillonnée de type A et non ancestrale à l'autre lignée de type a (voir la figure 4.3).

De plus, le modèle à une infinité de sites assume que le taux de mutation $\frac{\theta}{2}$ le long d'une lignée est petit de telle sorte que la probabilité d'une mutation

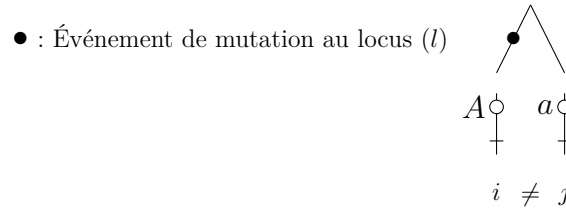


FIGURE 4.3. Événement établissant le lien entre l'hétérozygotie et la mutation.

correspond à la probabilité d'au moins une mutation. Ainsi,

$$\begin{aligned}
 \mathbb{E}\left(2p(1-p)\right) &= 2 \times \mathbb{E}\left(\text{probabilité d'au moins une mutation sur } T_{(1,0,0,1,0,0,0,0)}^A\right) \\
 &= 2 \times \mathbb{E}\left(1 - e^{-\frac{\theta}{2}T_3^A}\right) \\
 &= 2 \times \mathbb{E}\left(1 - 1 + \frac{\theta}{2}T_3^A - \frac{\theta^2}{2^2 \times 2!}T_3^{A^2} + \frac{\theta^3}{2^3 \times 3!}T_3^{A^3} - \dots\right) \\
 &\approx \theta \mathbb{E}\left(T_3^A\right).
 \end{aligned}$$

Les états ancestraux et les événements possibles à considérer à partir de ces états ainsi que les probabilités de transition associées pour deux lignées, une de type A et une autre de type a sont donnés dans le tableau (4.3) et la figure 4.4.

Les transitions à partir des trois états 1, 2, 3 et l'expression (4.4.5), nous permettent d'obtenir le système suivant :

$$\begin{aligned}
 \mathbb{E}(T_1^A) &= \frac{2}{R+2C} \left(1 + \frac{R}{2} \mathbb{E}(T_2^A)\right), \\
 \mathbb{E}(T_2^A) &= \frac{1}{1+2C} \left(1 + \mathbb{E}(T_1^A) + C \left(\mathbb{E}(T_2^A) + \mathbb{E}(T_3^A)\right)\right), \\
 \mathbb{E}(T_3^A) &= \frac{1}{1+2C} \left(1 + C \left(\mathbb{E}(T_2^A) + \mathbb{E}(T_3^A)\right)\right).
 \end{aligned}$$

On résout le système précédent et on trouve les espérances suivantes :

$$\begin{aligned}
 \mathbb{E}(T_1^A) &= \frac{(1+2C)(2+R)}{C(2+4C+R)}, \\
 \mathbb{E}(T_2^A) &= \frac{4C^2+4C+2CR+R+2}{C(2+4C+R)}, \\
 \mathbb{E}(T_3^A) &= \frac{2(2+2C+R)}{2+4C+R}.
 \end{aligned}$$

TABLEAU 4.3. Etats ancestraux et transitions d'état pour deux lignées, une de type A et une autre de type a .

Etat	\underline{n}	Evénements possibles	Prochain état
1	(0,0,0,0,0,1,0)	Rec	2
		$A\dot{\phi}$ $a\dot{\phi}$ ou $a\dot{\phi}$ $A\dot{\phi}$	1'
		Conv	1*
1'	(0,0,0,0,0,0,0)	absorbant	-
1*	(0,0,0,0,0,0,0)	$A, a\dot{\phi}$	-
		$a, A\dot{\phi}$	absorbant
2	(1,0,0,0,1,0,0)	Coa	1
		ou (0,1,0,1,0,0,0)	3
		Conv	2
3	(1,0,0,1,0,0,0)	Coa	1'
		ou (0,1,0,0,1,0,0)	1*
		Conv	3
		Conv	2

On en conclut que

$$\begin{aligned}
 \mathbb{E}(\pi_w) &= L\mathbb{E}(h_w) = \mathbb{E}(2p(1-p)) = L\theta\mathbb{E}(T_3^A) \\
 &= \frac{2\Theta(2+2C+R)}{2+4C+R},
 \end{aligned} \tag{4.4.6}$$

$$\begin{aligned}
 \mathbb{E}(\pi_b) &= L\mathbb{E}(h_b) = L\mathbb{E}(2p(1-q)) = L\theta\mathbb{E}(T_2^A) \\
 &= \frac{\Theta(4C^2+4C+2CR+R+2)}{C(2+4C+R)}.
 \end{aligned} \tag{4.4.7}$$

Avec un raisonnement analogue (voir le tableau 4.4 pour la notation) on trouve l'espérance du déséquilibre de liaison sous la forme

$$\begin{aligned}
 \mathbb{E}(D) &= \mathbb{E}(x_1x_4 - x_2x_3) \\
 &= \frac{\theta}{2}\mathbb{E}(T_{16}^A - T_{15}^A).
 \end{aligned}$$

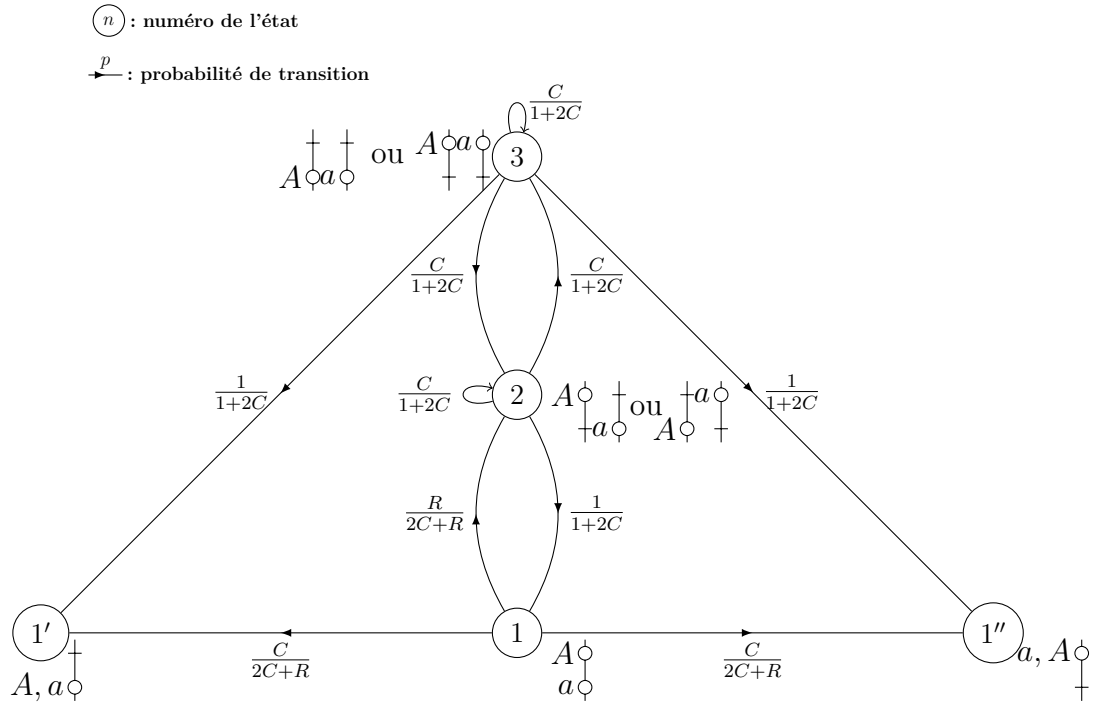


FIGURE 4.4. Diagramme de transitions d'état pour le processus de coalescence avec recombinaison et conversion à partir de deux lignées, une de type A et une autre de type a . Les transitions possibles entre les 5 états sont indiquées par des flèches associées aux probabilités de transition.

En effet, on a

$$\begin{aligned}
 \mathbb{E}(x_1 x_4) &= \mathbb{E}(\text{probabilité d'au moins une mutation sur } T_{(0,0,1,0,0,1,0,0)}^A) \\
 &= \mathbb{E}\left(1 - e^{-\frac{\theta}{2} T_{16}^A}\right) \\
 &\approx \frac{\theta}{2} \mathbb{E}\left(T_{16}^A\right),
 \end{aligned}$$

et

$$\begin{aligned}
\mathbb{E}(x_2x_3) &= \mathbb{E}\left(\text{probabilité d'au moins une mutation sur } T_{(0,0,0,0,0,0,1,1)}^A\right) \\
&= \mathbb{E}\left(1 - e^{-\frac{\theta}{2}T_{15}^A}\right) \\
&\approx \frac{\theta}{2}\mathbb{E}(T_{15}^A).
\end{aligned}$$

Les 16 états du système et les événements possibles avec leurs probabilités pour quatre lignées dont deux de type A et deux de type a sont donnés dans le tableau (4.4) et la figure 4.5.

Les 12 états du sous-système pour trois lignées dont deux exactement du même type et les transitions possibles à partir de ces états sont décrits dans le tableau (4.5) et la figure 4.6.

Les transitions entre les 12 états du sous-système, nous permettent de trouver le système d'équations suivant :

$$\begin{aligned}
\mathbb{E}(T_{\underline{8}}^A) &= \frac{2}{6 + 6C + R} \left(\mathbb{E}(T_{\underline{16}}^A) + \frac{R}{2} \mathbb{E}(T_{\underline{10}'}^A) + C \left(\mathbb{E}(T_{\underline{6}}^A) + \mathbb{E}(T_{\underline{8}}^A) + \mathbb{E}(T_{\underline{8}'}^A) \right) \right), \\
\mathbb{E}(T_{\underline{8}'}^A) &= \frac{2}{6 + 6C + R} \left(\mathbb{E}(T_{\underline{5}}^A) + \frac{R}{2} \mathbb{E}(T_{\underline{11}}^A) + \frac{C}{2} \left(\mathbb{E}(T_{\underline{6}'}^A) + \mathbb{E}(T_{\underline{6}^*}^A) + 2\mathbb{E}(T_{\underline{8}}^A) \right. \right. \\
&\quad \left. \left. + 2\mathbb{E}(T_{\underline{8}'}^A) \right) \right), \\
\mathbb{E}(T_{\underline{9}}^A) &= \frac{2}{6 + 6C + R} \left(\mathbb{E}(T_{\underline{16}}^A) + \frac{R}{2} \mathbb{E}(T_{\underline{10}'}^A) + C \left(\mathbb{E}(T_{\underline{7}}^A) + \mathbb{E}(T_{\underline{9}}^A) + \mathbb{E}(T_{\underline{9}'}^A) \right) \right), \\
\mathbb{E}(T_{\underline{9}'}^A) &= \frac{2}{6 + 6C + R} \left(\mathbb{E}(T_{\underline{4}}^A) + \frac{R}{2} \mathbb{E}(T_{\underline{11}'}^A) + \frac{C}{2} \left(\mathbb{E}(T_{\underline{7}'}^A) + \mathbb{E}(T_{\underline{7}^*}^A) + 2\mathbb{E}(T_{\underline{9}}^A) \right. \right. \\
&\quad \left. \left. + 2\mathbb{E}(T_{\underline{9}'}^A) \right) \right), \\
\mathbb{E}(T_{\underline{10}}^A) &= \frac{1}{6 + 4C} \left(\mathbb{E}(T_{\underline{6}'}^A) + \mathbb{E}(T_{\underline{7}'}^A) + C \left(2\mathbb{E}(T_{\underline{10}}^A) + \mathbb{E}(T_{\underline{11}}^A) + \mathbb{E}(T_{\underline{11}'}^A) \right) \right), \\
\mathbb{E}(T_{\underline{10}'}^A) &= \frac{1}{6 + 4C} \left(\mathbb{E}(T_{\underline{8}}^A) + \mathbb{E}(T_{\underline{9}}^A) + 2\mathbb{E}(T_{\underline{13}'}^A) + C \left(2\mathbb{E}(T_{\underline{10}'}^A) + \mathbb{E}(T_{\underline{11}}^A) + \mathbb{E}(T_{\underline{11}'}^A) \right) \right), \\
\mathbb{E}(T_{\underline{10}^*}^A) &= \frac{1}{6 + 4C} \left(\mathbb{E}(T_{\underline{6}^*}^A) + \mathbb{E}(T_{\underline{7}^*}^A) + 4\mathbb{E}(T_{\underline{13}}^A) + C \left(2\mathbb{E}(T_{\underline{10}^*}^A) + \mathbb{E}(T_{\underline{11}}^A) + \mathbb{E}(T_{\underline{11}'}^A) \right) \right), \\
\mathbb{E}(T_{\underline{11}}^A) &= \frac{1}{6 + 4C} \left(\mathbb{E}(T_{\underline{7}}^A) + \mathbb{E}(T_{\underline{8}'}^A) + 2\mathbb{E}(T_{\underline{12}}^A) + \frac{C}{2} \left(\mathbb{E}(T_{\underline{10}}^A) + 2\mathbb{E}(T_{\underline{10}'}^A) \right) \right)
\end{aligned}$$

$$\begin{aligned}
& + \mathbb{E}(T_{10^*}^A) + 4\mathbb{E}(T_{11}^A) \Big) \Big), \\
\mathbb{E}(T_{11'}^A) &= \frac{1}{6+4C} \left(\mathbb{E}(T_{\underline{6}}^A) + \mathbb{E}(T_{9'}^A) + 2\mathbb{E}(T_{12'}^A) + \frac{C}{2} \left(\mathbb{E}(T_{10}^A) + 2\mathbb{E}(T_{10'}^A) \right. \right. \\
& \left. \left. + \mathbb{E}(T_{10^*}^A) + 4\mathbb{E}(T_{11'}^A) \right) \right), \\
\mathbb{E}(T_{12}^A) &= \frac{2}{6+6C+R} \left(\mathbb{E}(T_{5'}^A) + \frac{R}{2}\mathbb{E}(T_{11}^A) + \frac{C}{2} \left(2\mathbb{E}(T_{12}^A) + \mathbb{E}(T_{13}^A) + \mathbb{E}(T_{13'}^A) \right) \right), \\
\mathbb{E}(T_{12'}^A) &= \frac{2}{6+6C+R} \left(\mathbb{E}(T_{4'}^A) + \frac{R}{2}\mathbb{E}(T_{11'}^A) + \frac{C}{2} \left(2\mathbb{E}(T_{12'}^A) + \mathbb{E}(T_{13}^A) + \mathbb{E}(T_{13'}^A) \right) \right), \\
\mathbb{E}(T_{13}^A) &= \frac{2}{6+6C+R} \left(\mathbb{E}(T_{4^*}^A) + \mathbb{E}(T_{5^*}^A) + \mathbb{E}(T_{14}^A) + \frac{R}{2}\mathbb{E}(T_{10^*}^A) + \frac{C}{2} \left(\mathbb{E}(T_{12}^A) \right. \right. \\
& \left. \left. + \mathbb{E}(T_{12'}^A) + 2\mathbb{E}(T_{13}^A) \right) \right), \\
\mathbb{E}(T_{13'}^A) &= \frac{2}{6+6C+R} \left(\mathbb{E}(T_{15}^A) + \frac{R}{2}\mathbb{E}(T_{10'}^A) + \frac{C}{2} \left(\mathbb{E}(T_{12}^A) + \mathbb{E}(T_{12'}^A) + 2\mathbb{E}(T_{13'}^A) \right) \right), \\
\mathbb{E}(T_{14}^A) &= \frac{1}{1+2C+R} \left(\mathbb{E}(T_{\underline{1}}^A) + R\mathbb{E}(T_{13}^A) \right), \\
\mathbb{E}(T_{15}^A) &= \frac{1}{1+2C+R} \left(R\mathbb{E}(T_{13'}^A) \right), \\
\mathbb{E}(T_{16}^A) &= \frac{1}{1+2C+R} \left(\frac{R}{2} \left(\mathbb{E}(T_{\underline{8}}^A) + \mathbb{E}(T_{\underline{9}}^A) \right) + C \left(\mathbb{E}(T_{\underline{4}}^A) + \mathbb{E}(T_{\underline{5}}^A) \right) \right).
\end{aligned}$$

Les transitions entre les 16 états pour quatres lignées nous permettent ensuite de trouver le système suivant :

$$\begin{aligned}
\mathbb{E}(T_{\underline{4}}^A) &= \frac{2}{2+4C+R} \left(\frac{R}{2}\mathbb{E}(T_{\underline{6}}^A) + \frac{C}{2} \left(2\mathbb{E}(T_{\underline{4}}^A) + \mathbb{E}(T_{\underline{2}}^A) + \mathbb{E}(T_{\underline{3}}^A) \right) \right), \\
\mathbb{E}(T_{\underline{4}'}^A) &= \frac{2}{2+4C+R} \left(\frac{R}{2}\mathbb{E}(T_{\underline{6}}^A) + \frac{C}{2} \left(2\mathbb{E}(T_{\underline{4}'}^A) + \mathbb{E}(T_{\underline{4}^*}^A) \right) \right), \\
\mathbb{E}(T_{\underline{4}^*}^A) &= \frac{2}{2+4C+R} \left(\mathbb{E}(T_{\underline{1}}^A) + \frac{R}{2}\mathbb{E}(T_{\underline{6}^*}^A) + \frac{C}{2} \left(\mathbb{E}(T_{\underline{4}'}^A) + \mathbb{E}(T_{\underline{4}^*}^A) \right) \right), \\
\mathbb{E}(T_{\underline{5}}^A) &= \frac{2}{2+4C+R} \left(1 + \frac{R}{2}\mathbb{E}(T_{\underline{7}}^A) + \frac{C}{2} \left(2\mathbb{E}(T_{\underline{5}}^A) + \mathbb{E}(T_{\underline{2}}^A) + \mathbb{E}(T_{\underline{3}}^A) \right) \right), \\
\mathbb{E}(T_{\underline{5}'}^A) &= \frac{2}{2+4C+R} \left(1 + \frac{R}{2}\mathbb{E}(T_{\underline{7}}^A) + \frac{C}{2} \left(2\mathbb{E}(T_{\underline{5}'}^A) + \mathbb{E}(T_{\underline{5}^*}^A) \right) \right),
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(T_{\underline{5}^*}^A) &= \frac{2}{2+4C+R} \left(1 + \mathbb{E}(T_{\underline{1}}^A) + \frac{R}{2} \mathbb{E}(T_{\underline{7}^*}^A) + \frac{C}{2} (\mathbb{E}(T_{\underline{5}'}^A) + \mathbb{E}(T_{\underline{5}^*}^A)) \right), \\
\mathbb{E}(T_{\underline{6}}^A) &= \frac{2}{6+6C} \left(\mathbb{E}(T_{\underline{4}}^A) + \mathbb{E}(T_{\underline{4}'}^A) + \frac{C}{2} (4\mathbb{E}(T_{\underline{6}}^A) + \mathbb{E}(T_{\underline{6}'}^A) + \mathbb{E}(T_{\underline{6}^*}^A)) \right), \\
\mathbb{E}(T_{\underline{6}'}^A) &= \frac{2}{6+6C} \left(\mathbb{E}(T_{\underline{3}}^A) + \frac{C}{2} (2\mathbb{E}(T_{\underline{6}}^A) + 3\mathbb{E}(T_{\underline{6}'}^A) + \mathbb{E}(T_{\underline{6}^*}^A)) \right), \\
\mathbb{E}(T_{\underline{6}^*}^A) &= \frac{2}{6+6C} \left(\mathbb{E}(T_{\underline{2}}^A) + 2\mathbb{E}(T_{\underline{4}^*}^A) + \frac{C}{2} (2\mathbb{E}(T_{\underline{6}}^A) + \mathbb{E}(T_{\underline{6}'}^A) + 3\mathbb{E}(T_{\underline{6}^*}^A)) \right), \\
\mathbb{E}(T_{\underline{7}}^A) &= \frac{2}{6+6C} \left(1 + \mathbb{E}(T_{\underline{5}}^A) + \mathbb{E}(T_{\underline{5}'}^A) + \frac{C}{2} (4\mathbb{E}(T_{\underline{7}}^A) + \mathbb{E}(T_{\underline{7}'}^A) + \mathbb{E}(T_{\underline{7}^*}^A)) \right), \\
\mathbb{E}(T_{\underline{7}'}^A) &= \frac{2}{6+6C} \left(1 + \mathbb{E}(T_{\underline{3}}^A) + \frac{C}{2} (2\mathbb{E}(T_{\underline{7}}^A) + 3\mathbb{E}(T_{\underline{7}'}^A) + \mathbb{E}(T_{\underline{7}^*}^A)) \right), \\
\mathbb{E}(T_{\underline{7}^*}^A) &= \frac{2}{6+6C} \left(1 + \mathbb{E}(T_{\underline{2}}^A) + 2\mathbb{E}(T_{\underline{5}^*}^A) + \frac{C}{2} (2\mathbb{E}(T_{\underline{7}}^A) + \mathbb{E}(T_{\underline{7}'}^A) + 3\mathbb{E}(T_{\underline{7}^*}^A)) \right).
\end{aligned}$$

En résolvant ces systèmes, on trouve que

$$\begin{aligned}
\mathbb{E}(D_{sum}) &= L\mathbb{E}(D) = \frac{\Theta}{2} \mathbb{E}(T_{\underline{16}}^A - T_{\underline{15}}^A) \\
&= \frac{2\Theta C}{2+4C+R}.
\end{aligned} \tag{4.4.8}$$

Les paramètres Θ , C et R peuvent être estimés par π_w , π_b et D_{sum} à partir des expressions (4.4.6), (4.4.7) et (4.4.8) :

$$\begin{aligned}
\hat{\Theta} &= \frac{\pi_w + 2D_{sum}}{2}, \\
\hat{C} &= \frac{\pi_w - 2D_{sum}}{2(\pi_b - \pi_w)}, \\
\hat{R} &= \frac{\pi_w^2 + 4D_{sum}^2 - 4\pi_b D_{sum}}{2(\pi_b - \pi_w)D_{sum}}.
\end{aligned}$$

Les estimations obtenues par le processus ancestral concordent avec celles obtenues par Innan (2002b). Nos résultats s'appuient sur la robustesse du processus ancestral limite qui est applicable à un large éventail de modèles.

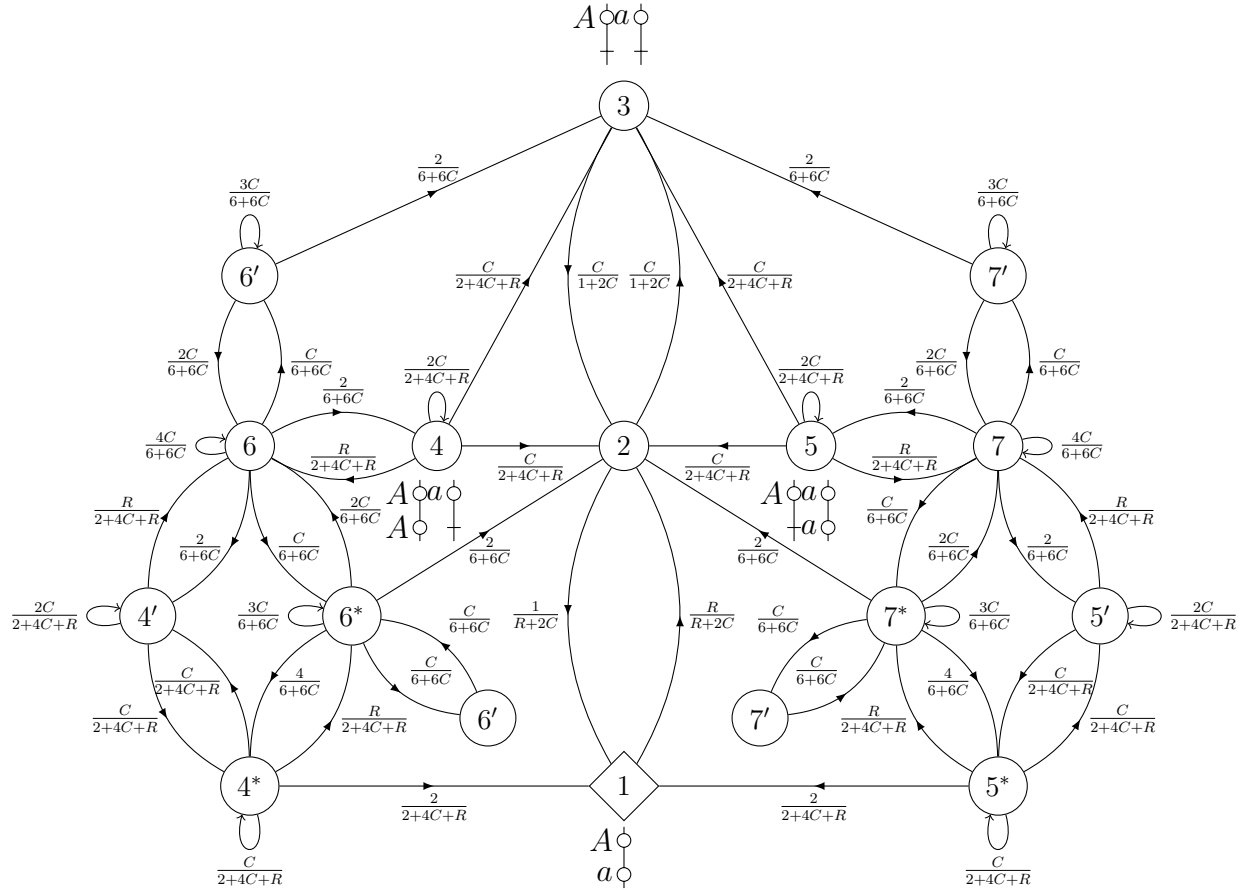


FIGURE 4.6. Diagramme de transitions d'état pour le processus de coalescence avec recombinaison et conversion à partir de trois lignées, deux du même type et un de l'autre type. Les transitions possibles du système sont indiquées par des flèches associées aux probabilités de transition.

TABLEAU 4.4. Etats ancestraux et transitions d'état pour quatre lignées, deux de type A et deux de type a .

Etat	\underline{n}	Evénements possibles	Prochain état	Etat	\underline{n}	Evénements possibles	Prochain état		
8	(1,1,0,0,0,1,0,0)	Coa	16	11	(2,0,0,1,1,0,0,0)	Coa	7		
		Rec	10'			Coa	8'		
		$A\phi \quad \dagger a\phi$ $\dagger A\phi a\phi$	Conv			8	$A\phi A\phi a\phi \quad \dagger$ $\dagger \quad \dagger \quad \dagger a\phi$	Coa	12
		Conv	8'			Conv	10		
8'	(2,0,0,0,0,1,0,0)	Conv	6	11'	(1,1,0,2,0,0,0,0)	Conv	10'		
		Coa	5			Conv	10*		
		Rec	11'			Conv	11		
		$A\phi A\phi a\phi$ $\dagger \quad \dagger a\phi$	Conv			8	Coa	6	
9	(0,0,1,1,1,0,0,0)	Conv	8'	12	(1,0,0,1,0,0,1,0)	Coa	9'		
		Conv	6'			$A\phi \quad \dagger a\phi a\phi$ $\dagger A\phi \quad \dagger \quad \dagger$	Coa	12'	
		Conv	6*			Conv	10		
		Coa	16			Conv	10'		
9'	(0,0,1,2,0,0,0,0)	Rec	10'	12'	(1,0,0,1,0,0,1,0)	Conv	10*		
		Conv	9			Conv	11'		
		$a\phi \quad \dagger A\phi$ $\dagger a\phi A\phi$	Conv			9'	Coa	5'	
		Conv	7			Rec	11		
9*	(0,0,1,2,0,0,0,0)	Conv	7*	12*	(1,0,0,1,0,0,1,0)	Conv	12		
		Coa	6'			$A\phi A\phi a\phi$ $a\phi \quad \dagger \quad \dagger$	Conv	13	
		Coa	7'			Conv	13'		
		Conv	7*			Coa	4'		
10	(2,0,0,2,0,0,0,0)	Conv	7*	13	(1,0,0,0,1,0,1,0)	Conv	12'		
		Coa	6'			$A\phi \quad \dagger \quad \dagger$ $a\phi A\phi a\phi$	Conv	13	
		Coa	7'			Conv	13'		
		Conv	10			Coa	4*		
10'	(1,1,0,1,1,0,0,0)	Conv	11	13'	(0,1,0,1,0,0,1,0)	Coa	5*		
		Conv	11'			$A\phi A\phi \quad \dagger$ $a\phi \quad \dagger a\phi$	Coa	14	
		Coa	8			Rec	10*		
		Coa	9			Conv	12		
10*	(2,0,0,0,2,0,0,0)	Conv	11	13*	(0,1,0,1,0,0,1,0)	Conv	12'		
		Conv	11'			Conv	13		
		Coa	6*			Coa	15		
		Coa	7*			Rec	10'		
11	(2,0,0,0,2,0,0,0)	Conv	11	14	(0,0,0,0,0,0,2,0)	Conv	12		
		Conv	11'			$A\phi \quad \dagger a\phi$ $a\phi A\phi \quad \dagger$	Conv	12'	
		Coa	13			Conv	13'		
		Conv	10*			Coa	1		
11'	(1,1,0,1,1,0,0,0)	Conv	11'	15	(0,0,0,0,0,0,1,1)	Rec	13		
		Coa	9			$A\phi A\phi$ $a\phi a\phi$	Rec	13'	
		Coa	13'			$A\phi a\phi$ $a\phi A\phi$	Rec	13'	
		Conv	10'			Conv	13		
11*	(2,0,0,0,2,0,0,0)	Conv	11	16	(0,0,1,0,0,1,0,0)	Rec	8		
		Conv	11'			Rec	9		
		Coa	13			$A\phi a\phi$ $A\phi a\phi$	Conv	4	
		Conv	10*			Conv	5		

TABLEAU 4.5. Etats ancestraux et transitions d'état pour trois lignées dont exactement deux de même type A ou a .

Etat	\underline{n}	Evénements possibles	Prochain état	Etat	\underline{n}	Evénements possibles	Prochain état
4	(0,0,1,1,0,0,0,0)	Rec	6	7'	(1,0,0,2,0,0,0,0)	Coa	3
	$A\phi a\phi$	Conv	4		$A\phi a\phi a\phi$	Conv	7
	$A\phi \uparrow$	Conv	2		$\uparrow \uparrow \uparrow$	Conv	7'
		Conv	3			Conv	7*
4'	(1,0,0,0,0,0,0,1)	Rec	6	7*	(1,0,0,0,2,0,0,0)	Coa	2
	$A\phi a\phi$	Conv	4'			Coa	5*
	$\uparrow A\phi$	Conv	4*		$A\phi \uparrow \uparrow$	Conv	7
4*	(0,1,0,0,0,0,0,1)	Coa	1		$\uparrow a\phi a\phi$	Conv	7'
	$\uparrow a\phi$	Rec	6*			Conv	7*
	$A\phi A\phi$	Conv	4'				
		Conv	4*				
5	(1,0,0,0,0,1,0,0)	Rec	7				
	$A\phi a\phi$	Conv	5				
	$\uparrow a\phi$	Conv	2				
		Conv	3				
5'	(0,0,0,1,0,0,1,0)	Rec	7				
	$A\phi a\phi$	Conv	5'				
	$a\phi \uparrow$	Conv	5*				
5*	(0,0,0,0,1,0,1,0)	Coa	1				
	$A\phi \uparrow$	Rec	7*				
	$a\phi a\phi$	Conv	5'				
		Conv	5*				
6	(1,1,0,1,0,0,0,0)	Coa	4				
		Coa	4'				
	$A\phi \uparrow a\phi$	Conv	6				
	$\uparrow A\phi \uparrow$	Conv	6'				
		Conv	6*				
6'	(2,0,0,1,0,0,0,0)	Coa	3				
		Conv	6				
	$A\phi A\phi a\phi$	Conv	6'				
	$\uparrow \uparrow \uparrow$	Conv	6*				
6*	(2,0,0,0,1,0,0,0)	Coa	2				
		Coa	4'				
	$A\phi A\phi \uparrow$	Conv	6				
	$\uparrow \uparrow a\phi$	Conv	6'				
		Conv	6*				
7	(1,0,0,1,1,0,0,0)	Coa	5				
	ou (0,1,0,1,1,0,0,0)	Coa	5'				
		Conv	7				
	$A\phi a\phi \uparrow$	Conv	7'				
	ou $a\phi \uparrow \uparrow$	Conv	7*				
		Conv	7*				

CONCLUSION

Le but de ce mémoire était d'utiliser les processus ancestraux pour étudier les mesures de polymorphisme de nucléotides.

L'analyse de la mesure de déséquilibre de liaison sous un modèle avec recombinaison intragénique utilise le processus ancestral exact. Cependant cette application peut être simplifiée en approximant le processus ancestral exact par le processus SMC (sequentially Markov chain). La correction d'une formule de covariance utilisée dans le calcul d'une mesure d'association à deux sites n'a pas changé significativement les conclusions. Une comparaison des deux processus a démontré une très faible diminution du déséquilibre de liaison sous le processus approximé.

La conversion génique intrachromosomique a les propriétés d'un mécanisme responsable de maintenir l'homogénéité d'une famille de gènes répétés en tandem, tout en permettant à un motif d'ADN variant de remplacer les séquences existantes. Nous avons donc considéré le processus ancestral exact sous le modèle de Moran avec conversion et recombinaison.

Ainsi, le processus ancestral limite sous un modèle à une infinité de sites avec recombinaison intragénique et conversion intrachromosomique a permis d'obtenir analytiquement, de façon robuste et simple comparée au processus de diffusion sous le modèle de Wright-Fisher, les espérances de trois mesures de variation génétique, π_w , π_b , D_{sum} . Ces mesures de variabilité sont par la suite utilisées pour estimer les paramètres de mutation, de recombinaison intragénique et de conversion intrachromosomique, responsables de cette variabilité génétique.

Notons qu'en étudiant un événement de conversion qui se produit entre une paire de séquences, la longueur du brin d'ADN de conversion génétique pourrait

être prise en considération (Wiuf et Hein 2002). La distribution de la longueur du brin de conversion génétique indique que les L sites de nucléotides sur les gènes dupliqués ne sont pas indépendants. Les trois estimateurs déduits ici, $\mathbb{E}(\pi_w)$, $\mathbb{E}(\pi_b)$, et $\mathbb{E}(D_{sum})$, tiennent sans l'hypothèse d'indépendance des L sites. Cependant ceci n'est pas le cas pour leurs variances.

BIBLIOGRAPHIE

- [1] Cargill, M., D. Altshuler, J. Ireland, P. Sklar, and K. Ardlie (1999), Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22, pp.231-238.
- [2] Drake, J.W., B. Charlesworth, and D. Charlesworth (1998), Rates of spontaneous mutations. *Genetics*, 148, pp. 1667-1686.
- [3] Ewens W.J. (2004), Mathematical Population Genetics. Theoretical Introduction. *Springer*; 2nd edition.
- [4] Feldman, M.W. (1966), On the offspring number distribution in a genetic population. *Journal of Applied Probability*, 3, pp. 129-141.
- [5] Fisher R A. (1930), The Genetical Theory of Natural Selection. *Oxford : Clarendon Press*, 26 272 p.
- [6] Griffiths, R.C. (1981), Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology*, 19, pp. 169-186.
- [7] Hill W.G., and Robertson, A. (1968), Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38, pp. 226-231.
- [8] Hudson R.R. (1983), Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23, pp. 183-201.
- [9] Hudson R.R. (1990), Gene genealogies and the coalescent process. *Oxford Survey in Evolutionary Biology*, 7, pp. 1-44.
- [10] Innan, H. (2002a), A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. *Genetics*, 161, pp. 865-872.
- [11] Innan, H. (2002b), The coalescent and infinite-site model of a small multigene family. *Genetics*, 163, pp. 803-810.

- [12] Inomata, N., H. Shibata, E. Okuyama and T. Yamazaki (1995), Evolutionary relationships and sequence variation of α -amylase variants encoded by duplicated genes in the *Amy* locus of *Drosophila melanogaster*. *Genetics*, 141, pp. 237-244.
- [13] Kimura, M. (1969), The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations. *Genetics*, 61, pp.893-903.
- [14] Kingman J.F.C. (1982a), The coalescent. *Stoch. Proc. Appl.* 13, pp. 235-248.
- [15] Kingman J.F.C. (1982b), On the genealogy of large populations. *J. Appl. Probab.*, 19A, pp. 27-43.
- [16] Kingman J.F.C. (1982c), Exchangeability and the evolution of large populations. *Exchangeability in Probability and Statistics*, pp. 97-112.
- [17] Muller, H.J. (1932), Some genetic aspects of sex. *The American Naturalist*, 66, pp. 118-138.
- [18] McVean, G.A.T. (2002), A genealogical interpretation of linkage disequilibrium. *Genetics*, 162, pp. 987-991.
- [19] McVean, G.A.T. and Cardin N.J. (2005), Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 360, pp. 1387-1393.
- [20] Moran, P.A.P. and G.A. Watterson (1959), The genetic effects of family structure in natural population. *Australian Journal of Biological Sciences*, 12, pp. 1-15.
- [21] Nagylaki T. and Petes T.(1982), Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. *Genetics*, 100, pp. 315-337.
- [22] Nei M., Roychoudhury A.K. (1974), Sampling variances of heterozygosity and genetic distance. *Genetics*, 76(2), pp. 379-390.
- [23] Nagylaki T. (1983), The evolution of multigene families under intrachromosomal gene conversion. *Genetics*, 106, pp. 529-548.

- [24] Ohta T. and Kimura M. (1971), Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics*, 68, pp. 571-580.
- [25] Pluzhnikov, A. and Donnelly P. (1996), Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, 144, pp. 1247-1262.
- [26] Simonsen K.L. and Churchill G.A. (1997), A Markov chain model of coalescence with recombination. *Theoretical Population Biology* 52, pp. 43-59.
- [27] Stahl F.W. (1994), The Holliday junction on its thirtieth anniversary. *Genetics* 138, pp. 241-246.
- [28] Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T., Stanley, S.E., Jiang, R., Messer, C.J., Chew, A., Han, J.H., *et al.* (2001), Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293, pp. 489-493.
- [29] Tajima F. (1983), Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105, pp. 437-460.
- [30] Tavaré S. (1984), Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26, pp. 119-164.
- [31] Wakeley, J. (2008), Coalescent Theory : An Introduction. *Roberts & Company Publishers, Greenwood Village, Colorado.*
- [32] Watterson, G.A. (1975), On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7, pp. 256-276.
- [33] Wiuf, C. and J.J. Hein (2000), The coalescent with gene conversion. *Genetics*, 155, pp. 451-462.
- [34] Wright S. (1931), Evolution in Mendelian populations. *Genetics*, 16, pp. 97-159.

Annexe A

PREUVES DÉTAILLÉES DE QUELQUES RÉSULTATS

A.1. MESURE DE POLYMORPHISME DE SÉQUENCES D'ADN - π

Dans le modèle de coalescent, le nombre de mutations le long d'une lignée ancestrale est un processus de Poisson d'intensité $\frac{\theta}{2}$. Considérons que la longueur d'une lignée est égale à t . Alors le nombre de mutations total K sur cette lignée, est de loi de Poisson de paramètre $\frac{\theta t}{2}$. Ainsi, on a

$$\mathbb{E}[K] = \frac{\theta t}{2}.$$

Il s'ensuit que

$$\begin{aligned}\mathbb{E}[\pi] &= \mathbb{E}\left[\frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=1+1}^n k_{ij}\right] = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=1+1}^n \mathbb{E}[k_{ij}] \\ &= \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=1+1}^n \frac{\theta}{2} \mathbb{E}[2T_{ij}] \\ &= \frac{\theta}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=1+1}^n \mathbb{E}[T_{ij}] = \theta,\end{aligned}$$

puisque le temps jusqu'à à un événement de coalescence dans un échantillon de taille 2 suit une loi exponentielle de paramètre 1.

A.2. MESURE DE LIAISON EN TERMES DE COVARIANCES DE TEMPS DE COALESCENCE

Pour deux loci (A) et (B), avec les allèles A, a , et B, b , respectivement, on considère la mesure du ratio des espérances (équation (1.2.1) dans Ohta & Kimura 1971) donnée par

$$\sigma_d^2 = \frac{\mathbb{E}[D^2]}{\mathbb{E}[p_A p_a p_B p_b]},$$

où $D = p_{AB} - p_A p_B$. Dans un contexte généalogique on considère cette mesure de liaison dans le cas où les deux loci sont polymorphes et que chaque polymorphisme est le résultat d'une seule mutation [single-nucleotide polymorphisms (SNP's)]. Lorsqu'il n'y a que deux allèles aux deux loci, D^2 est indépendant de la définition des allèles. Ici on suppose que le taux de mutation à chaque locus $\frac{\theta}{2}$ est petit et que les allèles mutants (plus récents) sont A et B aux loci (A) et (B), respectivement.

On prend un grand échantillon de n séquences choisies au hasard dans lequel tous les allèles sont représentés et on définit la variable $\delta_x^{(Y)}$ qui prend la valeur 1 si l'allèle Y est au locus (Y) de la séquence x , et qui prend la valeur 0 si l'allèle y est au locus (Y) de la séquence x étant donné que le locus (Y) se présente sous la forme d'un de deux allèles Y ou y . En particulier

$$\delta_i^{(A)} = \begin{cases} 1 & \text{si l'allèle } A \text{ est au locus } (A) \text{ de la séquence } i \quad \begin{matrix} A \circ \\ \vdots \\ \dagger \end{matrix}, \\ 0 & \text{sinon,} \end{cases}$$

$$\delta_k^{(B)} = \begin{cases} 1 & \text{si l'allèle } B \text{ est au locus } (B) \text{ de la séquence } k \quad \begin{matrix} B \circ \\ \vdots \\ \dagger \end{matrix}, \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, on a

$$\delta_i^{(A)}(1 - \delta_j^{(A)}) = \begin{cases} 1 & \text{si les deux séquences } i \text{ et } j \text{ dans cet ordre sont de types } \begin{matrix} A \circ & a \circ \\ \vdots & \vdots \\ \dagger & \dagger \end{matrix}, \\ 0 & \text{sinon.} \end{cases}$$

Pour le dénominateur de l'équation (1.2.1), on a

$$\mathbb{E}[p_{AP_a}p_{BP_b}] = \mathbb{E}\left(\mathbb{E}\left(\delta_i^{(A)}(1 - \delta_j^{(A)})\delta_k^{(B)}(1 - \delta_l^{(B)})\middle|p_A, p_B\right)\right), \quad (\text{A.2.1})$$

où i, j, k, l désignent quatre séquences distinctes.

• : Événement de mutation au locus (A)

○ : Événement de mutation au locus (B)

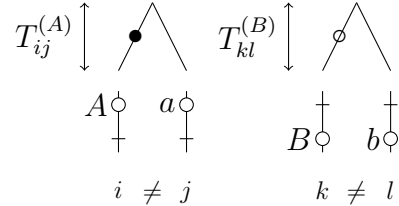


FIGURE A.1. Événements de mutation qui définissent l'espérance conditionnelle (A.2.1).

On peut exprimer l'espérance conditionnelle (A.2.1) par la probabilité qu'au moins une mutation se produise sur la branche ancestrale au locus (A) de la séquence i et non ancestrale au locus (A) de la séquence j et qu'au moins une mutation se produise sur la branche ancestrale au locus (B) de la séquence k et non ancestrale au locus (B) de la séquence l (voir la figure A.1), divisée par la probabilité qu'au moins une mutation se produise à chacun des loci (A) et (B) dans la généalogie de l'échantillon. Alors on a

$$\mathbb{E}[p_{AP_a}p_{BP_b}] = \frac{\mathbb{E}\left[1 - e^{-\frac{\theta}{2}(T_{ij}^{(A)} + T_{kl}^{(B)})}\right]}{\mathbb{E}\left[1 - e^{-\frac{\theta}{2}(\tau^{(A)} + \tau^{(B)})}\right]},$$

où $\tau^{(A)}$ et $\tau^{(B)}$ désignent les longueurs totales de toutes les branches ancestrales à l'échantillon aux loci (A) et (B), respectivement. En faisant tendre θ vers 0, on a

$$\begin{aligned} \mathbb{E}[p_{AP_a}p_{BP_b}] &= \frac{\mathbb{E}\left[T_{ij}^{(A)}T_{kl}^{(B)}\right]}{\mathbb{E}\left[\tau^{(A)}\tau^{(B)}\right]}, \\ &= \frac{\mathbb{E}[t]^2 + \text{Cov}\left[T_{ij}^{(A)}T_{kl}^{(B)}\right]}{\mathbb{E}\left[\tau^{(A)}\tau^{(B)}\right]}, \end{aligned} \quad (\text{A.2.2})$$

où t représente un temps de coalescence de deux branches à un locus. De façon analogue pour le numérateur de l'équation (1.2.1), on a

$$\mathbb{E}[D^2] = \mathbb{E}[p_{AB}^2] - 2\mathbb{E}[p_{AB}p_{APB}] + \mathbb{E}[p_{APB}^2], \quad (\text{A.2.3})$$

où les trois derniers termes sont respectivement, la probabilité que les séquences i et j soient identiques au locus (A) et au locus (B), la probabilité que les séquences i et j soient identiques au locus (A) et les séquences i et k au locus (B), et la probabilité que les séquences i et j soient identiques au locus (A) et les séquences k et l au locus (B).

On définit

$$\delta_i^{(AB)} = \begin{cases} 1 & \text{si la séquence } i \text{ est de type } \begin{array}{c} A \circ \\ B \circ \end{array}, \\ 0 & \text{sinon.} \end{cases}$$

Alors on a,

$$\mathbb{E}[p_{AB}^2] = \mathbb{E}\left(\mathbb{E}\left(\delta_i^{(AB)}\delta_j^{(AB)}\right)\middle|p_{AB}\right). \quad (\text{A.2.4})$$

On peut exprimer l'espérance conditionnelle (A.2.4) par la probabilité qu'au moins une mutation se produise sur la branche ancestrale au locus (A) des séquences i et j et qu'au moins une mutation se produise sur la branche ancestrale au locus (B) des séquences i et j (voir la figure A.2), divisée par la probabilité qu'au moins une mutation se produise à chacun des loci (A) et (B) dans la généalogie de l'échantillon. On a

$$\mathbb{E}[p_{AB}^2] = \frac{\mathbb{E}\left[1 - e^{-\frac{\theta}{2}(I_{ij}^{(A)} + I_{ij}^{(B)})}\right]}{\mathbb{E}\left[1 - e^{-\frac{\theta}{2}(\tau^{(A)} + \tau^{(B)})}\right]},$$

où $I_{ij}^{(A)}$ est la longueur de la branche à partir du MRCA des séquences i et j au locus (A) jusqu'au MRCA de tout l'échantillon au locus (A) (voir la figure A.2), et de façon analogue pour $I_{ij}^{(B)}$. En faisant tendre θ vers 0, on trouve

$$\mathbb{E}[p_{AB}^2] = \frac{\mathbb{E}\left[I_{ij}^{(A)} I_{kl}^{(B)}\right]}{\mathbb{E}\left[\tau^{(A)} \tau^{(B)}\right]}.$$

En faisant

$$I_{ij}^{(A)} = T^{(A)} - T_{ij}^{(A)},$$

où $T_{ij}^{(A)}$ est le temps jusqu'à la coalescence des séquences i et j au locus (A) et $T^{(A)}$ est le temps total jusqu'au MRCA de tout l'échantillon au locus (A) et de

façon analogue $I_{ij}^{(B)} = T^{(B)} - T_{ij}^{(B)}$, on trouve

$$\mathbb{E}[p_{AB}^2] = \frac{\mathbb{E}[(T^{(A)} - T_{ij}^{(A)})(T^{(B)} - T_{ij}^{(B)})]}{\mathbb{E}[\tau^{(A)}\tau^{(B)}]}. \quad (\text{A.2.5})$$

De façon analogue pour la seconde probabilité du numérateur (A.2.3), on a

$$\mathbb{E}[p_{AB}p_{APB}] = \mathbb{E}\left(\mathbb{E}\left(\delta_i^{(AB)}\delta_j^{(A)}\delta_k^{(B)}\right)\middle|p_{AB}, p_A, p_B\right). \quad (\text{A.2.6})$$

On peut exprimer l'espérance conditionnelle (A.2.6) par la probabilité qu'au moins une mutation se produise sur la branche ancestrale au locus (A) des séquences i et j et qu'au moins une mutation se produise sur la branche ancestrale au locus (B) des séquences i et k , divisée par la probabilité que l'échantillon contienne les deux mutants. On a

$$\begin{aligned} \mathbb{E}[p_{AB}p_{APB}] &= \frac{\mathbb{E}[I_{ij}^{(A)}I_{ij}^{(B)}]}{\mathbb{E}[\tau^{(A)}\tau^{(B)}]}, \\ &= \frac{\mathbb{E}[(T^{(A)} - T_{ij}^{(A)})(T^{(B)} - T_{ik}^{(B)})]}{\mathbb{E}[\tau^{(A)}\tau^{(B)}]}. \end{aligned} \quad (\text{A.2.7})$$

Finalement pour la troisième probabilité du numérateur (A.2.3), on a

$$\mathbb{E}[p_A^2p_B^2] = \mathbb{E}\left(\mathbb{E}\left(\delta_i^{(A)}\delta_j^{(A)}\delta_k^{(B)}\delta_l^{(B)}\right)\middle|p_A, p_B\right), \quad (\text{A.2.8})$$

$$= \frac{\mathbb{E}[(T^{(A)} - T_{ij}^{(A)})(T^{(B)} - T_{kl}^{(B)})]}{\mathbb{E}[\tau^{(A)}\tau^{(B)}]}, \quad (\text{A.2.9})$$

car l'espérance conditionnelle (A.2.8) peut être exprimée par la probabilité qu'au moins une mutation se produise sur la branche ancestrale au locus (A) des séquences i et j et qu'au moins une mutation se produise sur la branche ancestrale au locus (B) des séquences k et l , divisée par la probabilité que l'échantillon contienne les deux mutants.

Ainsi en remplaçant les expressions (A.2.5), (A.2.7), et (A.2.9) dans le numérateur (A.2.3), on a pour $\mathbb{E}[D^2]$ l'expression

$$\frac{\text{Cov}[T_{ij}^{(A)}, T_{ij}^{(B)}] - 2\text{Cov}[T_{ij}^{(A)}, T_{ik}^{(B)}] + \text{Cov}[T_{ij}^{(A)}, T_{kl}^{(B)}]}{\mathbb{E}[\tau^{(A)}\tau^{(B)}]},$$

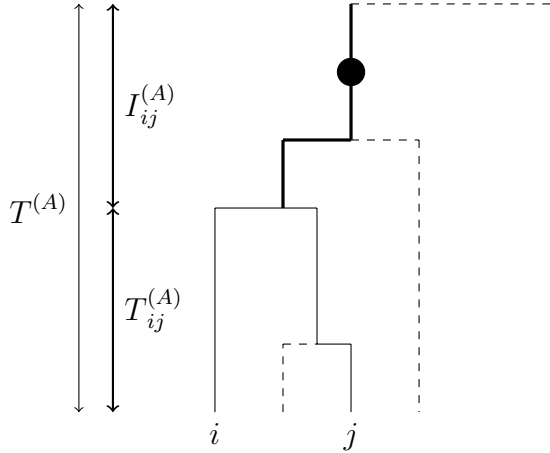


FIGURE A.2. Les statistiques de la généalogie au locus (A) de deux séquences i et j prises dans un échantillon.

Source: McVean 2002.

où

$$\begin{aligned} \mathbb{C}ov[T_{ij}^{(A)}, T_{ij}^{(B)}] &= \mathbb{E}[(T^{(A)} - T_{ij}^{(A)})(T^{(B)} - T_{ij}^{(B)})] \\ \mathbb{C}ov[T_{ij}^{(A)}, T_{ik}^{(B)}] &= \mathbb{E}[(T^{(A)} - T_{ij}^{(A)})(T^{(B)} - T_{ik}^{(B)})] \\ \mathbb{C}ov[T_{ij}^{(A)}, T_{kl}^{(B)}] &= \mathbb{E}[(T^{(A)} - T_{ij}^{(A)})(T^{(B)} - T_{kl}^{(B)})]. \end{aligned}$$

On rappelle d'autre part l'expression (A.2.2), c'est-à-dire

$$\mathbb{E}[p_{AP_a} p_{BP_b}] = \frac{\mathbb{E}[t]^2 + \mathbb{C}ov[T_{ij}^{(A)} T_{kl}^{(B)}]}{\mathbb{E}[\tau^{(A)} \tau^{(B)}]}.$$

Mcvean (2002) a ainsi pu exprimer l'équation (1.2.1) en termes de covariances des temps de coalescence aux deux loci pour différentes configurations de chromosomes :

$$\sigma_d^2 = \frac{C_{ij,ij} - 2C_{ij,ik} + C_{ij,kl}}{\mathbb{E}[t]^2 + C_{ij,kl}},$$

où $C_{ij,kl}$ est la covariance entre le temps de coalescence, au locus (A), entre deux séquences échantillonnées i et j et le temps de coalescence au locus (B) entre deux séquences échantillonnées k et l (voir les covariances des états 1, 2, et 3 de la section (2.2.1)). Ici $\mathbb{E}[t]$ est l'espérance du temps de coalescence d'une paire à un locus. Sous le coalescent standard, $\mathbb{E}[t] = 1$. Ces covariances sous le coalescent

standard sont dérivées en résolvant un système d'équations linéaires (Griffiths 1981 ; voir aussi Pluzhnikov et Donnelly 1996).

A.3. PROCESSUS DE DIFFUSION

L'approximation par un processus de diffusion est une solution pour étudier les effets de la dérive génétique ainsi que d'autres facteurs d'évolution sur les changements de la composition d'une population dans le sens conventionnel d'écoulement du temps.

Soit un processus stochastique à n composantes à temps continu $\{\mathbf{X}_t, t \geq 0\}$, où $\mathbf{X}_t = (X_t(1), \dots, X_t(n))$ pour $t \geq 0$. On définit $\Delta_h \mathbf{X}_t$, l'incrément dans le processus de l'instant t à l'instant $t + h$, c'est-à-dire :

$$\Delta_h \mathbf{X}_t = \mathbf{X}_{t+h} - \mathbf{X}_t.$$

Définition A.3.1. *Un processus markovien à temps continu $\{\mathbf{X}_t, t \geq 0\}$ est dit un processus de diffusion s'il vérifie les conditions suivantes :*

$$\mathbb{E} \left[\Delta_h X_t(i) \middle| \mathbf{X}_t = \mathbf{x} \right] = a_i(\mathbf{x})h + o(h) \text{ pour } i = 1, \dots, n, \quad (\text{A.3.1})$$

$$\mathbb{E} \left[\Delta_h X_t(i) \Delta_h X_t(j) \middle| \mathbf{X}_t = \mathbf{x} \right] = b_{i,j}(\mathbf{x})h + o(h) \text{ pour } i, j = 1, \dots, n, \quad (\text{A.3.2})$$

$$\mathbb{E} \left[\left(\Delta_h X_t(i) \right)^r \middle| \mathbf{X}(t) = \mathbf{x} \right] = o(h) \text{ pour } r \geq 3, \quad (\text{A.3.3})$$

où a_i et $b_{i,j}$ sont des fonctions continues de x , pour $i, j = 1, \dots, n$. Dans ce cas on appelle $\{a_i, i = 1, \dots, n\}$ les moyennes infinitésimales et $\{b_{i,j}, i, j = 1, \dots, n\}$ les covariances infinitésimales du processus $\{\mathbf{X}_t, t \geq 0\}$.

Soit g une fonction bornée assez régulière, en fait $g \in C^2(\mathbb{R}^n)$. Un développement de Taylor donne

$$\begin{aligned} g(\mathbf{X}_h) &= g(\mathbf{X}_0) + \sum_{i=1}^n (X_h(i) - X_0(i)) \frac{dg}{dx_i}(\mathbf{X}_0) \\ &+ \frac{1}{2} \sum_{i,j=1}^n (X_h(i) - X_0(i))(X_h(j) - X_0(j)) \frac{d^2g}{dx_i dx_j}(\mathbf{X}_0) \\ &+ O\left(|\Delta_h X_0(i) \cdot \Delta_h X_0(j) \cdot \Delta_h X_0(k)|\right). \end{aligned}$$

En prenant l'espérance conditionnelle par rapport à $\mathbf{X}_0 = \mathbf{x}$, ensuite en divisant par h et en faisant tendre h vers 0, on a

$$\begin{aligned} \lim_{h \rightarrow 0} \mathbb{E} \left[\frac{g(\mathbf{X}_h) - g(\mathbf{X}_0)}{h} \middle| \mathbf{X}_0 = \mathbf{x} \right] &= \lim_{h \rightarrow 0} \left[\sum_{i=1}^n \frac{\mathbb{E} \left[\Delta_h X_0(i) \middle| \mathbf{X}_0 = \mathbf{x} \right]}{h} \frac{dg}{dx_i}(\mathbf{x}) \right. \\ &+ \frac{1}{2} \sum_{i,j=1}^n \frac{\mathbb{E} \left[\Delta_h X_0(i) \cdot \Delta_h X_0(j) \middle| \mathbf{X}_0 = \mathbf{x} \right]}{h} \frac{d^2 g}{dx_i dx_j}(\mathbf{x}) \\ &\left. + \frac{O \left(\mathbb{E} \left[|\Delta_h X_0(i) \cdot \Delta_h X_0(j) \cdot \Delta_h X_0(k)| \middle| \mathbf{X}_0 = \mathbf{x} \right] \right)}{h} \right]. \end{aligned}$$

L'opérateur L donné par l'équation précédente est appelé le générateur infinitésimal du processus $\{\mathbf{X}_t, t \geq 0\}$. On a donc,

$$L(g(\mathbf{x})) = \frac{d}{dh} \mathbb{E} \left[g(\mathbf{X}_h) \middle| \mathbf{X}_0 = \mathbf{x} \right] \Big|_{h=0}.$$

En se basant sur les conditions de la définition (A.3.1), $\{\mathbf{X}_t, t \geq 0\}$ est un processus de diffusion de moyennes infinitésimales a_i et de variances infinitésimales $b_{i,j}$, si

$$L(g(\mathbf{x})) = \sum_{i=1}^n a_i(\mathbf{x}) \frac{dg}{dx_i}(\mathbf{x}) + \frac{1}{2} \sum_{i,j=1}^n b_{i,j}(\mathbf{x}) \frac{d^2 g}{dx_i dx_j}(\mathbf{x}), \quad (\text{A.3.4})$$

où

$$\begin{aligned} a_i(\mathbf{x}) &= \lim_{h \rightarrow 0} \frac{\mathbb{E} \left[X_h(i) - X_0(i) \middle| \mathbf{X}_0 = \mathbf{x} \right]}{h}, \text{ pour } i = 1, \dots, n, \\ b_{i,j}(\mathbf{x}) &= \lim_{h \rightarrow 0} \frac{\mathbb{E} \left[\Delta_h X_0(i) \cdot \Delta_h X_0(j) \middle| \mathbf{X}_0 = \mathbf{x} \right]}{h}, \text{ pour } i, j = 1, \dots, n. \end{aligned}$$