Université de Montréal

Effet de l'échantillonnage non proportionnel de cas et de témoins sur une méthode de vraisemblance maximale pour l'estimation de la position d'une mutation sous sélection

par

Luc Villandré

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

Le 1er décembre 2008

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Effet de l'échantillonnage non proportionnel de cas et de témoins sur une méthode de vraisemblance maximale pour l'estimation de la position d'une mutation sous sélection

présenté par

Luc Villandré

a été évalué par un jury composé des personnes suivantes :

Alejandro Murua

(président-rapporteur)

Sabin Lessard

(directeur de recherche)

Martin Bilodeau

(membre du jury)

Mémoire accepté le:

Le 1er décembre 2008

TABLE DES MATIÈRES

Liste des figures v

Sommaire ix

Summary x

Chapitre 1. Introduction 1

1.1. Contexte et énoncé de la problématique 1

1.2. Objectifs 6

1.3. Précisions sur la terminologie 7

Chapitre 2. Modèles de base en génétique des populations 8

2.1. Modèle de Wright-Fisher [**24**] 8

2.1.1. Approximation du modèle de Wright-Fisher par un processus de diffusion ... 10

2.1.2. Le modèle de Wright-Fisher vu de façon rétrospective 10

2.2. Le modèle coalescent de Kingman 15

2.3. Le graphe de recombinaison ancestral (ARG) [**24**] 19

2.4. Remarques additionnelles sur les modèles de généalogie 25

Chapitre 3. Implémentation d'une méthode de vraisemblance maximale dans le cadre du modèle coalescent avec sélection et recombinaison .. 27

3.1. Effets de la sélection 28

3.2. Première partie : génération de l'échantillon 37

3.2.1. Locus central soumis à un effet de sélection stabilisatrice 37

3.2.2. Locus central soumis à un effet de sélection génique 38

3.2.3. Approximation d'un processus de diffusion par un processus de naissance et de mort basé sur le modèle de Moran 41

3.3. Deuxième partie : Calcul de la vraisemblance de l'échantillon 49

3.3.1. Approche de Griffiths et Tavaré 49

3.3.2. Une méthode de vraisemblance maximale adaptée à la recombinaison et à la sélection 53

3.3.3. L'équation de récurrence : Dérivation et interprétation 54

3.4. Probabilités de transition et calcul de la vraisemblance 58

Chapitre 4. Description de l'échantillon utilisé et résultats 65

4.1. Considérations préliminaires 65

4.2. Résultats 67

4.2.1. Sélection stabilisatrice 67

4.2.2. Sélection génique 73

Chapitre 5. Discussion 78

Annexe A. Le concept de déséquilibre d'appariement (*linkage disequilibrium*) A-i

Annexe B. Le phénomène de recombinaison ..... B-i

B.0.3. La distribution de recombinaison-ségrégation ..... B-i

B.0.4. Les valeurs de *linkage* ..... B-iii

Annexe C. Méthodes d'estimation bayésiennes et non-bayésiennes de paramètres génétiques ..... C-i

Bibliographie ..... C-i

LISTE DES FIGURES

1 Représentation illustrant le lien entre la distance séparant deux loci et le taux de recombinaison entre ceux-ci. La croix désigne l'endroit où est apparue la césure et a été placée arbitrairement. Un carré vide désigne un locus non-ancestral. $R(1-3)$ désigne le taux de recombinaison entre les loci 1 et 3 et doit être compris comme une unité de mesure. $D(i-j)$ désigne la distance entre i et j. 4

2 Arbre représentant une généalogie possible générée à partir d'un modèle de Wright-Fisher avec $N = 9$ (Tiré des notes de Simon Tavaré **[24]**) 12

3 Plusieurs événements de coalescence simultanés..... 13

4 Réalisation du processus de coalescence sans mutation avec $n = 9$..... 16

5 Réalisation du processus de coalescence avec mutation avec $n = 9$. Le X représente une mutation..... 17

6 Une réalisation du graphe de recombinaison ancestral (ARG). "1" désigne un événement de coalescence, "2", un événement de recombinaison entre les loci numéro 2 et numéro 3 et "3", un événement de mutation. Le graphe doit être lu de façon rétrospective, c'est-à-dire de bas en haut..... 20

7 Représentation de différents événements pouvant affecter l'échantillon. Le numéro identifiant chaque événement peut être associé à la liste de taux précédente..... 24

8 Représentation graphique d'une séquence. M désigne le locus sous sélection, qui sera soumis soit à un effet de sélection stabilisatrice, soit à un effet de sélection génique. ρ est la distance entre les loci 1 et 2 et ρ_0 est la distance entre le locus 1 et le locus sous sélection..... 28

9 Cycle de vie des individus d'une population de la génération $\tau - 1$ à la génération τ. 30

10 Une réalisation du graphe de recombinaison ancestral quand il y a sélection au locus central. Le cloisonnement créé par la sélection ou, en d'autres mots, la séparation en deux famille distinctes, se doit d'être remarqué. 36

11 Les deux types d'événements de recombinaison. Les allèles dits «attribués» ne sont pas ancestraux à la séquence résultante. Ils sont choisis aléatoirement en fonction de la fréquence des séquences de la famille des cas et des témoins au moment de l'événement de recombinaison. Le M indique quelle séquence a transmis l'allèle au locus sous sélection. 38

12 Un événement de recombinaison. M indique l'allèle au locus sous sélection de la séquence i et de la séquence j. Par convention, la séquence-parent j transmettra toujours à la séquence i l'allèle au locus sous sélection et la séquence k aura toujours une configuration inconnue au locus sous sélection. Malgré tout, elle devra se voir attribuer une famille. 57

13 Graphique illustrant la progression du niveau de vraisemblance en fonction du nombre d'itérations. Les valeurs présentées sont arbitraires. 69

14 Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le premier graphique a été généré à partir d'une distance proposée de 2,5, le deuxième à partir d'une distance proposée de 2,7 et le troisième à partir d'une distance proposée de 3. 70

15 Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le premier graphique a été généré à partir d'une distance proposée de 4, le deuxième à partir d'une distance proposée de 3,8. Le troisième devrait aussi être généré à partir d'une distance de 2,5. Puisqu'il est déjà affiché à l'illustration (14), il n'est pas reproduit ici. 71

16 Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le premier graphique a été généré à partir d'une distance proposée de 2, le deuxième à partir d'une distance proposée de 2,3. Le troisième devrait aussi être généré à partir d'une distance de 3. Puisqu'il est déjà affiché à l'illustration (14), il n'est pas reproduit ici. 71

17 Courbes de vraisemblance. Chaque courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le premier graphique a été généré à partir d'une distance proposée de 2,5, le deuxième à partir d'une distance proposée de 2. Les séquences de l'échantillon comportent 7 loci. 72

18 Courbe de vraisemblance. La courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le graphique a été généré à partir d'une distance proposée de 1,5. Les séquences de l'échantillon comportent 7 loci. 72

19 Courbe de vraisemblance. La courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le graphique a été généré à partir d'une distance proposée de 2. Les séquences de l'échantillon comportent 11 loci. 73

20 Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Les deux courbes ont été générées à partir d'une distance proposée de 1. La courbe à gauche a été produite à partir d'un échantillon de séquences à 15 loci, tandis que celle à droite a été produite à partir d'un échantillon de séquences comportant 17 loci. 74

21 Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Les deux courbes ont été générées à partir d'une distance proposée de 1. La courbe à gauche a été produite à partir d'un échantillon de séquences à 19 loci, tandis que celle à droite a été produite à partir d'un échantillon de séquences comportant 21 loci. 74

22 Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Les deux courbes ont été générées à partir d'une distance proposée de 1. La courbe à gauche a été produite à partir d'un échantillon de séquences à 23 loci, tandis que celle à droite a été produite à partir d'un échantillon de séquences comportant 25 loci. 75

23 Courbes de vraisemblance. Chaque courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. La courbe de gauche a

été générée à partir d'une distance proposée de 2.85. La courbe à droite a été produite à partir d'une distance proposée de 3,05. 75

24 Courbes de vraisemblance. Chaque courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. La courbe de gauche a été générée à partir d'une distance proposée de 3.35. La courbe à droite a été produite à partir d'une distance proposée de 3,6..... 76

25 Courbe de vraisemblance. La courbe a été générée à partir de 3 000 000 de graphes. La ligne pointillée correspond à la distance véritable. La distance proposée était de 4..... 77

26 Deux types de recombinaison. Dans le scénario 1, un seul chiasme est formé. Nous observons donc ce qu'on appelle un événement de «crossover». Dans le scénario 2, deux chiasmes sont formés. Nous sommes donc en présence de conversion génétique. B-ii

SOMMAIRE

Après une synthèse de différents concepts et modèles élémentaires en génétique statistique, une nouvelle méthode de vraisemblance maximale en cartographie génétique, basée sur une formule récursive et améliorée grâce à l'utilisation du principe d'*importance sampling*, est introduite. Elle est en fait une extension d'une méthode conçue et testée par Larribe et Lessard [**15**] au cas où il y a sélection naturelle à un locus donné. Le but principal sera de déterminer si l'échantillonnage arbitraire de cas et de témoins qu'ils avaient proposé permet une estimation non biaisée de la position d'un site mutant sous sélection. Des simulations basées sur la méthode de Monte Carlo sont réalisées à cette fin. En théorie, la méthode requiert un échantillonnage entièrement aléatoire des cas et des témoins. L'échantillonnage arbitraire a été employé afin de pallier à la rareté de l'allèle sous sélection dans la population. On conclut que malgré une convergence lente, la méthode d'échantillonnage retenue permet tout de même de situer assez précisément la site mutant ciblé. L'inclusion du phénomène de sélection n'a donc pas d'impact sur sa nature non biaisée. L'ouvrage se termine par une discussion couvrant les différentes forces et faiblesses de la méthode, ainsi que des améliorations possibles à considérer dans des projets subséquents.

Mots clés : Cartographie génétique, échantillonnage, sélection naturelle, vraisemblance, Monte Carlo, récursion, importance sampling.

SUMMARY

After a summary of different elementary models and concepts in statistical genetics, a new maximum likelihood method in gene mapping based on a recursive formula and improved by the use of importance sampling is introduced. It is in fact an extension of a method previously conceived and tested by Larribe and Lessard [**15**] to the case where natural selection is acting at a given locus. The main goal will be to ascertain whether the arbitrary sampling of cases and controls they had proposed still leads to an unbiased estimation of the position of a mutant site under selection. Simulations based on the Monte Carlo method will be used for this purpose. In theory, the method requires random sampling of cases and controls. Arbitrary sampling has been employed in order to compensate for the rarity of the allele under selection. We are able to conclude that despite slow convergence, arbitrary sampling still allows us to estimate the position of the selected mutant site rather precisely. The inclusion of selection does not make arbitrary sampling a source of bias. The current work ends with a discussion about the different strengths and weaknesses of the method, as well as possible improvements worth considering in subsequent projects.

Key words : Gene mapping, sampling, natural selection, likelihood, Monte Carlo, recursion, importance sampling.

Chapitre 1

INTRODUCTION

1.1. CONTEXTE ET ÉNONCÉ DE LA PROBLÉMATIQUE

Depuis la formulation dans «The Origin of Species» des principes fondamentaux sous-tendant l'évolution au sens darwinien, de nombreux efforts ont été faits pour expliquer de façon rigoureuse l'origine du polymorphisme[1] [2] observé dans plusieurs populations d'êtres vivants. Ronald Fisher, le père des statistiques modernes, s'est assez tôt dans sa carrière académique intéressé à ce problème. En fait, en plus des méthodes d'analyse de variance, de test d'hypothèses et de calcul de vraisemblance, on lui doit aussi la paternité d'une des premières formulations mathématiques de la théorie darwinienne de l'évolution. Encore aujourd'hui, le modèle qu'il a conçu conjointement avec Sewall Wright est à la base de la plupart des travaux de modélisation effectués dans le domaine. En essence, ce modèle avait comme objectif premier de décrire l'évolution temporelle de la composition d'une population d'individus présentant un trait particulier. Sa formulation la plus élémentaire est donnée à la section 2.1. Bien sûr, au cours des dernières décennies, le modèle en question a été modifié afin de l'adapter à certains phénomènes propres à l'évolution des populations tels l'insularité[3], la croissance, la dérive génétique[4] et la sélection. Or, au cours des trente dernières années, deux modèles dérivés de celui de Wright-Fisher ont eu un impact particulièrement considérable dans le domaine qu'on appelle maintenant génétique statistique.

[1] Polymorphisme : Variabilité observée au niveau des phénotypes.
[2] Phénotype : Ensemble des caractéristiques physiques observables d'un être vivant.
[3] Insularité : Séparation physique de populations interreproductibles.
[4] Dérive génétique : Changement dans la composition d'une population finie dû à la reproduction aléatoire des individus.

D'un côté, Kingman [**12**] s'est attardé en 1982 à un cas-limite du modèle de Wright-Fisher, soit la situation où la taille de la population est très grande. Il a aussi changé l'approche temporelle. Il adopte une approche rétrospective, contrairement à ses prédécesseurs qui adoptaient plutôt une approche prospective. En d'autres mots, Kingman considère que l'instant 0 correspond au moment présent et que l'instant t, $t > 0$, correspond à un moment dans le passé. Le processus stochastique qu'il a élaboré pour décrire l'évolution d'un échantillon de gènes tiré de cette population prend la forme d'une chaîne de Markov en temps continu dont le domaine correspond à des partitions de l'ensemble des gènes contenus dans l'échantillon. Ce modèle, appelé «coalescent», reste encore aujourd'hui une référence incontournable en génétique statistique.

Depuis l'introduction du modèle de Wright-Fisher, on savait que le temps jusqu'à l'ancêtre commun de deux gènes choisis au hasard dans une population neutre dont les membres se reproduisent au hasard était distribué géométriquement avec paramètre $2N$ (N est la taille de la population). Puisque $2N$ est généralement grand, en rééchelonnant le temps de façon appropriée, on peut approximer cette distribution par une distribution exponentielle avec la même espérance. Kingman a su généraliser ce processus pour n gènes. Il s'est basé sur deux constats. D'une part, le phénomène de coalescence des lignées peut être généré indépendemment du phénomène de mutation des allèles[5] de ces lignées. D'autre part, il est possible de générer de façon rétrospective la généalogie d'un échantillon de gènes sans se soucier du reste de la population [**18**]. Le modèle coalescent est décrit plus précisément à la section 2.2.

Or, le coalescent, tel que défini par Kingman initialement, parvient difficilement à décrire de façon réaliste l'évolution temporelle dans le portrait génétique d'une population. Entre autres, il ne tient pas compte d'un mécanisme fondamental en génétique, soit le phénomène de recombinaison. Lors de la création des gamètes[6], les chromosomes pairés peuvent s'échanger du matériel génétique. On nomme cet échange «recombinaison». La recombinaison, aussi appelée «crossing-over» ou «enjambement», accroît le nombre d'agencements alléliques possibles et de ce fait, accroît la diversité phénotypique dans une population[7]. Le coalescent a été adapté au phénomène de recombinaison par Griffiths et Marjoram [**5**], qui ont donné le nom de «graphe de recombinaison ancestrale» (ARG) à leur processus. Cette approche, maintenant couramment

[5]Allèle : Variante d'un gène.

[6]Gamète : Cellule reproductrice.

[7]En fait, certaines théories évolutionnistes modernes [**21**] font de la recombinaison, et non de la mutation, le phénomène à l'origine de l'apparition de nouveaux caractères phénotypiques dans une population.

utilisée, est abordée plus en détails à la section 2.3. Le phénomène de recombinaison lui-même présente certaines particularités et est présenté à l'annexe B.

Aucun évolutionniste ne nie le rôle prépondérant de la sélection naturelle dans la propension des profils alléliques avantageux au sein d'une population. En effet, une classe d'individus mieux adaptés à leur environnement aura plus de descendance et verra la proportion qu'elle représente dans la population s'accroître jusqu'à ce qu'un nouvel équilibre soit atteint. Logiquement, les effets de la sélection naturelle devraient pouvoir se traduire au niveau des gènes eux-mêmes. Quelques approches, empruntant les principes de base du modèle coalescent, ont été proposées à cette fin. Krone et Neuhauser [**13**] ont dérivé comme cas limite d'un diagramme de percolation[8] un processus analogue à celui de Kingman, incluant toutefois les effets de sélection, qu'ils ont appelé le graphe de sélection ancestrale. Or, puisque le graphe en question ne présente pas les avantages, notamment la simplicité, du modèle coalescent de Kingman, il est rarement utilisé comme outil d'analyse. En effet, la détermination de la forme du graphe nécessite la connaissance de la configuration de l'ancêtre commun de l'échantillon prélevé (appelé MRCA pour «Most Recent Common Ancestor»). Malheureusement, trop souvent, sa configuration est tout simplement inconnue.

D'autre part, plusieurs années auparavant, Kaplan et Hudson [**9**] [**8**] avaient eux-mêmes proposé un modèle de propension allélique qui tenait compte de la sélection dite «stabilisatrice»[9]. Ils se sont basés sur le principe que sous certaines restrictions, le phénomène de sélection naturelle était analogue à celui de ségrégation allélique[10]. Dans un tel contexte, la configuration génétique au locus[11] sous sélection détermine la famille à laquelle appartient un individu et un événement de recombinaison devient analogue à un événement de migration (si l'on assume qu'une séquence migrante peut choisir sa propre classe comme destination). À l'intérieur d'une même famille, les effets de la sélection naturelle disparaissent et l'évolution temporelle de la population peut être décrite adéquatement avec les outils fournis par le modèle coalescent de Kingman.

---

[8]Diagramme de percolation : Représentation graphique d'une marche aléatoire multiple et coalescente.

[9]Un allèle soumis à un effet de sélection naturelle stabilisatrice voit sa fréquence fixée dans la population.

[10]Ségrégation allélique : Cloisonnement de populations basé sur un critère génétique, phénomène analogue à l'insularité.

[11]Locus : Position précise sur un chromosome, peut être identifié physiquement.

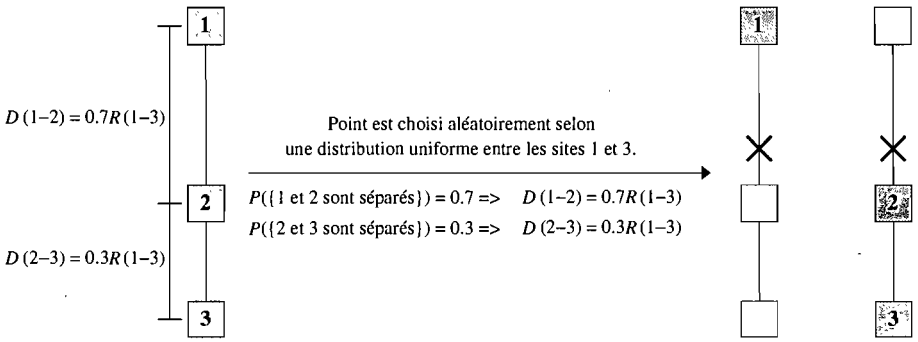

FIG. 1. Représentation illustrant le lien entre la distance séparant deux loci et le taux de recombinaison entre ceux-ci. La croix désigne l'endroit où est apparue la césure et a été placée arbitrairement. Un carré vide désigne un locus non-ancestral. $R(1-3)$ désigne le taux de recombinaison entre les loci 1 et 3 et doit être compris comme une unité de mesure. $D(i-j)$ désigne la distance entre i et j.

Tous ces outils avaient comme but premier de modéliser rigoureusement la généalogie d'un échantillon de séquences d'ADN afin d'en étudier les propriétés ou bien afin de déterminer par des méthodes d'estimation la valeur de paramètres, tels les taux de mutation, de recombinaison ou de sélection, régissant l'évolution génétique d'une population donnée. Or, l'accroissement de la puissance des ordinateurs, couplé avec le développement de ces modèles stochastiques, a rendu possible l'utilisation de méthodes de vraisemblance complète, i.e. basées directement sur le type des allèles que l'on retrouve à un nombre arbitraire de loci polymorphes, au lieu de seulement une statistique quelconque tel le taux d'appariement, défini comme le taux d'association entre des gènes situés à différents sites, ou l'hétérozygocie[12], pour calculer approximativement la valeur de ces paramètres. Le phénomène d'appariement, ayant un fort lien conceptuel avec le phénomène de recombinaison, est décrit plus en détails à l'annexe A.

Un des domaines ayant le plus bénéficié de ces progrès techniques et méthodologiques reste celui de la cartographie génétique. La cartographie génétique est définie comme l'ensemble des méthodes permettant de situer des gènes sur un brin d'ADN. Le code génétique étant composé de millions de bases azotées[13], il peut être difficile de cibler précisément une région comme étant impliquée dans la manifestation d'un caractère phénotypique particulier. Par conséquent, des techniques statistiques ont été développées afin de déterminer avec un certain degré de précision la position de cette région. Un paramètre revêt une importance particulière dans ce domaine : le taux de recombinaison. En effet, si le taux de recombinaison est assumé constant sur une

[12]Hétérozygocie : Fréquence des individus diploïdes hétérozygotes dans le population.
[13]Bases azotées : Substances chimiques composant les molécules d'ADN.

région donnée et s'il n'y a pas d'interférence[14], il existe une relation bijective entre celui-ci et la distance physique séparant deux loci (voir la figure 1). En d'autres mots, dans un tel cas, la distance physique séparant deux loci sera directement proportionelle au taux de recombinaison entre ceux-ci. Ce même taux peut donc être utilisé comme unité de mesure. Les populations étant la plupart du temps très grandes (ou, du moins, assumées très grandes), un échantillon doit être prélevé ou généré aux fins de calculs. Ensuite, en conditionnant sur un taux de recombinaison entre deux marqueurs[15], on doit associer une valeur de vraisemblance à l'échantillon par rapport à la position du gène impliqué dans la manifestation du caractère étudié, celui-ci étant bien sûr compris entre les deux marqueurs.

Dans le même esprit, Larribe, Lessard et Schork **[15]**, en se basant sur certains résultats de Griffiths et Tavaré **[6]**, ont développé une méthode pour déterminer la vraisemblance d'un échantillon de séquences d'ADN par rapport à la position d'un site mutant. Celle-ci fait appel à la méthode de Monte Carlo. Toutefois, elle présente un certain problème lié à l'échantillonnage. En effet, son application nécessiterait en théorie un échantillon aléatoire tiré de la population entière. Or, dans le cas où les allèles mutants au site à situer sont peu fréquents, un échantillon aléatoire contiendra avec une grande probabilité un faible nombre de séquences comportant la version mutante de cet allèle, si ce n'est pas tout simplement aucune. Ainsi, il contiendra très peu d'information sur le site en question et l'estimation de la valeur des paramètres qui y sont liés risque d'être ardue. Pour compenser cette faiblesse, Larribe et Lessard **[14]** ont proposé de tirer dans chaque sous-population de séquences, définie par le gène présent au locus qu'on désire situer, un nombre arbitraire de séquences. Les simulations qu'ils ont réalisées leur ont permis de conclure que le biais introduit par cette méthode d'échantillonnage semble être limité ou inexistant. Remarquons que les auteurs s'étaient concentrés sur une population neutre, i.e. non-affectée par les effets de la sélection naturelle. Or, ce phénomène étant fréquent et fondamental, une extension du modèle pour l'inclure serait souhaitable. De plus, vu la formulation du modèle, cette généralisation pourrait se faire sans trop de difficultés. Il serait intéressant de savoir si dans un tel cas, en adoptant la même méthode d'échantillonnage, la méthode proposée parviendra toujours à bien situer la mutation ciblée.

[14]Interférence : Phénomène caractérisé par une interaction entre des événements d'enjambement sur des segments distincts d'un chromosome.

[15]Marqueur : Gène ou séquence d'ADN, dont la position sur un chromosome est connue, associé à un trait phénotypique quelconque.

1.2. Objectifs

Le présent travail aura tout d'abord comme but de présenter une nouvelle méthode d'inférence en cartographie génétique qui permettra de situer un allèle mutant soumis à un effet de sélection situé entre deux ensembles de marqueurs et d'évaluer son efficacité et sa précision sur un échantillon donné quand celui-ci contient un nombre égal de séquences comportant le mutant sous sélection et son ancêtre primitif. La sélection sera tout d'abord assumée stabilisatrice[16]. Ensuite, nous travaillerons avec une population qui n'a pas encore atteint un état d'équilibre, i.e. la fréquence de l'allèle sous sélection changera avec le temps selon un principe qui sera défini à la section 3.2.3.

En fait, la méthode consiste en l'introduction d'une fonction de vraisemblance pour la position, calculée en unités du taux de recombinaison, du locus sous sélection qu'on tente de situer par rapport à deux ensembles de marqueurs donnés. L'hypothèse de non-interférence, expliquée à l'annexe B, sera posée. Ainsi, le résultat obtenu pourra facilement se traduire en distance physique sur la séquence. La méthode développée, basée sur une chaîne de Markov décrivant l'évolution rétrospective du matériel ancestral d'un échantillon de séquences, fait appel à la simulation de type Monte Carlo. Elle s'inspire beaucoup de celle développée par Larribe, Lessard et Schork [**15**]. Elle se démarque toutefois de cette dernière en permettant l'inclusion d'un phénomène de sélection au locus que nous tentons de situer. Cela se traduit concrètement par une subdivision de la population en deux familles caractérisées par l'allèle au locus sous sélection.

Nous chercherons également à illustrer que même dans cette situation, l'échantillonnage non-proportionnel de séquences dans la famille des allèles mutants et primitifs ne mène pas à une estimation biaisée de la position de la mutation. En d'autres mots, nous tenterons de déterminer si cette méthode d'échantillonnage est aussi souhaitable quand la sélection naturelle à un locus ciblé est incorporée au modèle, ce qui n'a pas été vérifié jusqu'à présent. À cette fin, des échantillons seront générés sous un ensemble de paramètres connus puis, en faisant des simulations prenant ces échantillons comme données, nous tenterons de vérifier si la méthode est en mesure de retrouver la valeur de la distance entre le site sous sélection et le marqueur choisi ayant été utilisée dans le cadre de la génération de l'échantillon. Les fréquences ancestrales dans

[16]Sélection stabilisatrice : Type de sélection naturelle maintenant constante la proportion d'un allèle dans une population. Cette proportion doit être supérieure à 0. En présence d'un tel effet de sélection, on dit qu'il y a «équilibre polymorphique» au locus correspondant.

la population de la mutation sous sélection seront aussi assumées connues.

Le corpus du présent mémoire commencera par un rappel succint de quelques modèles de base en génétique des populations (chapitre 2). Le modèle dont fait l'objet ce mémoire sera décrit en détails au chapitre 3. La technique utilisée pour générer l'échantillon requis (section 3.2) et le développement menant à la formulation de la nouvelle méthode de vraisemblance (section 3.3) seront décrits de façon exhaustive. Les résultats des simulations réalisées seront présentés et commentés par la suite (chapitre 4). Le présent travail se terminera par un exposé des différentes forces et faiblesses de la méthode ainsi que des possibles améliorations qui pourraient lui être apportées dans le cadre de projets subséquents (chapitre 5). Les concepts d'appariement et de recombinaison présentent aussi certaines caractéristiques dignes de mention, mais dont la compréhension n'est pas essentielle pour comprendre le modèle proposé. Ils seront donc abordés en annexe (annexes A, B)

1.3. Précisions sur la terminologie

Dans le présent travail, un «individu» sera défini, selon le contexte, comme un gène quelconque ou comme une séquence de gènes situés sur un chromosome à des loci pas nécessairement consécutifs. D'autre part, le terme «locus» sera parfois substitué par «site» : les deux mots sont synonymes. De plus, les expressions «allèle primitif», «allèle mutant», «allèle ancestral» et «allèle non-ancestral» seront couramment employées. L'allèle primitif est antérieur à l'allèle mutant, d'où son nom. Par allèle ancestral, on entend un allèle dont des descendants sont présents dans l'échantillon de séquences chromosomiques considéré.

Chapitre 2

MODÈLES DE BASE EN GÉNÉTIQUE DES POPULATIONS

2.1. Modèle de Wright-Fisher [24]

Le modèle de Wright-Fisher adopte une approche temporelle prospective. En d'autres mots, le temps 0 correspond à l'instant présent et le temps τ, $\tau > 0$, correspond à un moment postérieur. On considère tout d'abord une population de $2N$ individus haploïdes[1] dont un locus précis peut comporter un allèle de type A ou a. Sa taille reste constante dans le temps et on ignore les effets de la mutation, de la sélection et de la recombinaison. Puisqu'il n'y a pas de sélection, si les individus étaient diploïdes[2], l'agencement des allèles (i.e. AA, Aa ou aa) n'aurait aucun impact sur la descendance. Par conséquent, dans cette situation, le modèle haploïde est entièrement équivalent au modèle diploïde. En d'autres mots, le modèle développé pour $2N$ individus haploïdes sera identique à celui créé pour N individus diploïdes. Les générations sont discrètes et distinctes et l'accouplement se fait aléatoirement. Chaque individu meurt après avoir donné naissance.

À partir de ces quelques hypothèses, on peut calculer, entre autres, la probabilité de fixation d'un allèle dans une population, i.e. la probabilité qu'un allèle fasse entièrement disparaître l'autre éventuellement.

Considérons tout d'abord $Z_{2N}(\tau)$, dénotant le nombre d'individus de type A dans la population à la génération τ. Dans le modèle de Wright-Fisher, il est assumé que les gamètes sont choisis au hasard à chaque génération à partir d'une réserve infinie de gamètes reflétant les fréquences alléliques parentales. Si le temps est calculé en nombre de générations, nous obtiendrons

[1]Haploïde : Dont le code génétique est formé d'un ensemble de chromosomes uniques.

[2]Diploïde : Dont le code génétique est formé d'un ensemble de paires de chromosomes. Par exemple, l'être humain est diploïde. En effet, chaque cellule non sexuelle contient 23 paires de chromosomes.

un échantillonnage binomial avec :

$$P(Z_{2N}(\tau+1)=j|Z_{2N}(\tau)=i)=p_{ij}=\binom{2N}{j}\pi_i^j(1-\pi_i)^{2N-j}, \tag{1}$$

où $\pi_i=\frac{i}{2N}$.

On constate que ce processus est en fait une chaîne de Markov homogène dans le temps. En plus, $\frac{Z_{2N}(\tau)}{2N}$ est une martingale. En effet, la nature markovienne de $Z_{2N}(\tau)$ nous permet d'affirmer que

$$E\left[\frac{Z_{2N}(\tau+1)}{2N}\middle|Z_{2N}(0),Z_{2N}(1),...,Z_{2N}(\tau)\right]=E\left[\frac{Z_{2N}(\tau+1)}{2N}\middle|Z_{2N}(\tau)\right]=\frac{Z_{2N}(\tau)}{2N}. \tag{2}$$

Ce processus a deux états absorbants : 0 et $2N$. Le théorème d'arrêt optionel de Doob [**11**] nous permet de déduire que

$$P(\{\text{Absorption à } i\})=\begin{cases}\frac{E[Z_{2N}(0)]}{2N}, \text{ si } i=2N,\\ 1-\frac{E[Z_{2N}(0)]}{2N}, \text{ si } i=0.\end{cases} \tag{3}$$

Ce résultat est une conséquence directe du théorème suivant.

Theorème 1. *Si $X(\tau)$, $\tau\geq 0$, est une martingale bornée et T est un temps d'arrêt, alors $E[X(T)]=X(0)$.*

Définissons $X_{2N}(\tau)=\frac{Z_{2N}(\tau)}{2N}$. Dans le cas qui nous occupe, le temps d'arrêt T est défini comme $T=\min\{\tau : X_{2N}(\tau)=0 \text{ ou } 1\}$. Ainsi, si p est défini comme la probabilité de fixation de l'allèle A, alors on a

$$E[X_{2N}(T)]=E[X_{2N}(0)],$$

d'où

$$p=E[X_{2N}(0)]. \tag{4}$$

Le raisonnement précédent est basé sur le fait que $X_{2N}(\tau)$, une mesure de fréquence, ne peut prendre que deux valeurs au temps d'arrêt T : 0 si l'allèle A disparaît et 1 s'il élimine l'allèle compétiteur. Pour le modèle de Wright-Fisher, $E[X_{2N}(0)]$ correspond simplement à la fréquence de l'allèle A au temps 0.

2.1.1. Approximation du modèle de Wright-Fisher par un processus de diffusion[3]

Quand $2N$ devient très grand, le processus de propension allélique peut être modifié pour qu'il approche un processus de diffusion. Or, pour obtenir une limite non-dégénérée, il est nécessaire de rééchelonner le temps en unités de $2N$ générations ($\tau = \lfloor 2Nt \rfloor$). Définissons

$$Y_{2N}(t) = \frac{Z_{2N}(\lfloor 2Nt \rfloor)}{2N}, \quad t \geq 0. \tag{5}$$

Un processus de diffusion en une dimension est caractérisé uniquement par sa moyenne infinitésimale, $\mu(y)$, et sa variance infinitésimale, $\sigma^2(y)$. Quand la chaîne de Markov est homogène dans le temps, ces deux quantités prennent les valeurs

$$\mu(y) = \lim_{h \to 0} \frac{E[Y(t+h) - Y(t) | Y(t) = y]}{h}, \tag{6}$$

$$\sigma^2(y) = \lim_{h \to 0} \frac{E[(Y(t+h) - Y(t))^2 | Y(t) = y]}{h}. \tag{7}$$

Sachant que $Z_{2N}(\tau + 1)$ étant donné $Z_{2N}(\tau) = i$ est de loi binomiale avec $n = 2N$ et $\theta = i/2N$, on peut trouver les valeurs suivantes :

$$E\left[\frac{Z_{2N}(\tau+1)}{2N} - \frac{Z_{2N}(\tau)}{2N} \middle| \frac{Z^{(N)}(\tau)}{2N} = \frac{i}{2N}\right] = 0, \tag{8}$$

$$E\left[\left(\frac{Z_{2N}(\tau+1)}{2N} - \frac{Z_{2N}(\tau)}{2N}\right)^2 \middle| \frac{Z_{2N}(\tau)}{2N} = \frac{i}{2N}\right] = \frac{1}{2N}\frac{i}{2N}(1 - \frac{i}{2N}). \tag{9}$$

En considérant $\tau = \lfloor 2Nt \rfloor$ et $h = 1/2N$, on obtient que

$$\mu(y) = \lim_{h \to 0} \frac{E[Y_{2N}(t+h) - Y_{2N}(t) | Y_{2N}(t) = y]}{h} = 0 \tag{10}$$

$$\sigma^2(y) = \lim_{h \to 0} \frac{E[(Y_{2N}(t+h) - Y_{2N}(t))^2 | Y_{2N}(t) = y]}{h} = y(1-y), \ 0 < y < 1. \tag{11}$$

La probabilité de fixation d'un allèle dans la population restera égale à sa fréquence initiale, ce qui est logique, puisqu'il s'agit encore d'une approximation du modèle de Wright-Fisher.

2.1.2. Le modèle de Wright-Fisher vu de façon rétrospective

Bien que le modèle de Wright-Fisher ait été d'abord introduit pour décrire le développement d'une population vers le futur, il est pratique d'adopter une approche rétrospective. En effet, un modèle prospectif peut difficilement être utilisé pour expliquer comment s'est constitué

[3]Les formules reproduites ici ont été prises dans Karlin et Taylor **[11]**.

une population. Le manque d'information par rapport à la composition passée de la population est bien souvent un frein à son applicabilité. D'un autre côté, la méthode requiert qu'on génère l'évolution d'une population en entier. En pratique, les populations peuvent être de taille très grande et l'utilisation de ce modèle devient donc synonyme d'une énorme charge de calculs. Il est donc avisé de faire appel à un modèle rétrospectif. Les hypothèses de bases resteront les mêmes.

Tous les individus de la population à une génération donnée sont maintenant numérotés de 1 à $2N$ et le nombre de rejetons de l'individu i à la génération suivante est dénoté ν_i, $0 \leq i \leq 2N$. Le nombre total de rejetons doit être de $2N$. Sous les hypothèses du modèle de Wright-Fisher, le vecteur $(\nu_1, \nu_2, ..., \nu_{2N})$ a une distribution multinomiale décrite par

$$P(\nu_1 = m_1, \nu_2 = m_2, ..., \nu_{2N} = m_{2N}) = \frac{(2N)!}{m_1! m_2! ... m_{2N}!} \left(\frac{1}{2N}\right)^{2N}, \tag{12}$$

sous la condition que $\sum_{i=1}^{2N} m_i = 2N$, avec $0 \leq m_1, m_2, ..., m_{2N} \leq 2N$.

Ce modèle est bien celui de Wright-Fisher. Pour en être convaincu, il suffit de se rappeler tout d'abord qu'en l'absence de mutation, les individus produisent des rejetons qui leur sont identiques. Ainsi, s'il y a i individus de type A numérotés de 1 à i, on peut affirmer que le nombre de descendants de ces individus (qui seront tous de type A) sera égal à $\sum_{j=1}^{i} \nu_j$. Dénotons le résultat de cette somme par $Z_{2N}(\tau)$. Sachant que $(\nu_1, \nu_2, ..., \nu_{2N})$ a une distribution multinomiale, nous pouvons affirmer que $Z_{2N}(\tau)$ aura bel et bien une distribution binomiale avec paramètres $n = 2N$ et $p_i = \frac{i}{2N}$, ce qui correspond aux paramètres trouvés précédemment. Maintenant, si l'on voit ce processus de façon rétrospective, il est crucial de remarquer que d'une part, les individus choisiront leur parent indépendamment et entièrement au hasard dans la génération précédente et que les choix de parents seront indépendants d'une génération à l'autre.

Nous sommes maintenant en mesure de générer une généalogie pour une population donnée. On peut voir un exemple à la figure (2). Dans cette optique, il peut être intéressant de s'attarder au temps qu'il faudra pour trouver un ancêtre commun à deux individus. Dénotons ce temps, exprimé en nombre de générations, par T_2'. On a

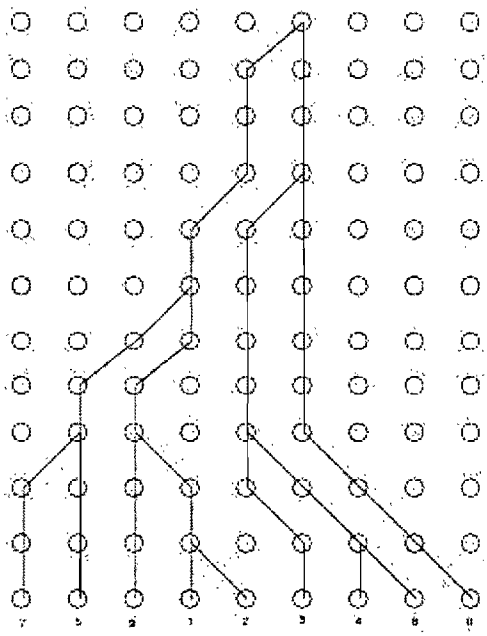

FIG. 2. Arbre représentant une généalogie possible générée à partir d'un modèle de Wright-Fisher avec $N = 9$ (Tiré des notes de Simon Tavaré [**24**])

$$P(T_2' > 1) = 1 - \frac{1}{2N} \tag{13}$$

et, par conséquent,

$$P(T_2' > \tau) = (1 - \frac{1}{2N})^{\tau}, \quad \tau \in \mathbb{N}, \ \tau \geq 0. \tag{14}$$

Si, en revanche, le temps t est échelonné en unités de $2N$ générations, on obtient

$$P(T_2 > t) = P(T_2' > \lfloor 2Nt \rfloor) = (1 - \frac{1}{2N})^{\lfloor 2Nt \rfloor}, \quad t \geq 0. \tag{15}$$

Maintenant, si on laisse $N \to \infty$, on obtient que

$$\lim_{N \to \infty} P(T_2 > t) = e^{-t}, \ t \geq 0.$$

On en déduit donc que T_2 suit asymptotiquement une loi exponentielle de paramètre $\lambda = 1$.

Quand deux séquences ou plus trouvent un parent en commun, on dit qu'il se produit un événement de *coalescence*. Si l'on définit T_i comme le temps entre le premier moment où un échantillon donné de la population comporte i ancêtres et le prochain événement de coalescence à l'intérieur de cet échantillon, on réalise que la nature markovienne du modèle de Wright-Fisher a comme effet de rendre les T_i $(i = 2, ..., n)$, où n correspond au nombre initial de séquences

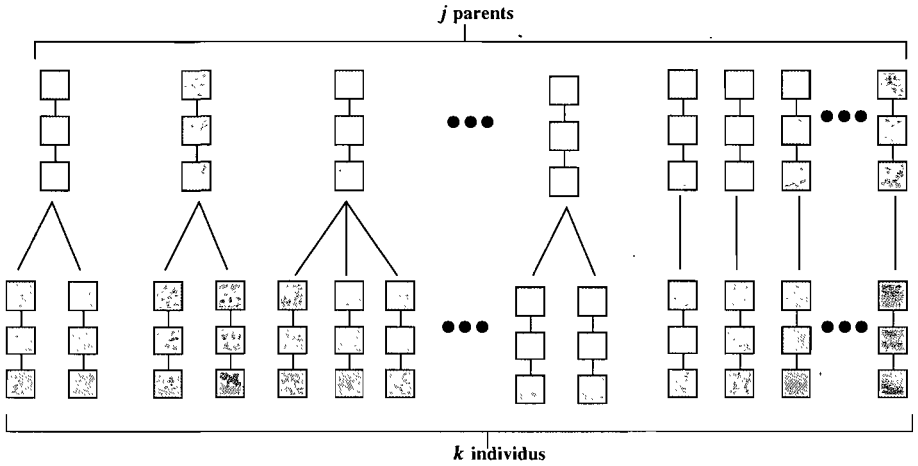

FIG. 3. Plusieurs événements de coalescence simultanés.

ancestrales à cet échantillon, indépendants.

En fait, il peut être intéressant de vérifier que la probabilité que k séquences trouvent j parents, $j < k-1$, devient négligeable quand $N \to \infty$. Un tel événement se produira s'il y a plusieurs événements de coalescence simultanés. On peut voir cette situation représentée à la figure (3).

Le problème en est un de combinatoire : il faut obtenir la probabilité que k objets soient répartis dans j contenants non-vides, $j \leq k$, s'il y a en tout k contenants. Empruntons la notation de Tavaré [**24**] :

$$g_{kj} = P(k \text{ individus aient } j \text{ parents distincts}) = 2N(2N-1)...(2N-j+1)S_k^{(j)}(2N)^{-k}, \quad (16)$$

où $S_k^{(j)}$ est un nombre de Stirling du deuxième type. Le nombre de Stirling de deuxième type $S_k^{(j)}$ indique le nombre de façons de partitionner un ensemble de k éléments en j sous-ensembles non-vides. Il correspond à

$$S_k^{(j)} = \frac{1}{j!}\sum_{i=1}^{j}(-1)^{j-i}\binom{j}{i}i^k. \quad (17)$$

Notons que $2N(2N-1)...(2N-j+1)$ correspond au nombre de façons de choisir j parents distincts, tandis que $S_k^{(j)}$ est égal au nombre de façons d'attribuer les k enfants aux j parents. Enfin $(2N)^k$ correspond au nombre total de façons de choisir les parents.

Un autre processus peut être défini à partir de l'information fournie par le modèle de Wright-Fisher. Dénotons-le par $A_n(t)$ et appelons-le *processus ancestral*. On l'appelle ainsi parce qu'il représente, t unités de temps en arrière, le nombre d'ancêtres communs à un échantillon de n individus. À la base, les éléments de sa matrice de transition sont donnés par (16).

Notons que

$$g_{kj} = \begin{cases} \binom{k}{2}\frac{1}{2N} + O(N^{-2}), & \text{si } j = k-1, \\ O(N^{-2}), & \text{si } j < k-1, \\ 1 - \binom{k}{2}\frac{1}{2N} + O(N^{-2}), & \text{si } j = k, \end{cases} \tag{18}$$

puisque $S_k^{(k-1)} = \binom{k}{2}$ pour $1 \leq j \leq k \leq n$.

Si l'on désigne par G_N la matrice de transition formée des éléments g_{kj} pour $1 \leq j \leq k \leq n$, on obtient que

$$G_N = I + \frac{Q}{2N} + O(N^{-2}), \tag{19}$$

I étant la matrice identité et $Q = \|q_{kj}\|_{k,j=1}^n$ étant une matrice dont toutes les entrées sont 0 à l'exception de

$$q_{kk} = -\binom{k}{2} \tag{20}$$

et

$$q_{k,k-1} = \binom{k}{2}, \tag{21}$$

pour $k = 2, 3, ..., n-1, n$.

En rééchelonnant de nouveau le temps en unités de $2N$ générations, on arrive à

$$G_{2N}^{\lfloor 2Nt \rfloor} = (I + \frac{Q}{2N} + O(N^{-2}))^{\lfloor 2Nt \rfloor} \to e^{Qt}, \tag{22}$$

lorsque $N \to \infty$, pour la matrice de transition de l'instant 0 à l'instant t dans le passé.

On réalise donc que le processus ancestral rééchelonné limite est un processus de mort, une chaîne de Markov à temps continu, dont le point de départ est $A_n(0) = n$, qui fait des chutes de 1 uniquement. En plus, le temps d'attente entre chaque chute est distribué de façon exponentielle avec moyenne $\frac{2}{k(k-1)}$, où k est le nombre de séquences ancestrales à l'échantillon

juste avant que la chute n'ait lieu. Les temps d'attente entre les chutes sont tous indépendants.

Ainsi, nous réalisons que sous les contraintes du modèle de Wright-Fisher, quand $N \to \infty$ et quand le temps est rééchelonné en unités de $2N$ générations, la probabilité que k séquences aient j parents, $j < k-1$, est de 0. Ce développement mène à la formulation du modèle coalescent de Kingman.

2.2. Le modèle coalescent de Kingman

Kingman [**12**] a défini un modèle fondamental en génétique des populations, le «n-coalescent», désormais dénoté par $C_n(t)$, $t \in \mathbb{R}^+$. Plutôt que de s'intéresser directement au nombre de lignées ancestrales, il s'attarde à un processus dont les états correspondent à des partitions en «classes d'équivalence». Chacune des classes de la partition à l'instant t contient un ensemble non ordonné d'individus, tirés d'un échantillon en comportant n, partageant le même ancêtre commun à l'instant t dans le passé. Dénotons la classe i par E_i. Ainsi, $C_n(0) = \{\{1\},\{2\},...,\{n\}\}$ (aucune classe n'a encore trouvé d'ancêtre commun et donc, chacune des classes ne contient qu'un individu) et $\lim_{t\to\infty} C_n(t) = \{\{1,2,...,n\}\}$ (toutes les séquences ont trouvé un même ancêtre commun : il ne reste par conséquent plus qu'une classe contenant tous les individus de l'échantillon).

La composition de la population évolue comme à la section 2.1.2 et le temps est exprimé en unités de $2N$ générations en laissant N tendre vers l'infini. Quand deux ancêtres de l'échantillon trouvent un ancêtre commun, les classes auxquelles ils appartiennent coalescent. Ainsi, le taux de coalescence de deux classes paritculières sera de 1. On réalise aussi que le processus $C_n(t)$ est markovien avec taux de la partition α à la partition β donné par

$$q_{\alpha\beta} = \begin{cases} -\binom{k}{2}, & \text{si } \alpha = \beta, |\alpha| = k, \\ 1, & \text{si } \alpha \prec \beta, \\ 0, & \text{autrement,} \end{cases} \tag{23}$$

où $|\alpha|$ dénote le nombre de classes de la partition α et $\alpha \prec \beta$ signifie que β peut être obtenu à partir de α en fusionnant deux classes de α. Notons aussi que $|C_n(t)| = A_n(t)$, le processus ancestral défini précédemment. On peut représenter aisément $C_n(t)$ par un arbre binaire avec racines. En effet, $C_n(t)$ nous donne toute l'information nécessaire pour le faire : la longueur des branches, l'identité des individus impliqués dans les événements de coalescence et l'ordre

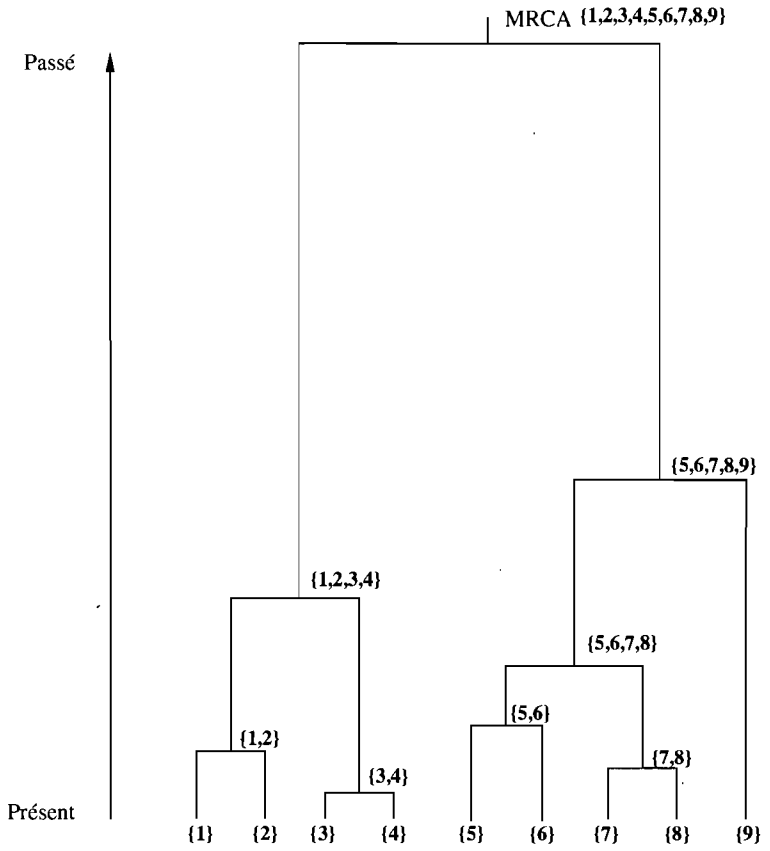

FIG. 4. Réalisation du processus de coalescence sans mutation avec $n = 9$.

de ces événements. L'illustration (4) représente une arborescence du processus coalescent bâtie à partir de neuf séquences.

Le modèle présenté ici peut être complexifié en y incluant le phénomène de mutation. Considérons un phénomène de mutation neutre et récurrente avec un taux de μ par génération. Le taux par unité de temps de $2N$ générations sera de $\theta/2$, où $\theta = 4N\mu$. Afin de ne pas obtenir divergence quand $N \to \infty$, la valeur de θ sera assumée constante quand $N \to \infty$. En d'autres mots, μ est inversement proportionnel à N ou

$$\lim_{N\to\infty} 2N\mu = \frac{\theta}{2}. \tag{24}$$

La figure (5) nous donne un exemple de mutation dans un arbre de coalescence.

Un des avantages du modèle coalescent est l'indépendance existant entre le processus stochastique à l'origine de la configuration arborescente et celui à l'origine de la mutation. Cette

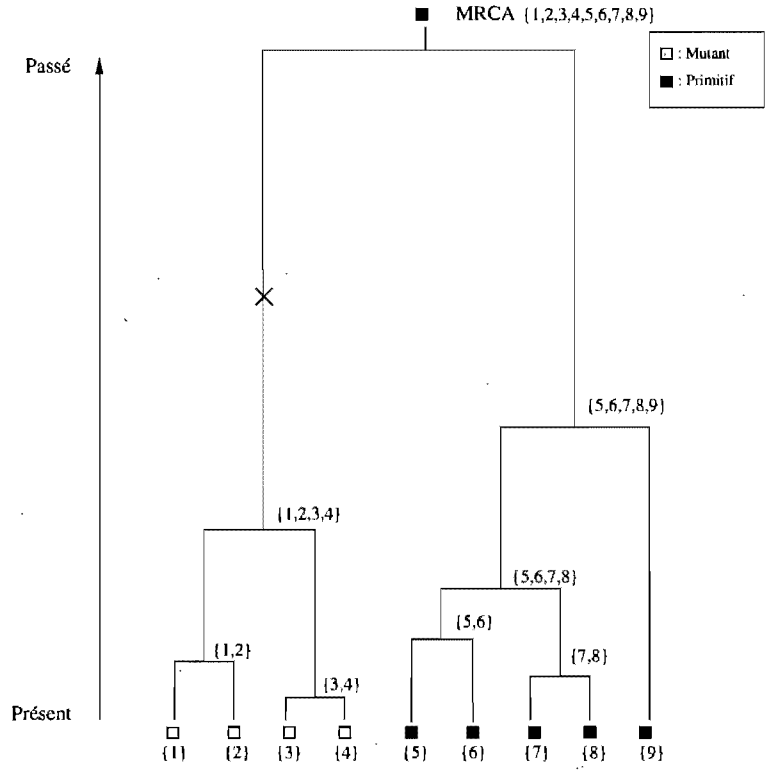

FIG. 5. Réalisation du processus de coalescence avec mutation avec $n = 9$. Le X représente une mutation.

indépendance nous permet de générer les deux processus séparément. En fait, elle peut être comprise intuitivement si on se rappelle que le modèle coalescent décrit le modèle de Wright-Fisher vu de façon rétrospective quand $N \to \infty$ avec $2N$ générations comme unité de temps. Considérons les liens généalogiques résultant de la reproduction dans ce contexte. Si l'on reprend un point de vue prospectif, les lignées se ramifient quand un individu produit plus d'un descendant et disparaissent quand il n'y a pas production de descendants. Ajouter une mutation neutre ne vient en rien affecter le phénomène de reproduction (les mutants continueront à produire des descendants au même taux). Si cela est vrai dans le cadre du modèle vu de façon prospective, cela doit également être vrai dans le modèle vu de façon rétrospective. Cette constation sera donc aussi applicable pour le modèle coalescent [**18**]. Par conséquent, on peut tout simplement superposer le processus de mutation à l'arbre de coalescence.

Le processus de mutation est aussi un processus de Poisson. On peut le comprendre en assumant tout d'abord un taux de mutation de μ par génération. Adoptons un point de vue rétrospectif et désignons par K' le temps nécessaire, exprimé en nombre de générations, pour qu'un événement de mutation se produise le long d'une lignée. Ainsi,

$$P(K' > k') = (1-\mu)^{k'}. \tag{25}$$

Si μ est encore soumis aux conditions énoncées précédemment, l'équation précédente prend la forme :

$$P(K' > k') = \left(1 - \frac{\theta}{4N}\right)^{k'}. \tag{26}$$

Maintenant, en rééchelonnant le temps en unités de $2N$ générations et en laissant $N \to \infty$, si l'on désigne par K le temps sous la nouvelle échelle ($k' = \lfloor 2Nk \rfloor$), on obtient

$$P(K > k) = \lim_{N\to\infty} \left(1 - \frac{\theta}{4N}\right)^{\lfloor 2Nk \rfloor} = e^{-\frac{\theta}{2}k}. \tag{27}$$

Tel que mentionné précédemment, nous sommes en présence d'un processus de Poisson avec taux $\theta/2$. Puisque chaque lignée mute indépendamment, si le nombre de lignées ancestrales de l'échantillon est a, alors le taux de mutation pour l'ensemble de ces lignée sera de $a\theta/2$.

De plus, si une mutation unique (i.e. non-récurrente) a rendu un locus donné polymorphe, alors la position de cette mutation sur l'arbre de coalescence jusqu'au MRCA sera distribuée uniformément, conditionnellement à la longueur de toutes les branches. Numérotons les branches de l'arbre de coalescence de 1 à $\frac{n(n+1)}{2} - 1$ et dénotons la longueur de la branche i par L_i. Soient la variable $L = \sum_{i=1}^{\frac{n(n+1)}{2}-1} L_i$, la longueur totale des branches de l'arbre avec $2N$ générations comme unité de longueur et la variable $N(l)$, désignant le nombre de mutations sur l'intervalle $[0, l)$. En se rappelant que la mutation apparaît sur chaque branche de l'arbre selon un processus de Poisson d'intensité $\frac{\theta}{2}$, on peut en déduire que, si B dénote la position de la mutation, alors

$$\begin{aligned} P(B < b | N(L) = 1, L = l) &= \frac{(\frac{\theta}{2}e^{-\frac{\theta}{2}b})e^{-\frac{\theta}{2}(l-b)}}{\frac{\theta}{2}le^{-\frac{\theta}{2}l}} \\ &= \frac{b}{l}, 0 \leq b \leq l, \end{aligned} \tag{28}$$

ce qui démontre que sa position est bel et bien distribuée selon une distribution uniforme sur l'intervalle [0, l].

2.3. Le graphe de recombinaison ancestral (ARG) [24]

La recombinaison[4] vient passablement compliquer les choses. Puisqu'elle a pour effet de séparer des loci autrement liés, la généalogie d'une séquence génétique n'est plus équivalente à la généalogie à un locus. Par conséquent, il est maintenant essentiel de différencier les arbres dits «marginaux», qui retracent les généalogies aux différents loci, et le graphe de recombinaison ancestral (ARG), qui contient aussi l'information sur l'appariement des allèles à ces loci [5]. Une représentation graphique de l'ARG est donnée par la figure (6). Les arbres marginaux seront dépendants et chacun aura son MRCA, qui ne sera pas obligatoirement le même. Voilà pourquoi on n'emploie pas le terme MRCA pour désigner le premier point dans le ARG où $A_n(t)$, dénotant le nombre d'individus ancestraux à l'échantillon au temps t, prend la valeur 1. On appelle plutôt cet individu «ancêtre ultime» (UA ou Ultimate Ancestor). À ce point, tous les arbres marginaux auront trouvé leur MRCA. Évidemment, si on désigne par τ_n le temps auquel apparaît ce UA à partir d'un échantillon de n individus et par M_i le temps d'apparition du MRCA au locus i, on déduit que

$$\tau_n \geq \max\{M_i : \ i = 1, 2, ...b\}, \tag{29}$$

où b désigne le nombre de loci dans la séquence considérée.

La recombinaison, en permettant de nouvelles combinaisons alléliques, provoquera invariablement un accroissement du polymorphisme dans la population. Dans le cadre du modèle coalescent, seuls des événements de coalescence venaient affecter la topologie de l'arbre généalogique des individus de l'échantillon. L'inclusion du phénomène de recombinaison vient y ajouter des événements de branchement.

Quand il n'y a pas de recombinaison, le processus ancestral, $A_n(t)$, atteint un état absorbant à 1. En effet, puisque la valeur de $A_n(t)$ ne peut pas augmenter et puisque cette valeur ne peut diminuer que s'il y a un événement de coalescence, lorsque $A_n(t)$ prend la valeur 1, celle-ci ne pourra plus varier. Après tout, il faut au moins deux individus pour qu'il puisse y avoir coalescence. Or, le processus ancestral pour l'ARG ne comporte aucun état absorbant. En effet, quand le processus ancestral atteint la valeur 1, invariablement, un événement de recombinaison

[4]Pour une introduction au phénomène de recombinaison, voir l'annexe B.

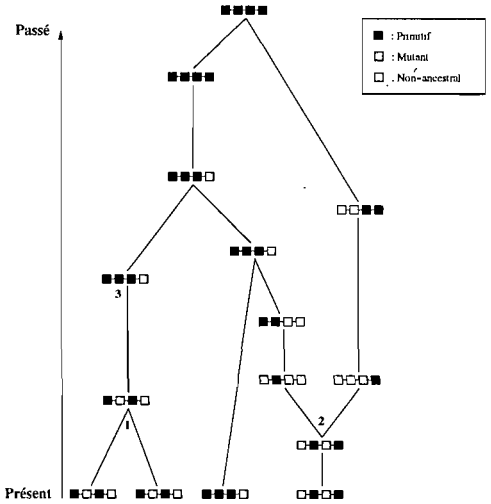

FIG. 6. Une réalisation du graphe de recombinaison ancestral (ARG). "1" désigne un événement de coalescence, "2", un événement de recombinaison entre les loci numéro 2 et numéro 3 et "3", un événement de mutation. Le graphe doit être lu de façon rétrospective, c'est-à-dire de bas en haut.

viendra la ramener à 2.

On suppose que les événements de recombinaison se produisent également selon un processus de Poisson, indépendamment des événements de coalescence et de mutation. Le taux de recombinaison r par génération pour chaque séquence aussi est assumé inversement proportionnel à $2N$, de façon à ce que $\lim_{N\to\infty} 2Nr = \rho/2$. Or, la recombinaison aura pour effet d'introduire du matériel dit «non-ancestral» dans les séquences étudiées. On le dit non-ancestral parce que sa descendance n'est pas représentée dans l'échantillon prélevé. Puisqu'un événement de recombinaison se produisant entre deux loci non-ancestraux n'aura aucun impact sur le graphe en question, le taux de recombinaison pour un échantillon devra être multiplié par un facteur β, la probabilité qu'un événement de recombinaison affecte le matériel ancestral. Dans le présent travail, puisqu'il n'y a pas d'interférence, ce facteur prend comme valeur

$$\beta = \sum_{i=1}^{s} \frac{n_i}{n} \sum_{j=1}^{b-1} I\{\text{Intervalle } j \text{ de la séq. de type } i \text{ est compris entre deux loci ancestraux}\} \frac{\rho_j}{\rho}, \tag{30}$$

où s correspond au nombre total de types de séquences, $\rho_j/2$ au taux de recombinaison rééchelonné entre les deux loci bordant l'intervalle j; $\rho/2$ au taux de recombinaison rééchelonné total sur la séquence et b correspond au nombre de loci dans la séquence. L'intervalle j correspond à l'espace séparant les loci j et $j+1$ d'une séquence donnée. Les n_i de l'équation précédente

correspondent au nombre de séquences ancestrales[5] de type i. Le type d'une séquence est tout simplement défini par les allèles qu'il comporte aux loci observés. Par exemple, l'échantillon utilisé pour créer l'ARG représenté par la figure (6) en comporte initialement 3. Enfin, mentionnons que $n = \sum_{i=1}^{s} n_i$. Notons aussi que toutes les quantités impliquées dans le calcul de β sont évaluées à un même instant t.

La formule donnant la valeur de β peut être comprise intuitivement. Si l'on assume que la position d'un point de césure (chiasme) est choisie uniformément sur une des séquences ancestrales, on peut déduire que la probabilité que cette césure apparaisse entre deux loci ancestraux sera donnée par

$$\frac{\text{Longueur cumulée des segments de séquences compris entre deux loci ancestraux}}{\text{Longueur totale cumulée des séquences ancestrales}}. \tag{31}$$

Ainsi, nous pouvons déduire que :

$$\text{Longueur totale cumulée des séquences ancestrales} = n\frac{\rho}{2}, \tag{32}$$

ce qui se comprend aisément, puisque $\frac{\rho}{2}$ est assumé comme étant la longueur de l'intervalle sur chaque séquence à l'intérieur de laquelle une césure peut apparaître[6] et n, le nombre total de séquences dans l'échantillon. Nous déduisons aussi que la longueur cumulée des segments de séquences compris entre deux loci ancestraux est donnée par

$$\sum_{i=1}^{s} n_i \sum_{j=1}^{b-1} I\{\text{Intervalle } j \text{ de la séquence de type } i \text{ est compris entre deux loci ancestraux}\}\frac{\rho_j}{2}, \tag{33}$$

$\frac{\rho_j}{2}$ correspondant à la distance entre les deux loci bordant l'intervalle j, et b, au nombre total de loci dans la séquence. Ainsi, le taux de recombinaison effectif total pour un échantillon constitué de n séquences ancestrales sera de $n\beta\frac{\rho}{2}$.

Comme pour le modèle coalescent, le processus de mutation peut encore être généré indépendamment de la topologie du graphe. Cependant, la présence de loci non-ancestraux aura

[5]Séquence ancestrale : Séquence dont le matériel génétique, par descendance, est représenté en partie ou en entier dans l'échantillon prélevé au temps 0.

[6]Il est important de se rappeler que la distance est toujours exprimée en fonction du taux de recombinaison.

pour effet de rendre possible que certaines mutations placées sur le graphe n'aient aucun impact sur l'échantillon au temps initial. En effet, dans le ARG, une branche donnée n'illustre pas nécessairement le parcours ancestral de tous les allèles d'une séquence, seulement celui des allèles qui lui sont ancestraux. C'est pourquoi quand il y a n séquences dans un échantillon, le taux effectif total de mutation n'est pas de $n\theta/2$, avec

$$\frac{\theta}{2} = \sum_{i=1}^{b} \frac{\theta_i}{2}, \tag{34}$$

$\theta_i/2$ correspondant au taux de mutation au locus i, mais bien de $n\alpha\theta/2$, où

$$\alpha = \sum_{i=1}^{s} \frac{n_i}{n} \sum_{j=1}^{b} I\{\text{Locus } j \text{ est ancestral}\} \frac{\theta_j}{\theta}, \tag{35}$$

avec n_i, n, s et b définis comme dans (30).

La quantité α doit être comprise comme la probabilité qu'une mutation affecte le matériel ancestral de l'échantillon. Notons enfin que θ_j est soumis aux mêmes conditions asymptotiques que le paramètre de mutation pour le processus de coalescence de base.

Il est important de réaliser que même si la recombinaison a pour effet de changer l'agencement des allèles des séquences d'un échantillon, elle n'a pas d'impact sur la longueur des arbres marginaux à chaque locus. En effet, en l'absence de sélection, ils seront toujours décrits adéquatement par le modèle coalescent de Kingman. Cela s'explique par le fait que l'introduction de matériel non-ancestral à l'échantillon étudié n'aura pas de répercussion sur le taux de coalescence à un locus donné, tout simplement parce que celui-ci est au fond approximativement égal au nombre de paires non ordonnées de séquences ancestrales au locus en question au temps indiqué.

Afin de décrire la dynamique du graphe de recombinaison ancestrale **[24]**, considérons maintenant une paire de loci, dénotés 1 et 2, et désignons les arbres marginaux qui leur sont associés par M_1 et M_2. Dénotons l'ensemble des vertices d'un graphe donné par $\xi(.)$ et l'ARG par le symbole G. Ainsi, $\xi(M_i)$ correspond à l'ensemble des vertices de l'arbre de coalescence à un locus i. Remarquons que $\xi(M_1)$ et $\xi(M_2)$ sont des sous-ensembles de $\xi(G)$ et que $\xi(M_i)'$ correspond aux vertices de G qui ne sont pas ancestrales au locus i. On peut partitionner $\xi(G)$ en quatre catégories distinctes :

- $A = \xi(M_1) \cap \xi(M_2)'$

- $B = \xi(M_1)' \cap \xi(M_2)$
- $C = \xi(M_1) \cap \xi(M_2)$
- $D = \xi(G) \cap \xi(M_1)' \cap \xi(M_2)'$

Dénotons le nombre de vertices appartenant au groupe I, $I = A, B, C, D$, au temps t par $n_I(t) = |\xi(G_t) \cap I|$, $\xi(G_t)$ représentant les vertices existant au temps t. On peut immédiatement déduire que

$$\begin{aligned} n_A(t) + n_B(t) + n_C(t) + n_D(t) &= |\xi(G_t)| = A_n(t), & (36) \\ n_A(t) + n_C(t) &= |\xi(T_1(t))| = A_n^{(1)}(t), & (37) \\ n_B(t) + n_C(t) &= |\xi(T_2(t))| = A_n^{(2)}(t), & (38) \end{aligned}$$

où $A_n(t)$ dénote le processus ancestral pour l'ARG, $A_n^{(1)}(t)$, le processus ancestral pour l'arbre marginal au locus 1 et $A_n^{(2)}(t)$, le processus ancestral pour l'arbre marginal au locus 2.

On définit le processus $\mathbf{m}(t)$ comme

$$\mathbf{m}(t) = (n_A(t), n_B(t), n_C(t), n_D(t)). \quad (39)$$

Ce processus aura les taux de transition suivant (voir l'illustration (7) pour savoir précisément quel type d'événement chaque transition désigne) :

$$\text{Taux } (i = (a,b,c,d) \to j) = \begin{cases} c\frac{\rho}{2} & \text{, si (1) } j = (a+1, b+1, c-1, d), \\ ab & \text{, si (5) } j = (a-1, b-1, c+1, d), \\ ac + \frac{a(a-1)}{2} & \text{, si (6) } j = (a-1, b, c, d), \\ bc + \frac{b(b-1)}{2} & \text{, si (4) } j = (a, b-1, c, d), \\ \frac{c(c-1)}{2} & \text{, si (7) } j = (a, b, c-1, d), \\ d(a+b+c) + \frac{d(d-1)}{2} & \text{, si (3) } j = (a, b, c, d-1), \\ (a+b+d)\frac{\rho}{2} & \text{, si (2) } j = (a, b, c, d+1). \end{cases} \quad (40)$$

Puisque nous n'avons aucun intérêt à garder en mémoire le nombre de séquences ne comportant aucun gène ancestral, il est plus pertinent de comptabiliser les valeurs du processus $\mathbf{n}(t)$:

$$\mathbf{n}(t) = (n_A(t), n_B(t), n_C(t)) \quad (41)$$

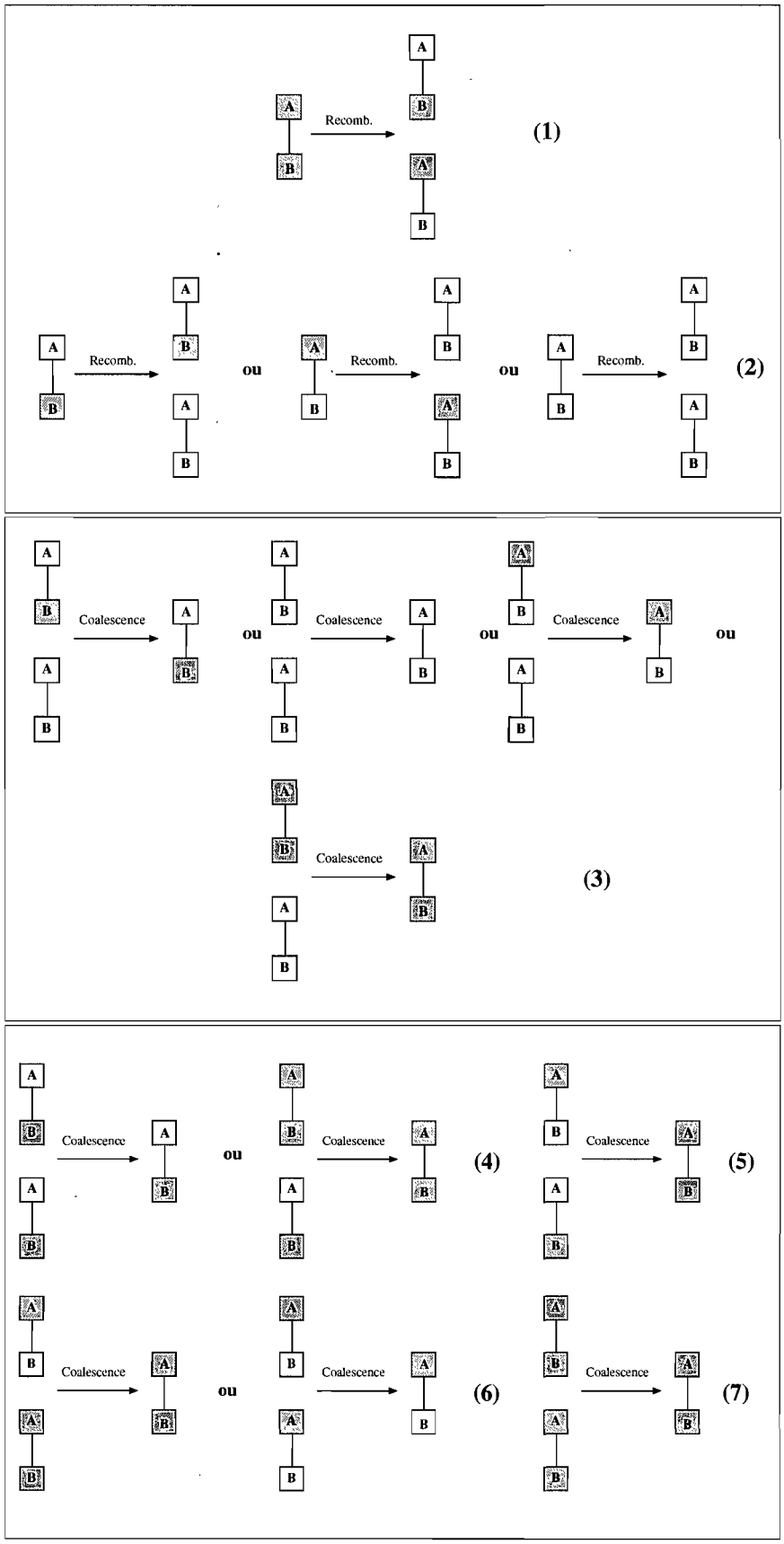

FIG. 7. Représentation de différents événements pouvant affecter l'échantillon. Le numéro identifiant chaque événement peut être associé à la liste de taux précédente.

avec les taux de transition

$$\text{Taux } (i=(a,b,c)\to j)=\begin{cases} c\frac{\rho}{2} & \text{, si } j=(a+1,b+1,c-1),\\ ab & \text{, si } j=(a-1,b-1,c+1),\\ ac+\frac{a(a-1)}{2} & \text{, si } j=(a-1,b,c),\\ bc+\frac{b(b-1)}{2} & \text{, si } j=(a,b-1,c),\\ \frac{c(c-1)}{2} & \text{, si } j=(a,b,c-1). \end{cases} \tag{42}$$

Maintenant, nous ne prenons en compte que les événements de recombinaison se produisant entre les loci 1 et 2 quand les gènes à ces deux loci sont ancestraux. Puisque les valeurs de $n_A(t)+n_C(t)$ et $n_B(t)+n_C(t)$ ne peuvent que descendre avec le temps, nous pouvons conclure qu'éventuellement, ces deux sommes prendront la valeur 1. Cela signifie que le processus $\mathbf{n}(t)$ atteindra un état absorbant. Cela n'est pas surprenant. Après tout, les événements de coalescence se produisent à un taux quadratique, tandis que les événements de recombinaison se produisent à un taux linéaire. Ceci a pour effet de faire diminuer le nombre de vertices avec le temps. Ainsi, tous les gènes à un même locus trouveront éventuellement un ancêtre commun.

2.4. Remarques additionnelles sur les modèles de généalogie

Bien que trouver un moyen de décrire de façon rigoureuse une généalogie puisse être une fin en soi, les modèles développés pour ce faire ont un autre objectif. En effet, en raison d'un manque d'information, il nous est impossible de reconstruire précisément la généalogie d'une population, qui pourtant est unique. Nous espérons donc que grâce à nos méthodes, nous puissions nous en approcher. Nous pouvons affirmer par conséquent que les problèmes d'inférence découlant du polymorphisme observé dans une population sont des problèmes de données manquantes [**18**].

Ainsi, nous serons plutôt intéressés à faire de l'inférence sur la valeur d'un paramètre quelconque ayant façonné la population jusqu'à l'instant présent. Nous espérons que la généalogie bâtie pourra nous fournir de l'information par rapport à celui-ci. En fait, on peut affirmer que formellement, la simulation de généalogies possibles constitue un moyen d'extraire l'information contenue dans la configuration actuelle d'une population.

Il est aussi important de remarquer que cette généalogie nous intéressera seulement si elle contient suffisamment d'information sur la valeur du paramètre que nous tentons d'estimer. Parfois, ce n'est tout simplement pas le cas. Ce manque d'information est fréquent dans le cadre du modèle coalescent. N'oublions pas que celui-ci fait appel à un échantillon de taille arbitraire pris dans une population quelconque dont tous les membres sont dépendants à un certain niveau en raison de leur généalogie commune. Malheureusement, pour cette raison, augmenter la taille de l'échantillon dans certains cas n'a qu'un effet marginal sur la précision de l'inférence.

Chapitre 3

IMPLÉMENTATION D'UNE MÉTHODE DE VRAISEMBLANCE MAXIMALE DANS LE CADRE DU MODÈLE COALESCENT AVEC SÉLECTION ET RECOMBINAISON

La présentation de différents modèles avait pour but de jeter les bases théoriques pour l'élaboration d'une nouvelle méthode de vraisemblance ayant comme objectif la détermination de la position d'une mutation sous sélection quand il y a recombinaison. Cette méthode sera de nature non-bayésienne[1]. Or, les phénomènes de sélection et de recombinaison seront soumis à quelques contraintes qui se doivent d'être précisées. Tout d'abord, la sélection agira sur un seul locus, qui sera flanqué à gauche et à droite d'un nombre égal de loci neutres et polymorphes. Notons en passant que ces derniers ne sont pas nécessairement contigus. Une représentation graphique d'une séquence est donnée par la figure (8). Remarquons aussi qu'un événement de recombinaison ne peut se produire qu'entre les loci 1 et 2. Tous les loci pairs, tout comme tous les loci impairs, sont liés et il n'y a pas d'interférence. Voilà pourquoi la distance séparant les loci périphériques n'a aucune importance. D'autre part, lorsqu'un événement de recombinaison se produit, un point de rupture est choisi au hasard (selon une distribution uniforme) entre les loci 1 et 2. De plus, nous travaillerons dans le cadre du modèle à une infinité de sites. Cela implique concrètement que toutes les mutations à un site donné partagent le même et unique ancêtre. Enfin, conformément au modèle développé par Hudson et Kaplan [8], la sélection au locus central, d'abord assumée stabilisatrice, ensuite assumée génique, aura pour effet de diviser l'échantillon en deux familles, celle des cas, comportant l'allèle mutant au site sous sélection, et celle des témoins, comportant l'allèle primitif à ce même locus. C'est ce que nous appelerons

[1]Pour une comparaison des approches bayésiennes et non-bayésiennes en génétique des populations, voir l'annexe C.

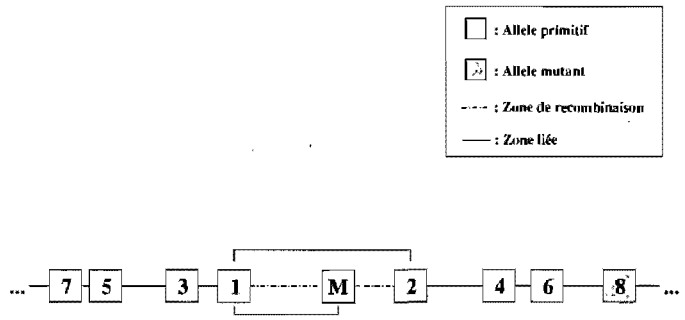

FIG. 8. Représentation graphique d'une séquence. M désigne le locus sous sélection, qui sera soumis soit à un effet de sélection stabilisatrice, soit à un effet de sélection génique. ρ est la distance entre les loci 1 et 2 et ρ_0 est la distance entre le locus 1 et le locus sous sélection.

l'«hypothèse de ségrégation allélique». La justification de cette séparation ainsi que ses répercussions sur l'évolution d'une population seront expliquées à l'instant.

3.1. EFFETS DE LA SÉLECTION

Dans un article écrit pour le *Handbook of Statistical Genetics*, Norborg [**18**] fait remarquer qu'une population génétique polymorphique soumise à un effet de sélection naturelle peut être séparée en familles distinctes analogues aux îlots dans un modèle de propension allélique avec cloisonnement géographique. L'identité au site sous sélection détermine à quelle famille appartient un individu et un événement de recombinaison peut provoquer un événement de «migration» d'une famille à l'autre. Cette approche a tout d'abord été développée par Kaplan et al. [**9**] et étendue au cas où il y a recombinaison par Hudson et Kaplan [**8**].

Afin d'expliquer ce phénomène, il peut être pertinent d'expliquer comment la sélection naturelle vient affecter la généalogie. Dans le développement du modèle de Wright-Fisher à la section 2.1, il a été assumé que chaque gène parent produisait des descendants à la génération suivante selon une même loi, conditionnellement à ce que le nombre de descendants soit égal à la taille de la population. Or, cette hypothèse ne tient pas quand il y a des allèles soumis à un effet de sélection. En effet, un avantage sélectif se traduira par une probabilité plus grande de voir représentés à la génération suivante les gènes avantagés. Si on adopte maintenant une approche rétrospective, de façon analogue, on affirmera que la sélection viendra faire en sorte que tous les gènes parents n'ont plus la même probabilité d'être choisis. Définissons tout d'abord les hypothèses du modèle ainsi que quelques variables.

Nous travaillerons cette fois avec une population diploïde de N individus comportant les génotypes A_1A_1, A_1A_2 et A_2A_2[2] à un locus A. Chaque génotype aura une valeur d'adaptation[3] donnée respectivement par w_{11}, w_{12} et w_{22}. La valeur d'adaptation moyenne à la génération τ sera dénotée par $\bar{w}(\tau)$. Des événements de mutation pourront se produire : A_1 deviendra A_2 avec probabilité u par génération et A_2 deviendra A_1 avec probabilité v par génération. Enfin, des événements de recombinaison entre le locus A, assumé sous sélection, et un locus neutre dénoté B se produiront avec probabilité r par séquence par génération. La fréquence de A_1 à la génération τ sera $X_{2N}(\tau)$. La fréquence de A_2 sera donc $(1 - X_{2N}(\tau))$. Les paramètres et variables précédemment mentionnés prendront les expressions suivantes :

$$w_{11} = w_{12} = w_{22} = 1 + O(\frac{1}{N}), \tag{43}$$

$$\bar{w}(\tau) = X_{2N}(\tau)^2 w_{11}(\tau) + 2X_{2N}(\tau)(1 - X_{2N}(\tau))w_{12}(\tau) + (1 - X_{2N}(\tau))^2 w_{22}(\tau), \tag{44}$$

$$u = \frac{\beta_1}{2N} + O(\frac{1}{N^2}), \tag{45}$$

$$v = \frac{\beta_2}{2N} + O(\frac{1}{N^2}), \tag{46}$$

$$r = \frac{R}{2N} + O(\frac{1}{N^2}), \tag{47}$$

où $\beta_1 > 0$, $\beta_2 > 0$ et $R > 0$.

Définisson $f_{A_j}(A_k, \tau)$ comme la probabilité qu'un gène au locus B de la génération τ choisi au hasard soit lié à un gène de type A_k et que son parent à la génération $\tau - 1$ soit lié à un gène de type A_j. Le cycle menant d'une génération à une autre est représenté à la figure 9. La probabilité $f_{A_j}(A_k, \tau)$ prendra les valeurs suivantes :

[2] A_1A_2 est équivalent à A_2A_1.

[3] Valeur d'adaptation : Mesure de l'aptitude relative à la survie et à la reproduction d'un génotype donné.

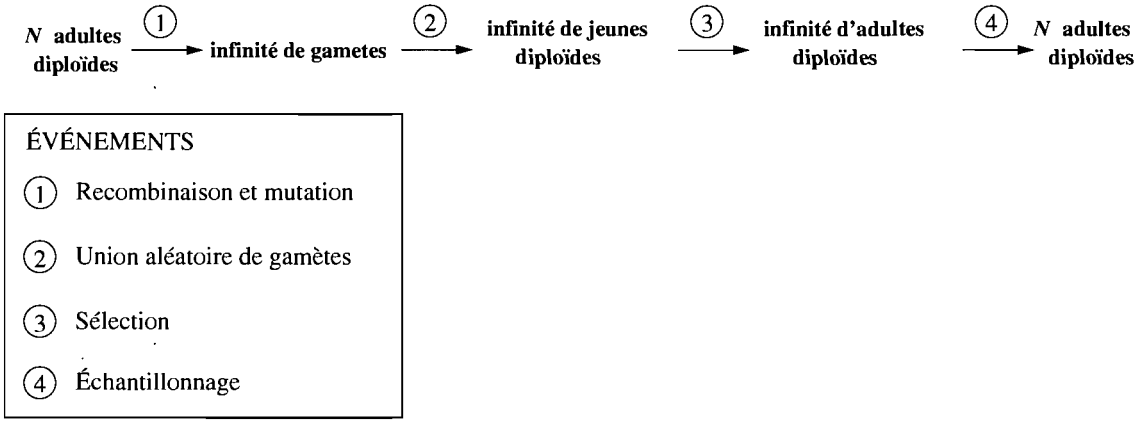

FIG. 9. Cycle de vie des individus d'une population de la génération $\tau - 1$ à la génération τ.

$$
\begin{aligned}
f_{A_1}(A_1,\tau) &= \frac{1}{\bar{w}(\tau-1)}\left(X_{2N}(\tau-1)^2 w_{11} + X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12}\right) + O\left(\frac{1}{N}\right) \\
&= X_{2N}(\tau-1) + O\left(\frac{1}{N}\right), \qquad (48) \\
f_{A_2}(A_2,\tau) &= (1-X_{2N}(\tau-1)) + O\left(\frac{1}{N}\right), \qquad (49)
\end{aligned}
$$

$$
\begin{aligned}
f_{A_1}(A_2,\tau) &= \frac{1}{\bar{w}(\tau-1)}\left(uX_{2N}(\tau-1)^2 w_{11} + (u+r)X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12}\right) + \\
&+ O\left(\frac{1}{N^2}\right) \qquad (50) \\
&= \frac{X_{2N}(\tau-1)(\beta_1 + R(1-X_{2N}(\tau-1)))}{2N} + O\left(\frac{1}{N^2}\right), \qquad (51) \\
f_{A_2}(A_1,\tau) &= \frac{(1-X_{2N}(\tau-1))(\beta_2 + RX_{2N}(\tau-1))}{2N} + O\left(\frac{1}{N^2}\right). \qquad (52)
\end{aligned}
$$

Ces équations peuvent être comprises intuitivement. Pour qu'un gène au locus B lié à A_1 ait comme parent un gène lié aussi à A_1, il faut que le gamète sélectionné pour se reproduire ne mute pas au locus A et n'acquiert pas un allèle A_2 par recombinaison. Or, ces deux événements ont des probabilités d'ordre $\frac{1}{N}$ (comme l'indiquent (45) et (47)). Ainsi, ils entreront dans le terme $O(\frac{1}{N})$. Puisque les gamètes porteurs de A_1 ont une fréquence $X_{2N}(\tau-1)$ à la génération $t-1$, les homozygotes A_1A_1 ont une fréquence $X_{2N}(\tau-1)^2$. Les hétérozygotes ont une fréquence de $2X_{2N}(\tau-1)(1-X_{2N}(\tau-1))$, le facteur 2 provenant du fait que A_1A_2 soit considéré équivalent à A_2A_1. Afin de refléter la valeur sélective de chaque combinaison, les facteurs w_{11} et w_{12} sont respectivement multipliés à la fréquence des homozygotes A_1A_1 et des hétérozygotes. Si

l'individu produit est homozygote A_1A_1, il est certain qu'une séquence choisie parmi les deux le composant comportera A_1. En revanche, pour l'hétérozygote, cette probabilité sera de $\frac{1}{2}$. Ainsi, le taux associé à l'événement «choisir une séquence comportant A_1 provenant d'un parent homozygote A_1A_1» sera de $X_{2N}(\tau-1)^2w_{11}$. Le taux associé à l'événement «choisir une séquence comportant A_1 provenant d'un parent hétérozygote» sera de

$$\frac{1}{2}2X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12} = X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12} \tag{53}$$

et $O(\frac{1}{N})$ regroupera les probabilités des événements de recombinaison et de mutation qui auraient pu faire en sorte que le scénario recherché se produise. Ainsi, le taux total pour l'événement recherché (un gène au locus B lié à A_1 a comme parent un gène lié à A_1) sera de $X_{2N}(\tau-1)^2w_{11} + X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12} + O(\frac{1}{N})$. Pour en faire une probabilité, il nous faut un facteur de normalisation. Celui-ci correspondra au total des taux pour tous les événements. Celui-ci prendra la valeur

$$X_{2N}(\tau-1)^2w_{11} + 2X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12} + (1-X_{2N}(\tau))^2w_{22} = \bar{w}(\tau). \tag{54}$$

Deux événements supplémentaires ont été inclus dans le facteur de normalisation, soient «choisir une séquence comportant A_2 provenant d'un parent homozygote A_2A_2» (avec taux $(1-X_{2N}(\tau))^2w_{22}$) et «choisir une séquence comportant A_2 provenant d'un parent hétérozygote» (avec taux $X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12}$). En divisant $X_{2N}(\tau-1)^2w_{11} + X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12} + O(\frac{1}{N})$ par le facteur de normalisation, on obtient (48). L'équation (49) peut être trouvée de façon similaire.

D'autre part, pour qu'un gène au locus B lié à A_2 ait comme parent un gène lié à A_1, il faut que le gamète sélectionné pour se reproduire ait muté au locus A ou recombiné. La fréquence des gamètes comportant A_2 provenant d'homozygotes A_1A_1 sera de $uX_{2N}(\tau-1)^2$, u étant la probabilité de mutation et $X_{2N}(\tau-1)^2$ étant la fréquence des homozygotes A_1A_1 à la génération $\tau-1$. Notons que la recombinaison ici ne peut pas changer la configuration des gamètes puisque les homozygotes A_1A_1 n'ont pas à la base de gamètes comportant A_2. Bien sûr, certains apparaîtront suite à des événements de mutation, mais l'événement consistant en une mutation suivie d'une recombinaison a un taux d'ordre $\frac{1}{N^2}$. Il sera donc regroupé dans le terme $O(\frac{1}{N^2})$. Le taux trouvé sera par la suite multiplié par la valeur d'adaptation w_{11} afin

de refléter sa force sélective. Ainsi, la probabilité associée à l'événement «choisir une séquence comportant un gène au locus B lié à A_2 dont le parent était un homozygote A_1A_1» sera de $uX_{2N}(\tau-1)^2w_{11}+O(\frac{1}{N^2})$. Si le parent est hétérozygote, des gamètes comportant A_2 seront produits à partir de gamètes comportant A_1 s'il y a mutation au locus A ou recombinaison avec un taux proportionnel à $u+r$. Encore une fois, le taux associé à des événements multiples sera d'ordre $\frac{1}{N^2}$ et sera inclus dans le terme $O(\frac{1}{N^2})$. Puisque les hétérozygotes produisent, avant mutation et recombinaison, des gamètes comportant A_1 et A_2 en nombre égal, la probabilité de choisir un parent porteur de A_1 sera de $\frac{1}{2}$. La fréquence des hétérozygotes dans la population est de $2X_{2N}(\tau-1)(1-X_{2N}(\tau-1))$. Enfin, le paramètre d'adaptation qui leur est associé est de w_{12}. Ainsi, le taux associé à l'événement «choisir une séquence comportant un gène B lié à un gène A_2 dont le parent était lié à A_1» sera de

$$\begin{aligned} &uX_{2N}(\tau-1)^2w_{11}+(u+r)2X_{2N}(\tau-1)(1-X_{2N}(\tau-1))\frac{1}{2}+O(\frac{1}{N^2}) \quad = \\ &uX_{2N}(\tau-1)^2w_{11}+(u+r)X_{2N}(\tau-1)(1-X_{2N}(\tau-1))w_{12}+O(\frac{1}{N^2}). \end{aligned} \tag{55}$$

Afin d'obtenir une probabilité, ce taux sera divisé par le facteur de normalisation (54). Ainsi, nous obtenons l'expression (50). D'une façon analogue, on peut arriver à l'expression (52).

Considérons maintenant le processus stochastique $Q(\tau)=(i,j)$, dénotant, à un moment τ, par i le nombre de gènes ancestraux au locus B dans un échantillon liés à A_1 et par j le nombre de ceux liés à A_2. Tout d'abord, nous sommes intéressés à trouver les taux de coalescence, i.e. le taux des événements diminuant de 1 la valeur de i, sans affecter j ou vice-versa. Ceci correspond au cas où $|Q(\tau-1)|\neq|Q(\tau)|$, $|Q(\tau)|$ étant le nombre total de séquences dans l'échantillon à la génération τ. La proportion des gènes au locus B de la génération τ fournie par un certain parent porteur de A_1 sera de

$$\frac{X_{2N}(\tau-1)w_{11}+(1-X_{2N}(\tau-1))w_{12}}{2N\bar{w}(\tau-1)}+O(\frac{1}{N^2}) \tag{56}$$

$$=\frac{f_{A_1}(A_1,\tau)}{2NX_{2N}(\tau-1)}+O(\frac{1}{N^2}). \tag{57}$$

L'équation (56) devient plus claire quand on réalise que le numérateur représente l'espérance de la valeur d'adaptation d'un gène parent dont le locus B est lié à A_1 à la génération $\tau-1$ et le dénominateur, l'espérance de la valeur d'adaptation totale de la population. En effet, à la génération $\tau-1$, un gène lié à A_1 sera couplé avec un autre gène lié à A_1 avec probabilité $X_{2N}(\tau-1)$

et dans ce cas, il aura comme valeur d'adaptation w_{11}. S'il est couplé avec un gène lié à A_2, avec probabilité $(1 - X_{2N}(\tau - 1))$, il aura comme valeur d'adaptation w_{12}. Puisqu'il y a $2N$ gènes au locus B dans la population et puisque chaque gène a $\bar{w}(\tau - 1)$ comme valeur d'adaptation espérée, l'espérance de la valeur totale d'adaptation de la population sera de $2N\bar{w}(\tau - 1)$. Notons enfin que $O(\frac{1}{2N^2}$ regroupe les termes pour les événements de mutation et de recombinaison.

Par conséquent, la probabilité que deux gènes au locus B liés à A_1 aient le même parent aussi lié à A_1 sera de

$$2NX_{2N}(\tau - 1)\left(\frac{f_{A_1}(A_1, \tau)}{2NX_{2N}(\tau - 1)f(A_1, \tau)}\right)^2 + O(\frac{1}{N^2}) = \frac{1}{2NX_{2N}(\tau - 1)} + O(\frac{1}{N^2}), \quad (58)$$

où $f(A_1, \tau)$ est la probabilité qu'un certain gène à la génération τ soit lié à A_1 ($f(A_j, \tau) = f_{A_1}(A_j, \tau) + f_{A_2}(A_j, \tau)$, $j = 1, 2$), peu importe la configuration du parent. L'équation précédente se comprend aisément si on réalise que $\left(\frac{f_{A_1}(A_1, \tau)}{2NX_{2N}(\tau - 1)f(A_1, \tau)}\right)^2$ représente approximativement la probabilité que deux gènes aient un parent donné lié à A_1, à condition que ces deux gènes soient liés à A_1, et que puisqu'il y a $2NX_{2N}(\tau - 1)$ gènes liés à A_1 à la génération $\tau - 1$, la probabilité que deux gènes aient un même parent A_1 sera de $2NX_{2N}(\tau - 1)\left(\frac{f_{A_1}(A_1, \tau)}{2NX_{2N}(\tau - 1)f(A_1, \tau)}\right)^2 + O(\frac{1}{N^2})$.

Puisqu'il y a $\binom{i}{2}$ paires de gènes ancestraux au locus B liés à A_1 à la génération τ, on a

$$P(Q(\tau - 1) = (i - 1, j)|Q(\tau) = (i, j), X_{2N}(\tau - 1)) = \binom{i}{2}\frac{1}{2NX_{2N}(\tau - 1)} + O(\frac{1}{N^2}) \quad (59)$$

et, puisqu'il y en a $\binom{j}{2}$ liés à A_2, on a également

$$P(Q(\tau - 1) = (i, j - 1)|Q(\tau) = (i, j), X_{2N}(\tau - 1)) = \binom{j}{2}\frac{1}{2N(1 - X_{2N}(\tau - 1))} + O(\frac{1}{N^2}). \quad (60)$$

La présence du facteur $\binom{i}{2}$ ou $\binom{j}{2}$ s'explique par le fait qu'on cherche à savoir la probabilité de tirer une paire de gènes ou plus partageant un parent parmi $\binom{i}{2}$ ou $\binom{j}{2}$ paires au total. Plus précisément, cette probabilité serait, pour i fixe, de (sachant que la probabilité que deux paires ou plus de gènes dans l'échantillon aient un parent en commun sera regroupée dans $O(1/N^2)$)

$$\begin{aligned} P(Q(\tau-1) &= (i,j-1)|Q(\tau)=(i,j), X_{2N}(\tau-1)) \\ &= \binom{j}{2}\left(\frac{1}{2N(1-X_{2N}(\tau-1))}+O(\frac{1}{N^2})\right)\times \\ &\quad \left(1-(\frac{1}{2N(1-X_{2N}(\tau-1))}+O(\frac{1}{N^2}))\right)^{\binom{j}{2}-1} \\ &= \binom{j}{2}\frac{1}{2N(1-X_{2N}(\tau-1))}(1+O(\frac{1}{N})+O(\frac{1}{N^2})) \\ &= \binom{j}{2}\frac{1}{2N(1-X_{2N}(\tau-1))}+O(\frac{1}{N^2}). \end{aligned} \tag{61}$$

Les probabilités données précédemment sont bien sûr des probabilités de coalescence. On remarque donc que la probabilité de coalescence dans la sous-population des gènes liés à A_1 ou à A_2 ne dépend pas du nombre d'individus dans l'autre sous-population. Ainsi, nous parvenons à prouver que la sélection sur le processus de coalescence a un effet analogue à celui du cloisonnement géographique.

Ensuite, nous devons calculer le taux de recombinaison dans chaque sous-population. Un événement de recombinaison est analogue à un événement de migration, i.e. $Q(\tau)=(i,j)$ avec $Q(\tau-1)=(i-1,j+1)$ ou $Q(\tau-1)=(i+1,j-1)$. Ces probabilités de transition sont données par

$$\begin{aligned} P(Q(\tau-1) &= (i+1,j-1)|Q(\tau)=(i,j), X_{2N}(\tau-1)) \\ &= \frac{jf_{A_1}(A_2,\tau)}{f(A_2,\tau)}+O(\frac{1}{N^2}) \qquad (62) \\ &= j\left(\frac{X_{2N}(\tau-1)}{1-X_{2N}(\tau-1)}\right)\frac{\beta_1+R(1-X_{2N}(\tau-1))}{2N}+O(\frac{1}{N^2}), \qquad (63) \end{aligned}$$

et par

$$\begin{aligned} P(Q(\tau-1) &= (i-1,j+1)|Q(\tau)=(i,j), X_{2N}(\tau-1)) \\ &= i\left(\frac{1-X_{2N}(\tau-1)}{X_{2N}(\tau-1)}\right)\frac{\beta_2+R(X_{2N}(\tau-1))}{2N}+O(\frac{1}{N^2}). \end{aligned} \tag{64}$$

On peut comprendre l'équation (62) en se rappelant que $\frac{f_{A_1}(A_2,\tau)}{f(A_2,\tau)}+O(\frac{1}{N^2})$ correspond à la probabilité qu'un gène au locus B en particulier de la génération τ ait un parent lié à un gène A_1 à condition que le gène tiré soit lié à A_2. On multiplie le tout par j puisqu'il y a j gènes ancestraux liés à A_2 à la génération τ (le raisonnement expliquant la présence de j est le même

que celui conduisant à (61)). Un raisonnement similaire permet d'arriver à (64).

Maintenant que ces probabilités ont été calculées, il est possible de calculer la distribution du temps entre événements pour le processus $Q(\tau)$ quand $N \to \infty$ et quand le temps est exprimé en unités de $2N$ générations. Commençons par calculer $P(Q(\tau-1) = Q(\tau)|Q(\tau) = (i,j), X_{2N}(\tau-1))$. On a

$$\begin{aligned} P(Q(\tau-1) &= Q(\tau)|Q(\tau) = (i,j), X_{2N}(\tau-1)) \\ &= 1 - \frac{h_{ij}(X_{2N}(\tau-1))}{2N} + O(\frac{1}{N^2}), \end{aligned} \tag{65}$$

où

$$h_{ij}(x) = \frac{\binom{i}{2}}{x} + \frac{\binom{j}{2}}{1-x} + \frac{j(\beta_1 + R(1-x))x}{1-x} + \frac{i(\beta_2 + Rx)(1-x)}{x}. \tag{66}$$

Maintenant, si l'on désire obtenir un taux de changement instantané, i.e. sur un intervalle de temps de longueur infinitésimale, on doit rééchelonner le temps. Réexprimons donc le temps en unités de $2N$ générations ($\tau = \lfloor 2Nt \rfloor$) et laissons la taille de la population tendre vers l'infini. Ainsi, nous obtenons comme taux instantané $h_{ij}(Y(t))$ à l'instant $t \geq 0$ où $Y(t) = \lim_{N\to\infty} X_{2N}(\lfloor 2Nt \rfloor)$.

Six types d'événements peuvent se produire indépendamment :

(1) Coalescence dans la famille des gènes liés à A_1 ;

(2) Coalescence dans la famille des gènes liés à A_2 ;

(3) Mutation de A_1 à A_2 ;

(4) Mutation de A_2 à A_1 ;

(5) Recombinaison d'un gène lié à A_1 avec un gène lié à A_2 ;

(6) Recombinaison d'un gène lié à A_2 avec un gène lié à A_1.

Chacun de ces événements se voit attribuer un taux lié à sa contribution à $h_{ij}(Y(t))$. Ainsi, conformément à la liste précédente, nous pouvons trouver les taux suivants :

(1) $\frac{\binom{i}{2}}{Y(t)}$;

(2) $\frac{\binom{j}{2}}{1-Y(t)}$;

(3) $\frac{i\beta_2(1-Y(t))}{Y(t)}$;

(4) $\frac{j\beta_1 Y(t)}{1-Y(t)}$;

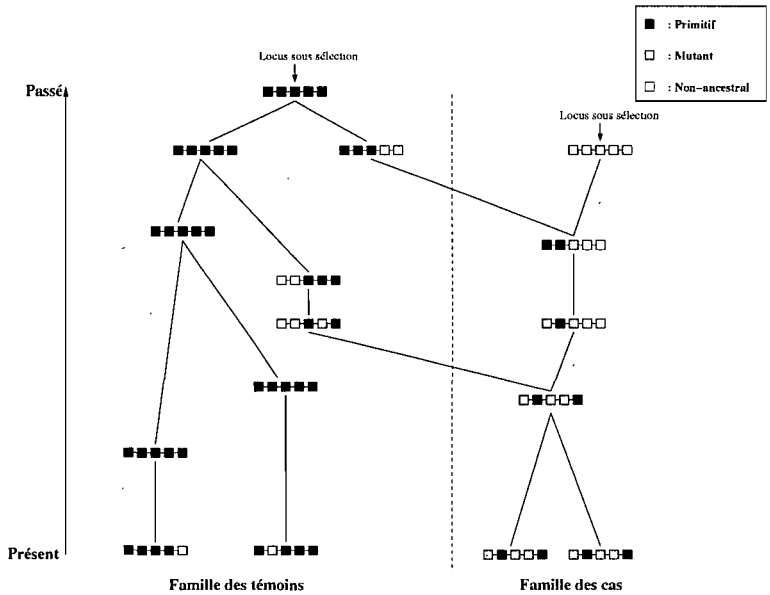

FIG. 10. Une réalisation du graphe de recombinaison ancestral quand il y a sélection au locus central. Le cloisonnement créé par la sélection ou, en d'autres mots, la séparation en deux famille distinctes, se doit d'être remarqué.

(5) $iR(1 - Y(t))$;

(6) $jRY(t)$.

L'hypothèse de sélection stabilisatrice implique que la mutation au locus sous sélection est ancienne. En effet, si elle était assumée récente, nous observerions une variation dans sa fréquence, ce qui contredirait notre hypothèse de départ selon laquelle la fréquence du mutant ne change plus. Si elle est en plus assumée unique, un nouvel événement de mutation devient impossible au site sous sélection. Il ne peut donc pas y avoir d'événements de migration causés par une nouvelle mutation. Nous considérerons alors que β_1 et β_2 sont égaux à 0. La figure (10) est un exemple du graphe de recombinaison ancestrale construit sous l'effet de l'hypothèse de sélection à un locus sous sélection.

3.2. Première partie : génération de l'échantillon

3.2.1. Locus central soumis à un effet de sélection stabilisatrice

L'absence de données concrètes à l'aide desquelles on pourrait tester la validité de notre modèle nous oblige à générer aléatoirement un échantillon de segments chromosomiques qui sera utilisé à cette fin. L'échantillon comptera n séquences, $\frac{n}{2}$ dans la famille des cas, et $\frac{n}{2}$ dans la famille des témoins, comportant chacune b loci. Si l'on assume qu'un locus est soumis à un effet de sélection stabilisatrice, l'échantillon sera généré à l'aide de l'algorithme suivant :

(1) Un temps inter-événement est généré. Ce temps est distribué exponentiellement avec moyenne $\frac{1}{\lambda}$ où λ correspond à la somme des taux pour chaque événement possible. Quatre types d'événements peuvent se produire :

- Coalescence dans la famille des cas ;
- Coalescence dans la famille des témoins ;
- Recombinaison avec une séquence de la famille des cas (voir cas #1 dans la figure (11)) ;
- Recombinaison avec une séquence de la famille des témoins (voir cas #2 dans la figure (11)).

Le taux pour chaque événement est (en lien avec la liste précédente) :

- $\frac{n_C(n_C-1)}{2p}$;
- $\frac{n_T(n_T-1)}{2(1-p)}$;
- $\frac{n\beta\rho}{2}p$;
- $\frac{n\beta\rho}{2}(1-p)$;

où p, constant, correspond à la prévalence des cas dans la population et n_C (n_T) au nombre d'individus dans la famille des cas (témoins) au moment de l'événement. La quantité ρ est définie de la même façon que dans la section 2.3 et β est défini par l'équation (30). Notons que les taux précédents ont été obtenus à la section 3.1[4]. Les taux de recombinaison comportent toutefois maintenant un facteur β, ajouté, tel qu'énoncé précédemment, afin de rendre compte du fait qu'un événement de recombinaison ne se produisant pas entre deux sites ancestraux n'aura pas d'impact sur la forme de l'ARG.

[4] A_1 correspond au gène mutant et p, à la valeur constante sur t de $Y(t)$. La convention sur l'écoulement du temps est toutefois inversée, i.e. $t_1 > t_2$ signifie que le point t_1 est antérieur à t_2.

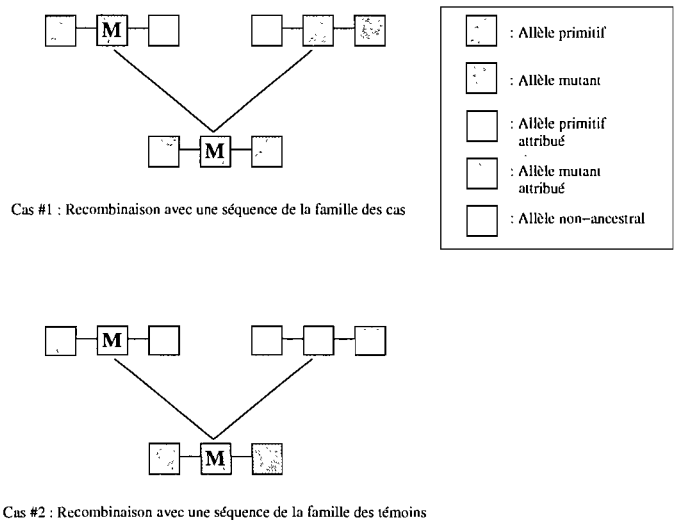

FIG. 11. Les deux types d'événements de recombinaison. Les allèles dits «attribués» ne sont pas ancestraux à la séquence résultante. Ils sont choisis aléatoirement en fonction de la fréquence des séquences de la famille des cas et des témoins au moment de l'événement de recombinaison. Le M indique quelle séquence a transmis l'allèle au locus sous sélection.

(2) Un événement est généré. Chaque événement a comme probabilité le taux de l'événement divisé par λ.

(3) Les étapes 1 et 2 seront répétées jusqu'à ce que le nombre total de gènes ancestraux soit égal à $b + 1$, et non b, puisque la sélection stabilisatrice empêchera l'allèle mutant au site sous sélection de disparaître. Ainsi, le matériel ancestral final comportera deux séquences, une dans la famille des cas et l'autre dans la famille des témoins, se partageant les allèles périphériques.

(4) Une mutation est ajoutée à chaque locus (sauf au locus sous sélection). Notons que les mutations sont placées sur les arbres marginaux correspondant à chaque locus, et non sur le graphe de recombinaison ancestral lui-même. La position de la mutation sur chaque arbre est distribuée de façon uniforme sur toutes ses branches, ce qui représente une conséquence directe de l'hypothèse selon laquelle le polymorphisme observé à un locus est le résultat d'un seul événement de mutation (voir section 2.2). La mutation sera par la suite transférée à toutes les séquences des générations suivantes liées au point choisi.

3.2.2. Locus central soumis à un effet de sélection génique

L'abandon de l'hypothèse de sélection stabilisatrice nous oblige à modéliser la fréquence de l'allèle sous sélection à chaque instant. La propension de cet allèle sera modélisée par un

processus de diffusion décrit à la section 3.2.3. L'hypothèse de ségrégation allélique peut être maintenue. Les événements suivants se produiront :

- Événement de coalescence dans la famille F ($F = M$ pour les cas et A pour les témoins) au temps t avec taux
$$\binom{n_F}{2}\frac{1}{Y^{(F)}(t)}, \tag{67}$$
- Événement de recombinaison vers la famille F au temps t avec taux
$$\frac{n\beta\rho Y^{(F)}(t)}{2}. \tag{68}$$

Notons que $Y^{(F)}(t)$ correspond à la fréquence de la famille F dans la population au temps t et n_F, au nombre de séquences à ce moment dans la famille F.

La mutation au locus sous sélection disparaîtra quand $Y^{(M)}(t)$ atteindra la valeur 0 en remontant le temps. Puisque la probabilité de coalescence tend vers 1 quand la fréquence tend vers 0, nous n'obtenons pas de contradiction avec l'hypothèse selon laquelle la mutation doit être unique lors de sa disparition. En d'autres mots, plus on se rapproche du temps de disparition de l'allèle mutant, plus le taux de coalescence dans la famille des cas augmente. Ceci aura pour conséquence de ramener à 1 le nombre de gènes de type mutant immédiatement avant leur disparition. Ainsi, la disparition n'affectera qu'une séquence et l'hypothèse d'unicité de l'ancêtre de l'allèle mutant sous sélection ne sera pas contredite. Quand le UA (l'ancêtre ultime) aura été trouvé (i.e. quand le nombre de séquences ancestrales sera de 1 pour la première fois), la simulation d'événements de coalescence et de recombinaison sera interrompue et les mutations aux différents loci seront alors simulées.

Pour déterminer quel événement se produira, on procède comme suit. Tout d'abord, un temps d'arrivée T_i pour chaque événement i possible sera calculé. Puis, $T = \min_{i=1,2,\ldots,\Omega}(T_i)$, Ω correspondant au nombre total d'événements possibles, déterminera le temps de l'événement se produisant. Ensuite, la nature de l'événement se produisant sera choisie aléatoirement, chaque type d'événement se voyant attribuer une probabilité proportionelle à sa contribution au taux total. Puisque les processus générant les événements sont des processus de Poisson indépendants, le taux total correspondra à la somme des taux pour les événements individuels, i.e.

$$\lambda(t) = \sum_{i=1}^{\Omega} \lambda_i(t). \tag{69}$$

De plus, tel qu'énoncé précédemment, on aura

$$P(T = T_i | T = t) = \frac{\lambda_i(t)}{\lambda(t)}. \tag{70}$$

Les processus de recombinaison[5] et de mutation ne dépendent pas des fréquences alléliques au locus sous sélection. Ils sont donc des processus de Poisson homogènes. Générer un temps d'arrivée pour le processus de coalescence n'est pas aussi simple, vu la relation directe entre le taux de coalescence et la fréquence de l'allèle sous sélection dans la population à chaque moment. Le processus de coalescence est donc un processus de Poisson non-homogène. Pour ce type d'événement, la génération d'un temps d'arrivée, T_i, se fera par interpolation basée sur la relation suivante :

$$\begin{aligned} U_i &= 1 - \exp\left(- \int_{t_0}^{T_i} \lambda_i(v) dv \right) \\ &= 1 - \exp\left(- \int_{t_0}^{T_i} \binom{n_F}{2} \frac{1}{Y^{(F)}(v)} dv \right), \end{aligned} \tag{71}$$

et donc

$$\frac{-\log(1 - U_i)}{\binom{n_F}{2}} = \int_{t_0}^{T_i} \frac{1}{Y^{(F)}(v)} dv, \tag{72}$$

où

- U_i suit une loi uniforme sur $[0, 1]$;
- n_F désigne le nombre de séquences dans la famille F à laquelle appartient les séquences de type i avec

$$F = \begin{cases} C, \text{ si } i \text{ est dans la famille des cas,} \\ T, \text{ si } i \text{ est dans la famille des témoins;} \end{cases} \tag{73}$$

- t_0 est le temps présent ;
- $Y^{(F)}(v)$ est la fréquence de la famille F dans la population au temps v.

Ainsi, pour obtenir les temps de coalescence, on doit tout d'abord générer $Y^{(F)}(v)$. Ensuite, on obtient une valeur pour la variable aléatoire U_i et par une technique d'interpolation, on trouve la valeur de T_i qui y est associée.

[5]Pris globalement, c'est-à-dire nonobstant la famille à laquelle appartiendra la séquence dont le gène au locus sous sélection est non-ancestral

3.2.3. Approximation d'un processus de diffusion par un processus de naissance et de mort basé sur le modèle de Moran

En général, on peut générer la fréquence du type M dans une grande population ($Y^{(C)}(t)$ dans la section 3.2.2) en fonction du temps à l'aide d'un processus de diffusion circonscrit entre 0 et 1 dont le générateur prend la forme [**11**]

$$L = \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2} + \mu(x)\frac{\partial}{\partial x}, \tag{74}$$

$\sigma^2(x)$ étant la fonction de diffusion et $\mu(x)$ étant la fonction de dérive.

Or, Griffiths [**4**] propose d'approximer ce processus de diffusion par un processus de naissance et de mort basé sur le modèle de Moran ayant une population de base de $2N$ (N devant être assumé très grand).

Le modèle de Moran [**2**] en temps discret est une chaîne de Markov. Assumons une population de taille constante formée de $2N$ individus haploïdes comportant à un locus déterminé soit un allèle de type A (primitif) ou de type M (mutant). Quand un individu choisi au hasard se reproduit, son descendant vient prendre la place d'un autre individu, lui aussi choisi au hasard. Il peut même remplacer son parent. L'approche temporelle est pour l'instant prospective.

Soit $Z_{2N}^{(M)}(\tau)$, le nombre d'individus de type M au moment τ. Si $Z_{2N}^{(M)}(\tau) = j$ et s'il n'y a pas de sélection, alors, on a

$$Z_{2N}^{(M)}(\tau+1) = \begin{cases} j+1 & \text{, avec probabilité } \frac{j}{2N}\left(\frac{2N-j}{2N}\right) = \lambda_j^{(M)}, \\ j-1 & \text{, avec probabilité } \frac{j}{2N}\left(\frac{2N-j}{2N}\right) = \mu_j^{(M)}, \\ j & \text{, avec probabilité } 1-\lambda_j^{(M)}-\mu_j^{(M)}. \end{cases} \tag{75}$$

Notons que 0 et $2N$ sont des états absorbants. Par conséquent, dans la famille M, on peut voir un événement de naissance ($j \to j+1$) comme un événement de reproduction d'un gène de type M, dont le descendant résultant viendra prendre la place d'un gène de type A et un événement de mort ($j \to j-1$) comme un événement de reproduction d'un gène de type A, dont le descendant résultant viendra prendre la place d'un gène de type M.

La sélection naturelle en faveur de M peut être incorporée au modèle de Moran en multipliant $\lambda_j^{(M)}$ par un facteur $(1 + s_{2N}(j)/2)$ et $\mu_j^{(M)}$ par un facteur de $(1 - s_{2N}(j)/2)$, $s_{2N}(j) \geq 0$ étant appelé «coefficient de sélection». Dans ce cas, on a

$$\lambda_j^{(M)} = (1 + \frac{s_{2N}(j)}{2})\frac{j}{2N}\frac{2N-j}{2N}, \tag{76}$$

$$\mu_j^{(M)} = (1 - \frac{s_{2N}(j)}{2})\frac{j}{2N}\frac{2N-j}{2N}. \tag{77}$$

Ainsi, on remarque que le taux de naissance pour le type M étant plus élevé que le taux de mort, il aura tendance à se répandre dans la population. Remarquons aussi que $(1 - s_{2N}(j)/2)$ correspond au taux de reproduction pour les gènes de type A tandis que $(1 + s_{2N}(j)/2)$ correspond à ce taux pour les gènes de type M.

Dans le modèle de Moran sans sélection, si on laisse $N \to \infty$ et si on rééchelonne le temps en unité de $2N^2$ événements de naissance et de mort ($\tau = \lfloor 2N^2 t \rfloor$), on obtient que $\frac{Z_{2N}^{(M)}(\lfloor 2N^2 t \rfloor)}{2N} = Y_{2N}^{(M)}(t)$ converge en loi vers un processus de diffusion $Y^{(M)}(t)$ avec

$$\mu(y) = 0, \tag{78}$$

$$\sigma^2(y) = y(1-y). \tag{79}$$

Remarquons que ces quantités sont égales à celles données par les équations (10) et (11). Il est intéressant ici de remarquer qu'en temps continu, quand $N \to \infty$ et quand il n'y a pas d'effet de sélection, malgré des hypothèses de base différentes, une population décrite par le modèle de Moran évolue de façon analogue à une population décrite par le modèle de Wright-Fisher. Notons aussi que l'analogie entre les deux modèles subsiste même si l'on y incorpore les phénomènes de mutation et de sélection [**2**].

Tel que prévu, le paramètre de dérive $\mu(y)$ prendra la valeur 0 sous neutralité. En effet, la neutralité se traduisant concrètement par une égalité entre le taux de naissance et le taux de mort, il est normal que nous n'observions pas de dérive dans le cas limite. On peut dériver ce résultat en notant tout d'abord que si $\Delta Y_{2N}^{(M)}(t) = Y_{2N}^{(M)}(t + \delta t) - Y_{2N}^{(M)}(t)$ et $\delta t = \frac{1}{2N^2}$, alors

$$E[\Delta Y_{2N}^{(M)}(t)|Y_{2N}^{(M)}(t)=y] = N(\lambda_j^{(M)}-\mu_j^{(M)})\delta t, \tag{80}$$

$$Var[\Delta Y_{2N}^{(M)}(t)|Y_{2N}^{(M)}(t)=y] = \frac{\lambda_j^{(M)}+\mu_j^{(M)}}{2}\delta t + o(\delta t), \tag{81}$$

où $y=\frac{j}{2N}$.

Ces valeurs peuvent se trouver ainsi :

$$\begin{aligned} E[\Delta Y_{2N}^{(M)}(t)|Y_{2N}^{(M)}(t)=y] &= \frac{1}{2N}\lambda_j^{(M)} - \frac{1}{2N}\mu_j^{(M)} \\ &= N(\lambda_j^{(M)}-\mu_j^{(M)})\delta t, \end{aligned} \tag{82}$$

$$\begin{aligned} Var[\Delta Y_{2N}^{(M)}(t)|Y_{2N}^{(M)}(t)=y] &= \left(\frac{1}{2N}\right)^2\lambda_j^{(M)} + \left(\frac{1}{2N}\right)^2\mu_j^{(M)} - (N(\lambda_j^{(M)}-\mu_j^{(M)})\delta t)^2 \\ &= \frac{\lambda_j^{(M)}+\mu_j^{(M)}}{2}\delta t + o(\delta t). \end{aligned} \tag{83}$$

$$\tag{84}$$

Ensuite, quand $\delta t \to 0$, c'est-à-dire $N \to \infty$, on parvient à

$$\begin{aligned} \lim_{N\to\infty} N(\lambda_j^{(M)}-\mu_j^{(M)}) &= \lim_{N\to\infty} N\left(\frac{j}{2N}\left(\frac{2N-j}{2N}\right) - \frac{j}{2N}\left(\frac{2N-j}{2N}\right)\right) \qquad (85) \\ &= 0 \qquad (86) \end{aligned}$$

et

$$\begin{aligned} \lim_{N\to\infty} \frac{\lambda_j^{(M)}+\mu_j^{(M)}}{2} &= \lim_{N\to\infty} 2N^2\left(\frac{i}{2N}\left(\frac{2N-i}{2N}\right) + \frac{i}{2N}\left(\frac{2N-i}{2N}\right)\right. \\ &= \lim_{N\to\infty}\left(\frac{i}{2N}\left(\frac{2N-i}{2N}\right)\right) \\ &= y(1-y). \end{aligned} \tag{87}$$

On peut bien sûr ajouter un effet de sélection au modèle de diffusion. Dans ce cas, une hypothèse supplémentaire est requise pour s'assurer de l'existence du processus limite. Nous devrons assumer que

$$Ns_{2N}(j) \to \frac{1}{2}\beta(y), \tag{88}$$

quand $N \to \infty$. La quantité $\beta(y)$ est appelée «taux de sélection». Avec ceci en tête, il est possible de déduire, par une méthode similaire à celle employée ci-haut, que les paramètres du processus de diffusion prendront les expressions

$$\mu(y) = \frac{1}{2}\beta(y)y(1-y), \tag{89}$$

$$\sigma^2(y) = y(1-y). \tag{90}$$

Si $\beta(y)$ n'est pas égal à 0, on considère que l'allèle M est soumis à un effet de sélection. Si $\beta(y)$ n'est pas égal à 0 et est constant, alors la sélection sera qualifiée de «génique».

Considérons maintenant le processus de naissance et de mort à temps continu $\tilde{Z}_{2N}^{(\cdot)}(t)$, pouvant prendre les valeurs $0, 1, ..., 2N$. Les paramètres $\tilde{\lambda}_j^{(\cdot)}$ et $\tilde{\mu}_j^{(\cdot)}$ représentent maintenant des taux instantanés de naissance et de mort. Pour qu'il puisse y avoir convergence de $\frac{\tilde{Z}_{2N}^{(\cdot)}(t)}{2N}$ vers un processus de diffusion avec le générateur (74), les taux du processus doivent satisfaire [4],

$$\frac{\tilde{\lambda}_j^{(\cdot)} - \tilde{\mu}_j^{(\cdot)}}{2N} \to \mu(y), \tag{91}$$

et

$$\frac{\tilde{\lambda}_j^{(\cdot)} + \tilde{\mu}_j^{(\cdot)}}{4N^2} \to \sigma^2(y), \tag{92}$$

ce qui sera le cas lorsque

$$\tilde{\lambda}_j^{(\cdot)} = N(2N\sigma^2(y) + \mu(y)), \tag{93}$$

et

$$\tilde{\mu}_j^{(\cdot)} = N(2N\sigma^2(y) - \mu(y)). \tag{94}$$

Cette constatation est à la base de l'approximation qui sera utilisée par la suite. En effet, elle nous permettra de générer approximativement un processus de diffusion en générant plutôt un processus de naissance et mort.

Concentrons-nous sur le processus d'approximation $\tilde{Z}_{2N}^{(M)}(t)$ pour le nombre de gènes mutants. Quand il y a dans l'échantillon j gènes de type mutant, on peut trouver les taux de

naissance et de mort appropriés en incorporant à (93) et (94) les expressions (89) et (90). Nous aurons donc

$$\tilde{\lambda}_j^{(M)} = (1 + \tfrac{s_{2N}(j)}{2})\tfrac{j}{2}(2N - j),$$

$$\tilde{\mu}_j^{(M)} = (1 - \tfrac{s_{2N}(j)}{2})\tfrac{j}{2}(2N - j),$$

où $s_{2N}(j) = \frac{\beta(y)}{2N}$. Il est à noter que ces taux correspondent aux taux de mort et de naissance, respectivement, dans la famille des gènes non-mutants, c'est-à-dire primitifs. On a donc $\tilde{\lambda}_j^{(M)} = \tilde{\mu}_{2N-j}^{(A)}$ et $\tilde{\mu}_j^{(M)} = \tilde{\lambda}_{2N-j}^{(A)}$.

Puisque $\tilde{\lambda}_j^{(M)} > \tilde{\mu}_j^{(M)}$, on dira que l'allèle mutant a un avantage sélectif sur l'allèle primitif.

Pour obtenir les fréquences alléliques dans le cadre du modèle dont fait l'objet ce mémoire, nous avons plutôt besoin de générer le processus de façon rétrospective. Or, le processus inversé, représenté par $\tilde{Z}_{2N}^{(M)*}(t)$, a la même distribution que $\tilde{Z}_{2N}^{(M)}(t)$ étant donné une éventuelle absorption à 0. Cette constatation est basée sur le principe de réversibilité temporelle du processus de naissance et de mort.

Les taux de naissance et de mort de $\tilde{Z}_{2N}^{(M)*}(t)$ seront[6]

$$\tilde{\lambda}_j^{(M)*} = \tilde{\lambda}_j^{(M)} \frac{u_{j+1}}{u_j}, \tag{95}$$

$$\tilde{\mu}_j^{(M)*} = \tilde{\mu}_j^{(M)} \frac{u_{j-1}}{u_j}, \tag{96}$$

où

$$u_j = \frac{1 - \alpha^{2N-j}}{1 - \alpha^{2N}}, \tag{97}$$

avec

$$\alpha = \frac{1 + \frac{\beta}{4N}}{1 - \frac{\beta}{4N}}, \tag{98}$$

où β correspond encore au taux de sélection.

Le principe de réversibilité temporelle ne sera pas prouvé ici. Or, la valeur des taux énoncés précédemment peut être trouvée à l'aide du raisonnement suivant.

[6]La sélection est assumée génique.

Afin de dériver les valeurs précédentes, on peut se concentrer plutôt sur le processus stochastique intrinsèque $\tilde{Z}_{2N}^{(M)}(\tau)$, une chaîne de Markov en temps discret comportant deux états absorbants, 0 et $2N$, pour laquelle le temps est calculé en événements de naissance et de mort. Ewens [1] cite comme résultat

$$p_{ji}^{*} = p_{ji}\frac{\pi_i}{\pi_j}, \tag{99}$$

p_{ji} correspondant à la probabilité de transition de l'état j à l'état i de la chaîne, p_{ji}^{*}, à cette même probabilité à condition que $2N$ soit atteint éventuellement, et π_i, à la probabilité que la chaîne atteigne éventuellement $2N$ si elle est présentement dans l'état i.

En effet, en définissant $D(\tau)$, une chaîne de Markov avec la matrice de transition p_{ji}, comme étant égal à $\tilde{Z}_{2N}^{(M)}(\tau)$, on a que

$$\begin{aligned} p_{ji}^{*} &= P(D(\tau+1) = i | D(\tau) = j, \{\text{Absorption à } 2N\}) \\ &= \frac{P(D(\tau+1) = i, \{\text{Absorption à } 2N\} | D(\tau) = j)}{P(\{\text{Absorption à } 2N\})} \\ &= p_{ji}\frac{\pi_i}{\pi_j}, \end{aligned} \tag{100}$$

ce qui est le résultat cité.

De là, on peut déduire que

$$\begin{aligned} p_{j,j+1}^{*} &= p_{j,j+1}\frac{\pi_{j+1}}{\pi_j}, \\ \frac{\tilde{\lambda}_j^{*}}{\tilde{\lambda}_j^{*} + \tilde{\mu}_j^{*}} &= \frac{\tilde{\lambda}_j}{\tilde{\lambda}_j + \tilde{\mu}_j}\frac{\pi_{j+1}}{\pi_j}, \\ \frac{\tilde{\lambda}_j^{*}}{\tilde{\lambda}_j + \tilde{\mu}_j} &= \frac{\tilde{\lambda}_j}{\tilde{\lambda}_j + \tilde{\mu}_j}\frac{\pi_{j+1}}{\pi_j}, \\ \tilde{\lambda}_j^{*} &= \tilde{\lambda}_j\frac{\pi_{j+1}}{\pi_j}, \end{aligned} \tag{101}$$

puisque $\tilde{\lambda}_j^{*} + \tilde{\mu}_j^{*} = \tilde{\lambda}_j + \tilde{\mu}_j$ (le conditionnement n'affectera pas le taux de changement total). Notons que le raisonnement précédent nous permet aussi de déduire que

$$\tilde{\mu}_j^{*} = \tilde{\mu}_j\frac{\pi_{j-1}}{\pi_j}. \tag{102}$$

Considérons maintenant un autre processus stochastique, $\tilde{Z}_{2N}^{(A)}(\tau)$, donnant plutôt au temps τ le nombre de gènes primitifs. Puisqu'un événement de mort dans la famille des gènes mutants constitue un événement de naissance dans la famille des gènes primitifs, nous pouvons affirmer qu'un événement de naissance dans $\tilde{Z}_{2N}^{(M)}(\tau)$ étant donné une absorption à $2N$ correspond à un événement de mort dans $\tilde{Z}_{2N}^{(A)}(\tau)$ étant donné une absorption à 0 et vice versa. Ainsi, si à l'instant τ, il y a j gènes mutants dans la population et $2N-j$ gènes primitifs, le taux de mortalité dans le processus $\tilde{Z}_{2N}^{(A)}(\tau)$, $\tilde{\mu}_{2N-j}^{(A)}$, correspond au taux de naissance pour $\tilde{Z}_{2N}^{(M)}(\tau)$, $\tilde{\lambda}_{j}^{(M)}$.

Or, quel est le taux de mortalité pour $\tilde{Z}_{2N}^{(A)}(\tau)$ étant donné une absorption à 0 ? Conformément au résultat (102), celui-ci est égal à

$$\tilde{\mu}_{2N-j}^{(A)*} = \tilde{\mu}_{2N-j}^{(A)} \frac{\pi_{2N-j-1}^{(A)}}{\pi_{2N-j}^{(A)}}, \tag{103}$$

où $\pi_{2N-j}^{(A)}$ et $\pi_{2N-j}^{(A)*}$ sont définis comme dans l'équation (99).

Maintenant, il faut déterminer la valeur de $\pi_{2N-j-1}^{(A)}$. Nous pouvons déduire que

$$\pi_{2N-j-1}^{(A)} = \frac{1-\alpha^{2N-(j+1)}}{1-\alpha^{2N}}, \tag{104}$$

où

$$\alpha = \frac{\tilde{\mu}_{2N-j}^{(A)}}{\tilde{\lambda}_{2N-j}^{(A)}}. \tag{105}$$

Notons que cette formule est analogue à celle développée dans le contexte du problème de la ruine du joueur. En effet, il est facile de prouver que si un joueur possède une fortune de a, s'il arrête de jouer quand sa fortune atteint $2N$ ou 0 et s'il a, à chaque ronde de jeu, r comme probabilité de victoire (équivalent à un gain de 1), $1-r$ comme probabilité de défaite (équivalent à une perte de 1), alors

$$P(\{\text{Fortune atteigne } 2N \text{ avant d'atteindre } 0 | \text{Fortune de } a \text{ initialement}\}) = \frac{1-\gamma^a}{1-\gamma^{2N}}, \tag{106}$$

où $\gamma = \frac{1-r}{r}$. Si on utilise ce concept dans le cadre du modèle dont il est question, on déduit que

$$\gamma = \frac{\frac{\tilde{\mu}^{(A)}_{2N-j}}{\tilde{\lambda}^{(A)}_{2N-j}+\tilde{\mu}^{(A)}_{2N-j}}}{\frac{\tilde{\lambda}^{(A)}_{2N-j}}{\tilde{\mu}^{(A)}_{2N-j}+\tilde{\lambda}^{(A)}_{2N-j}}} = \frac{\tilde{\mu}^{(A)}_{2N-j}}{\tilde{\lambda}^{(A)}_{2N-j}} = \alpha. \tag{107}$$

Ainsi, en substituant γ dans l'équation (106), on arrive à (104).

Maintenant, en remplaçant $\tilde{\mu}^{(A)}_{2N-j}$ par $\tilde{\lambda}^{(M)}_{j}$ et $\tilde{\mu}^{(A)*}_{2N-j}$ par $\tilde{\lambda}^{(M)*}_{j}$ dans (103), on arrive à

$$\tilde{\lambda}^{(M)*}_{j} = \tilde{\lambda}^{(M)}_{j} \frac{\frac{1-\alpha^{2N-(j+1)}}{1-\alpha^{2N}}}{\frac{1-\alpha^{2N-j}}{1-\alpha^{2N}}}. \tag{108}$$

Enfin, si on réalise que, dans le modèle dont il est question,

$$\alpha = \frac{\tilde{\mu}^{(A)}_{2N-j}}{\tilde{\lambda}^{(A)}_{2N-j}} = \frac{1+\frac{s_{2N}(j)}{2}}{1-\frac{s_{2N}(j)}{2}} = \frac{1+\frac{\beta}{4N}}{1-\frac{\beta}{4N}}, \tag{109}$$

on arrive à

$$\tilde{\lambda}^{(M)*}_{j} = \tilde{\lambda}^{(M)}_{j} \frac{u_{j+1}}{u_j}, \tag{110}$$

ce qui correspond précisément à (95).

Une démarche similaire peut être employée pour prouver l'exactitude de (96).

3.3. Deuxième partie : Calcul de la vraisemblance de l'échantillon

Après avoir généré un échantillon avec les paramètres appropriés, il faudra développer une méthode qui aura comme but de retrouver la distance entre les loci périphériques et la mutation sous sélection. Dans le cas qui nous occupe, le taux de mutation, le taux de sélection, la taille effective de la population, le taux de recombinaison sur toute la séquence ainsi que la fréquence de l'allèle sous sélection dans la population à tout moment seront assumés connus. De nombreuses méthodes d'estimation paramétrique ont été développées en génétique des populations. Toutefois, celle proposée d'abord par Griffiths et Tavaré [**6**], bien qu'elle se concentre sur l'estimation d'un taux de mutation, revêt une importance particulière, puisque c'est elle qui a tout d'abord inclus l'utilisation de chaînes de Markov dans le calcul de la vraisemblance.

3.3.1. Approche de Griffiths et Tavaré

En 1994, Griffiths et Tavaré [**6**] avaient déjà pensé à une méthode de vraisemblance maximale ayant pour but la détermination de la valeur du paramètre de mutation à l'origine d'un échantillon de segments chromosomiques. Plutôt que de tenter d'estimer la vraisemblance associée à une statistique descriptive par rapport au paramètre de mutation θ, elle cherche plutôt à calculer la vraisemblance de l'échantillon lui-même en fonction de ce paramètre. Voilà pourquoi on la classifie comme une méthode de vraisemblance complète. Plus précisément, à l'aide du modèle coalescent, les auteurs reconstruisent un grand nombre de généalogies possibles d'un échantillon prélevé dans une population neutre dans laquelle il n'y a pas de recombinaison. Ils calculent la probabilité associée à chacune de ces généalogies et, à grâce à la méthode de Monte Carlo, ils obtiennent par la suite une estimation de la fonction de vraisemblance pour le paramètre de mutation.

Il est aussi important de mentionner que les auteurs se concentrent sur le modèle à une infinité de sites. En essence, le modèle à une infinité de sites [**25**] est fondé sur l'hypothèse simplificatrice selon laquelle les mutations sont tellement rares et les sites tellement nombreux qu'on peut considérer que les mutations se produisent toujours à des sites différents. Puisqu'un chromosome peut comporter des milliards de bases azotées, cette hypothèse est raisonnable. S'il y a un événement de mutation, sa position, aléatoire, est choisie uniformément sur le segment.

Ainsi, un tel événement crée invariablement un nouveau type[7] de séquence, ce qui a pour conséquence d'empêcher un locus mutant de retrouver sa configuration non-mutante. En effet, cela impliquerait un nouvel événement de mutation au même endroit. Or, puisque la position de la mutation est déterminée aléatoirement à partir d'une distribution uniforme, il est impossible qu'un deuxième événement de mutation vienne annuler le premier. En d'autres mots, chaque mutation n'a qu'un seul et unique ancêtre.

À la base de leur modèle se trouve une formule permettant de trouver la vraisemblance d'un arbre génétique bâti à partir d'un échantillon ordonné, celle-ci étant dénotée $p(T, \mathbf{n})$. Ici, T indique les types de séquences présents dans l'échantillon et $\mathbf{n}$, le nombre de séquences de chacun de ces types dans l'échantillon. Plus précisément, $T = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d\}$, où $\mathbf{x}_i = (x_{i_0}, x_{i_1}, ...)$, $i = 1, 2, ...d$, est un vecteur indiquant le numéro des sites où il y a eu mutation du plus récent au plus ancien, e.g. $\mathbf{x}_1 = (5, 1, 2, 4, 3)$, et d correspond au nombre total de types de séquences dans l'échantillon. La formule se lit comme suit :

$$\begin{aligned} n(n-1+\theta)p(T,\mathbf{n}) &= \sum_{i:n_i\geq 2} n_i(n_i-1)p(T,\mathbf{n}-\mathbf{e}_i) \\ &+ \theta \sum_{\substack{i:n_i=1, x_{i_0} \ unique, \\ S\mathbf{x}_i \neq \mathbf{x}_j \forall j}} p(S_iT,\mathbf{n}) \\ &+ \theta \sum_{\substack{i.n_i=1 \\ x_{i_0} \ unique}} \sum_{j:S\mathbf{x}_i=\mathbf{x}_j} p(R_iT, R_i(\mathbf{n}+\mathbf{e}_j)), \end{aligned} \tag{111}$$

où :

- $\mathbf{x}_i$ désigne le i^e type de séquence de l'échantillon ;
- $\mathbf{e}_i$ est le vecteur avec 1 à la i^e position et 0 ailleurs, e.g. $\mathbf{e}_3 = \{0, 0, 1, 0, 0, ..., 0\}$;
- S est un opérateur qui efface la première coordonnée d'une séquence ;
- S_iT efface la première coordonnée de la i^e séquence de T[8] ;
- R_iT efface la i^e séquence de T ;
- «x_{i_0} unique» signifie que seulement une séquence dans l'échantillon comporte encore un allèle mutant au site x_{i_0}.

Chaque expression de la partie droite de l'équation (111) correspond à un événement en particulier. La première somme correspond aux événements de coalescence entre deux séquences de

[7] Le type d'une séquence est déterminé par les allèles qu'elle comporte. Ainsi, un nouveau type correspond simplement à une configuration allélique jamais vue jusqu'à maintenant.

[8] Puisque les sites mutants dans $\mathbf{x}_i$ sont mentionnés en ordre de disparition, quand un événement de mutation se produira, il fera disparaître la première mutation de la séquence $\mathbf{x}_i$.

type i, tandis que la deuxième désigne des événements de mutation affectant la seule séquence de type i et résultant en une séquence d'un type encore non-représenté dans l'échantillon. La troisième partie correspond aussi à des événements de mutation affectant la seule séquence de type i, mais dont la séquence résultante est d'un type j déjà présent dans l'échantillon. Puisqu'on travaille dans le cadre du modèle à une infinité de sites et puisque le temps est considéré de manière rétrospective, il est essentiel que l'allèle mutant soit unique dans l'échantillon avant de muter et de revenir à sa forme primitive ou sinon, on obtiendrait une contradiction avec l'hypothèse d'unicité de l'ancêtre de la mutation, d'où l'ajout de la condition «x_{i_0} unique».

La validité de (111) découle des propriétés des chaînes de Markov. Une simple décomposition nous indique que

$$p(T_i, \mathbf{n}_i) = \sum_y q_{xy} p(T_{i+1}, \mathbf{n}_{i+1}), \tag{112}$$

l'état x étant décrit par $(T_i, \mathbf{n}_i)$ (i étant un indicateur de temps dont la mention est omise dans (111)) et l'état y par $(T_{i+1}, \mathbf{n}_{i+1})$. Ici, q_{xy} désigne la probabilité de transition de l'état x à l'état y. Remarquons que y désigne ici uniquement les états compatibles avec x, i.e. qui peuvent être obtenus après un seul événement de coalescence ou de mutation, et que $x \neq y$. Après avoir divisé (111) par $n(n-1+\theta)$, chaque facteur nous donne la probabilité de l'événement de transition associé (voir paragraphe précédent), e.g. $\frac{n_k(n_k-1)}{n(n-1+\theta)}$ correspond à la probabilité qu'un événement de coalescence se produise dans la famille k.

Maintenant, en se basant sur le théorème suivant, on parvient à comprendre comment on peut utiliser une chaîne de Markov, définie par la suite, pour trouver la vraisemblance de l'échantillon.

Theorème 2. [24] *Soit $\{X_k\}$, $k \geq 0$, une chaîne de Markov définie sur l'espace S et comportant la matrice de transition P. Soit A, un ensemble d'état pour lesquels le temps d'arrêt*

$$\kappa = inf\{k \geq 0 : X_k \in A\} \tag{113}$$

est fini avec probabilité 1 si l'on part de n'importe quel état $x \in T \equiv S \backslash A$. Soient $f \geq 0$, une fonction définie sur S, et

$$u_x(f) = \mathbb{E}_x \prod_{k=0}^{\kappa} f(X_k) \tag{114}$$

si $X_0 = x \in S$ *et*

$$u_x(f) = f(x) \tag{115}$$

si $x \in A$. *Alors, pour tout* $x \in T$, *on a*

$$u_x(f) = f(x) \sum_{y \in S} p_{xy} u_y(f). \tag{116}$$

Ce théorème se prouve ainsi :

$$\begin{aligned}
u_x(f) &= E_x \left[\prod_{k=0}^{\kappa} f(X_k) \right] \\
&= f(x) E_x \left[\prod_{k=1}^{\kappa} f(X_k) \right] \\
&= f(x) E_x \left[E_x \left(\prod_{k=1}^{\kappa} f(X_k) \right) \middle| X_1 \right] \\
&= f(x) E_x \left[E_{X_1} \left(\prod_{k=0}^{\kappa} f(X_k) \right) \right] \text{ (Propriété de Markov)} \\
&= f(x) E_x \left[u_{X_1}(f) \right] \\
&= f(x) \sum_{y \in S} u_y(f) p_{xy}.
\end{aligned} \tag{117}$$

Ainsi, si notre but est de résoudre une équation de la forme de (117), on peut s'y prendre en calculant $E_x \left[\prod_{k=0}^{\kappa} f(X_k) \right]$. On sait aussi que la valeur d'une espérance peut être approximée à l'aide de la méthode de Monte Carlo. Ainsi, on peut déduire que

$$E_x \left[\prod_{k=0}^{\kappa} f(X_k) \right] = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \prod_{k=0}^{\kappa} f(X_k^{(i)}). \tag{118}$$

Or, l'équation (111) a exactement la forme de (117), si l'on considère :

(1) $u_x(f) = p(T, \mathbf{n})$,

(2) $u_y(f) = p(T, \mathbf{n} - \mathbf{e}_i)$ ou $p(S_i T, \mathbf{n})$ ou $p(R_i T, R_i(\mathbf{n} + \mathbf{e}_j))$,

(3) $p_{xy} = \frac{n_i(n_i - 1)}{\chi(T, \mathbf{n}) n(n - 1 + \theta)}$ ou $\frac{\theta}{\chi(T, \mathbf{n}) n(n - 1 + \theta)}$,

(4) $f(x) = \chi(T, \mathbf{n})$,

avec $\chi(T, \mathbf{n})$ étant un facteur de normalisation prenant la valeur

$$\chi(T, \mathbf{n}) = \sum_{i=1}^{d} \frac{n_i(n_i - 1)}{n(n + \theta - 1)} + \frac{\theta m}{n(n + \theta - 1)}. \tag{119}$$

Ici, d représente le nombre de types de séquences et m, le nombre de loci comportant un seul allèle mutant dans tout l'échantillon, ce qui est la condition pour qu'il puisse y avoir mutation

à un locus donné.

En faisant toutes ces substitutions, on retrouve précisément (111). Par conséquent, on peut obtenir la vraisemblance d'un échantillon donné en générant un grand nombre de réalisation de la chaîne de Markov $\{X_k\}$ et en calculant la valeur de (118). Le temps d'arrêt κ est donné ici par le temps d'atteinte du premier ancêtre commun.

L'équation (111) nous a permis de déduire les valeurs à donner aux probabilités de transition de la chaîne de Markov :

- $(T, \mathbf{n}) \rightarrow (T, \mathbf{n} - \mathbf{e}_i)$ avec probabilité $\frac{n_i - 1}{\chi(T,\mathbf{n})(n+\theta-1)}$
- $(T, \mathbf{n}) \rightarrow (S_i T, \mathbf{n})$ avec probabilité $\frac{\theta}{\chi(T,\mathbf{n})n(n+\theta-1)}$
- $(T, \mathbf{n}) \rightarrow (R_i T, R_i(\mathbf{n} + \mathbf{e}_j))$ avec probabilité $\frac{\theta}{\chi(T,\mathbf{n})n(n+\theta-1)}$

Bien que Griffiths et Tavaré [**6**] se soient concentrés sur un modèle ne comportant pas de recombinaison (leur but étant essentiellement de déterminer la vraisemblance d'un échantillon pour le paramètre de mutation θ), rien ne nous empêche d'y inclure ce phénomène. Griffiths et Marjoram [**5**] l'ont fait dans leur article en introduisant le graphe de recombinaison ancestral. L'équation de récurrence sera différente de (111) et, par conséquent, la chaîne de Markov à considérer sera aussi affectée, mais la méthode de base de détermination de la vraisemblance restera la même. L'équation de récurrence adaptée au phénomène de recombinaison et de sélection est décrite à la section 3.3.3.

3.3.2. Une méthode de vraisemblance maximale adaptée à la recombinaison et à la sélection

La vraisemblance par rapport au paramètre ρ_0, soit la distance calculée en unités du taux de recombinaison entre le locus sous sélection et le premier locus retenu se trouvant à sa gauche, sera calculée grâce aux méthodes de Monte Carlo. Tel que mentionné précédemment, cette unité de mesure a été retenue pour la simple raison qu'en absence d'interférence (ce qui est assumé ici), il existe une relation directe entre le taux de recombinaison entre deux sites et la distance physique les séparant. Comme dans les travaux de Griffiths [**5**], une équation de récurrence exprimant la vraisemblance d'un échantillon après un événement quelconque en fonction de sa

vraisemblance avant cet événement sera créée. Celle-ci sera par la suite utilisée pour élaborer un système d'équations fixant les probabilités de transition de l'échantillon.

3.3.3. L'équation de récurrence : Dérivation et interprétation

Tout d'abord, on sélectionne séparément n_C cas et n_T témoins dans la population. L'échantillonnage se fait de manière ordonnée. Pour calculer la probabilité d'obtenir un certain échantillon ordonné, on doit tout d'abord déterminer le taux associé à chaque type d'événement possible. La quantité $p_F(t)$ correspond à la fréquence dans la population des allèles de type F, F pouvant être égal à C (pour «cas») et T (pour «témoin»). Quand la sélection sera assumée stabilisatrice, $p_F(t)$ sera constant pour $t \geq 0$. Cependant, quand la sélection sera assumée génique, cette fonction verra sa valeur varier avec t avant d'atteindre éventuellement 0 si $F = C$ et 1 si $F = T$.

- Le taux associé à un événement de coalescence entre deux séquences ancestrales de type i de famille F est de $\frac{n_i(n_i-1)}{2}\frac{1}{p_F(t)}$.
- Le taux associé à un événement de coalescence entre deux séquences ancestrales compatibles de types i et j respectivement et de famille F est de $\frac{n_i n_j}{p_F(t)}$.
- Un événement de recombinaison coupera une séquence de type i ancestrale entre les loci m et $m+1$ avec taux $\frac{n_i \rho_m}{2} I\{m \in B_{(i)}\}$, où $B_{(i)}$ désigne l'ensemble des intervalles d'une séquence de type i compris entre deux loci ancestraux.
- Un événement de mutation affectera une séquence de type i ancestrale au locus m dans $A_{(i)}$, $A_{(i)}$ étant l'ensemble des loci ancestraux d'une séquence de type i, avec taux $\frac{\theta_m}{2} I\{\text{Nombre de séquences mutantes au locus ancestral } m = 1\}$.

On désigne par $q(H_\tau)$ la vraisemblance d'un échantillon ordonné H_τ au temps τ, le temps, discret, étant ici exprimé en nombre d'événements. Cette valeur de vraisemblance sera ajustée après chaque événement, qui se verra attribuer un code :

- C_i désigne un événement de coalescence impliquant deux séquences ancestrales de type i.
- C_{ij}^k désigne un événement de coalescence impliquant une séquence ancestrale de type i et une séquence ancestrale de type j, $i \neq j$, les deux séquences étant compatibles. La séquence résultante est de type k.

- $R_i^{jk}(m)$ désigne un événement de recombinaison affectant une séquence de type i ancestrale entre les loci m et $m+1$, j et k étant les types des séquences résultantes.
- $M_i^j(m)$ désigne un événement de mutation affectant l'unique séquence de type i mutante au locus m. La séquence résultante est de type j.

Notons, entre autres, que pour qu'un événement de recombinaison soit défini sans équivoque, les valeurs de i et k doivent être spécifiées simultanément. En effet, si l'on assume que k désigne toujours le type de la séquence non-ancestrale au locus sous sélection, sa mention nous permettra de déduire m. Le type j pourra être déterminé en combinant l'information de i et de k.

Ainsi, pour un échantillon ordonné, comme l'avaient fait Griffiths et Tavaré [**6**], nous pouvons exprimer la vraisemblance après τ événements par une formule de récurrence. Elle prendra la forme suivante :

$$\begin{aligned}
\eta q(H_\tau) &= \sum_F \sum_{i \in F} \binom{n_i}{2} \frac{1}{p_F(t)} q(H_\tau + C_i) \\
&+ \sum_{\substack{i,j \in F \\ i,j \text{ comp.} \\ i<j}} \frac{2n_i n_j}{p_F(t)} q(H_\tau + C_{ij}^k) \\
&+ \sum_F \sum_{i \in F} \sum_{m \in B_{(i)}} \frac{n_i \rho_m}{2} q(H_\tau + R_i^{jk}(m)) \\
&+ \sum_i \sum_{\substack{m \in A_{(i)}, \\ \textit{Mutant unique au locus } m}} \frac{\theta_m}{2} q(H_\tau + M_i^j(m)),
\end{aligned} \tag{120}$$

avec

$$\eta = \frac{n_T(n_T-1)}{2p_T(t)} + \frac{n_C(n_C-1)}{2p_C(t)} + \frac{n\rho\beta}{2} + \frac{n\alpha\theta}{2}, \tag{121}$$

où

$$\alpha = \sum_{i=1}^{s} \frac{n_i}{n} \sum_{j=1}^{b} I\{\text{Locus } j \text{ est ancestral}\} \frac{\theta_j}{\theta} \tag{122}$$

et

$$\beta = \sum_{i=1}^{s} \frac{n_i}{n} \sum_{j=1}^{b-1} I\{\text{Intervalle } j \text{ de la séq. de type } i \text{ est compris entre deux loci ancestraux}\} \frac{\rho_j}{\rho}, \tag{123}$$

s étant encore le nombre total de types de séquences et b, le nombre total de loci dans chaque séquence.

La validité de (120) repose sur les même bases que celle de (111). Sa formulation lui est d'ailleurs analogue. La logique utilisée pour dériver les deux équations est la même.

Pour obtenir la vraisemblance d'un échantillon non-ordonné, il ne suffit que d'appliquer les relations suivantes :

- $Q(H_\tau) = \frac{n_C! n_T!}{n_1! n_2! ... n_s!} q(H_\tau)$,
- $Q(H_\tau + C_i) = \frac{(n_F - 1)! n_{F'}!}{n_1! n_2! ... (n_i - 1)! ... n_s!} q(H_\tau + C_i)$, $i \in F$,
- $Q(H_\tau + C_{ij}^k) = \frac{(n_F - 1)! n_{F'}!}{n_1! n_2! ... (n_i - 1 + \delta_{ik})! (n_j - 1 + \delta_{jk})!} \times \frac{1}{((n_k + 1)!(1 - \delta_{ik} - \delta_{jk}) + \delta_{ik} + \delta_{jk}) ... n_s!} q(H_\tau + C_{ij}^k)$, $k \in F$,
- $Q(H_\tau + R_i^{jk}(m)) = \frac{(n_F + 1)! n_{F'}!}{n_1! n_2! ... (n_i - 1 + \delta_{ik})! (n_j + 1)! ((n_k + 1)!(1 - \delta_{ik}) + \delta_{ik}) ... n_s!} \times q(H_\tau + R_i^{jk}(m))$, $k \in F$,
- $Q(H_\tau + M_i^j(m)) = \frac{n_C! n_T!}{n_1! n_2! ... (n_k + 1)! ... n_s!}$,

où F' désigne l'autre famille, (i.e. si $F = C$, alors $F' = T$) s, le nombre total de familles de séquences et δ_{ij} prend la valeur 1 si i est égal à j et 0 sinon. Ainsi, $i \in C$ signifie que la séquence est dans la famille des cas tandis que $i \in T$ signifie qu'elle appartient à la famille des témoins.

Le coefficient de la partie droite du premier item de la liste précédente se comprend facilement : il correspond simplement au nombre de façons d'ordonner les séquences d'un échantillon séparé en deux familles (cas et témoins) et comportant en tout s types de séquences.

Le coefficient de la partie droite du deuxième item correspond au nombre de façons d'ordonner les séquences d'un même échantillon après un événement de coalescence entre deux séquences de type i. Cela résultera en une diminution de 1 de la valeur de n_i (désignant encore le nombre de séquences de type i juste avant l'événement) et donc en une diminution de 1 de n_F, $i \in F$.

Le coefficient de la partie droite du troisième item nous donne le nombre de façons d'ordonner l'échantillon après un événement de coalescence impliquant une séquence de type i et une séquence de type j. Après l'événement, il y aura une séquence de moins de type i et de type j, à condition que $i \neq k$ et que $j \neq k$. Voilà pourquoi on retrouve au dénominateur les indicateurs δ_{ik} et δ_{jk}. Or, nonobstant cela, un événement de coalescence résulte toujours en l'élimination

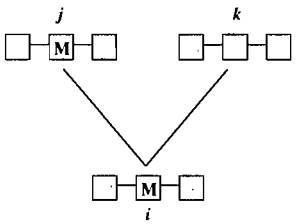

FIG. 12. Un événement de recombinaison. M indique l'allèle au locus sous sélection de la séquence i et de la séquence j. Par convention, la séquence-parent j transmettra toujours à la séquence i l'allèle au locus sous sélection et la séquence k aura toujours une configuration inconnue au locus sous sélection. Malgré tout, elle devra se voir attribuer une famille.

d'une séquence, d'où le facteur $(n_F - 1)!$.

Le coefficient de la partie droite du quatrième item correspond au nombre de façons d'ordonner l'échantillon de séquences après un événement de recombinaison. Puisqu'on doit obligatoirement attribuer aux deux séquences résultantes une famille, i.e. une configuration au locus sous sélection, il est possible que la séquence de type k, ne comportant pas l'allèle sous sélection des séquences de type i (voir illustration (12)), soit de type i. Dans un tel cas, le nombre de séquences de type $i = k$ ne changera pas, d'où l'expression $((n_k + 1)!(1 - \delta_{ik}) + \delta_{ik})$. Or, il y aura tout de même une séquence de plus dans la famille à laquelle appartient k, ce qui explique le facteur $(n_F + 1)!$.

Enfin, un événement de mutation éliminera une séquence de type i et en créera une nouvelle de type j. Voilà pourquoi l'expression dénotant le nombre de façons de réorganiser l'échantillon comporte au dénominateur $(n_j + 1)!$. Elle ne comporte pas de $(n_i - 1)!$ tout simplement parce que pour qu'il puisse y avoir mutation, n_i doit avoir 1 comme valeur. Ainsi, nous pouvons déduire que $(n_i - 1)! = 1$.

L'équation de récurrence pour un échantillon non-ordonné prendra donc la forme suivante :

$$
\begin{aligned}
Q(H_\tau) &= \sum_F P(C_F)\Big(\sum_{i\in F}\frac{(n_i-1)}{n_F-1}Q(H_\tau+C_i)\\
&+ 2\sum_{\substack{i,j\in F\\ i,j\ comp.\\ i<j}}\frac{(n_k+1-\delta_{ik}-\delta_{jk})}{n_F-1}Q(H_\tau+C_{ij}^k)\Big)\\
&+ \sum_F\sum_{i\in F}\sum_{m\in B_{(i)}}P(R_F)\frac{(n_j+1)(n_k+1-\delta_{ik})}{n(n_F+1)}\frac{\rho_m}{\beta\rho}Q(H_\tau+R_i^{jk}(m))\\
&+ P(M)\sum_i\sum_{\substack{m\in A_{(i)},\\ \textit{Mutant unique au locus } m}}\frac{n_j+1}{n}\frac{\theta_m}{\theta\alpha}Q(H_\tau+M_i^j(m)),
\end{aligned}
\tag{124}
$$

avec

$$
\begin{aligned}
P(C_F) &= \frac{\frac{n_F(n_F-1)}{p_F(t)}}{\frac{n_T(n_T-1)}{p_T(t)}+\frac{n_C(n_C-1)}{p_C(t)}+n\rho\beta+n\alpha\theta},\\
P(R_F) &= \frac{n\rho\beta p_F(t)}{\frac{n_T(n_T-1)}{p_T(t)}+\frac{n_C(n_C-1)}{p_C(t)}+n\rho\beta+n\alpha\theta},\\
P(M) &= \frac{n\alpha\theta}{\frac{n_T(n_T-1)}{p_T(t)}+\frac{n_C(n_C-1)}{p_C(t)}+n\rho\beta+n\alpha\theta}.
\end{aligned}
\tag{125}
$$

Puisque le changement de fréquence de l'allèle sous sélection dans la population n'affecte pas le nombre de façons d'ordonner les séquences, l'hypothèse de sélection génique ne change pas la forme de l'équation récursive. Seuls les facteurs $P(.)$ seront affectés. En effet, les $p_F(t)$ ne seront plus constants sur t.

3.4. Probabilités de transition et calcul de la vraisemblance[9]

La formule de récurrence nous sert de tremplin pour déterminer la valeur des taux de transition d'un état H_τ à un autre état $H_{\tau+1}$ (voir section 3.3.1) . En effet, pour déterminer le taux associé à un événement, il est suffisant de regarder le coefficient du $Q(.)$ qui y est associé. Par exemple, le taux associé à C_{ij}^k, $k\in T$, sera de $P(C_T)\dfrac{(n_k+1-\delta_{ik}-\delta_{jk})}{n_T-1}$.

[9]De grands pans de la démarche proposée ici ont été décrits dans Larribe et al. 2002 **[15]** et dans Larribe et Lessard 2008 **[14]**.

Les taux de transition résultants seront :

$$a(H_\tau, H_{\tau+1}) = \begin{cases} P(C_F)\dfrac{(n_i-1)}{n_F-1}, \text{ si } H_{\tau+1} = H_\tau + C_i, \text{ où } i \in F, \\ 2P(C_F)\dfrac{(n_k+1-\delta_{ik}-\delta_{jk})}{n_F-1}, \text{ si } H_{\tau+1} = H_\tau + C_{ij}^k, \text{ où } i,j \in F, \\ P(R_F)\dfrac{(n_j+1)(n_k+1-\delta_{ik})}{n(n_F+1)}\dfrac{\rho_m}{\beta\rho}, \text{ si } H_{\tau+1} = H_\tau + R_i^{jk}(m), \text{ où } k \in F, \\ P(M)\dfrac{n_j+1}{n}\dfrac{\theta_m}{\theta\alpha} \text{ si } H_{\tau+1} = H_\tau + M_i^j(m). \end{cases} \tag{126}$$

Afin d'utiliser ces taux pour générer une chaîne de Markov, une fonction de probabilité est requise. En appliquant un facteur de normalisation $N(H_\tau)$ à $a(H_\tau, H_{\tau+1})$, on obtient

$$P(H_{\tau+1}|H_\tau) = \frac{a(H_\tau, H_{\tau+1})}{N(H_\tau)}, \tag{127}$$

où

$$N(H_\tau) = \sum_{H_{\tau+1}} a(H_\tau, H_{\tau+1}). \tag{128}$$

Des événements seront donc générés à partir de $P(H_{\tau+1}|H_\tau)$ jusqu'à ce que chaque locus considéré, à l'exception du locus sous sélection dans le cas où il y a sélection stabilisatrice[10], ait trouvé un ancêtre commun.

On peut maintenant reformuler l'équation de récurrence (124) :

$$Q(H_\tau) = \sum_{H_{\tau+1}} N(H_\tau)P(H_{\tau+1}|H_\tau)Q(H_{\tau+1}). \tag{129}$$

La vraisemblance sera calculée à l'aide des méthodes de Monte Carlo accélérée par la technique d'*importance sampling*. L'utilisation de la méthode d'*importance sampling* pour déterminer la vraisemblance d'un échantillon de séquences génétiques a été proposée par Griffiths et Tavaré **[6]**. Supposons que l'on souhaite déterminer la vraisemblance par rapport au paramètre β, i.e. $L(\beta) = P(D|\beta)$, où D correspond à l'échantillon considéré. On peut exprimer $L(\beta)$ comme suit :

$$L(\beta) = \int_H P_\beta(D|H)P_\beta(H)dH, \tag{130}$$

où l'intégrale est par rapport à toute l'histoire ancestrale de D, représentée par H. Dans le cadre du modèle coalescent, on dira que H correspond à la topologie du graphe ainsi qu'aux mutations qui y sont associées. À partir des hypothèses du modèle coalescent, $P_\beta(H)$ peut être

[10]La mutation au locus sous sélection étant ancienne, il y aura toujours polymorphisme à ce locus, ce qui empêche donc l'atteinte d'un ancêtre commun.

calculé. L'espace de H étant infini, l'approche de Monte Carlo est utilisée pour estimer la valeur de l'intégrale. On obtient

$$L(\beta) \approx \frac{1}{M} \sum_{i=1}^{M} P_\beta(D|H^{(i)}), \tag{131}$$

où les $H^{(i)}$, $i = 1, 2, ..., M$, sont des histoires ancestrales échantillonnées à partir de $P_\beta(H)$. De plus, remarquons que $P_\beta(D|H)$ (et donc $P_\beta(D|H^{(i)})$) ne peut prendre que 0 ou 1 comme valeur. En effet, H détermine précisément la configuration de D. Par conséquent, si H est compatible avec D, la fonction prendra la valeur 1 ou sinon, elle prendra la valeur 0. Malheureusement, un très grand nombre de généalogies sont possibles et très peu d'entre elles sont compatibles avec D. Ainsi, il faudrait générer un très grand nombre d'histoires ancestrales afin d'obtenir une estimation précise de $L(\beta)$. On peut améliorer la méthode en question en se concentrant sur les généalogies compatibles avec D. Ceci revient à appliquer le principe d'*importance sampling*. On peut réexprimer la vraisemblance comme

$$L(\beta) = \int P_\beta(D|H) \frac{P_\beta(H)}{Q_\beta(H)} Q_\beta(H) dH, \tag{132}$$

où $Q_\beta(H)$ est une distribution qu'on appellera «distribution proposée». Ce peut être n'importe quelle distribution satisfaisant la condition

$$Q_\beta(H) > 0 \Leftrightarrow P_\beta(D|H) P_\beta(H) > 0, \tag{133}$$

cette condition étant nécessaire pour éviter qu'il ne puisse y avoir division par 0.

En estimant la valeur de l'intégrale par la formule de Monte Carlo, on obtient

$$L(\beta) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P_\beta(D|H^{(i)}) P_\beta(H^{(i)})}{Q_\beta(H^{(i)})}, \tag{134}$$

où les $H^{(i)}$, $i = 1, 2, ..., M$, sont des histoires ancestrales échantillonées à partir de $Q_\beta(H)$. Il est maintenant légitime de se questionner sur la forme que devrait prendre $Q_\beta(H)$ dans les circonstances. Idéalement, cette distribution serait

$$Q^*_\beta(H) = P_\beta(H|D) = \frac{P_\beta(H, D)}{P_\beta(D)}. \tag{135}$$

Cette distribution générera seulement des arbres compatibles avec l'échantillon D, ce qui est une nette amélioration. Or, dans ce cas, les termes de la somme (134) prendront la valeur

$$\frac{P_\beta(D|H^{(i)})P_\beta(H^{(i)})}{Q^*_\beta(H^{(i)})} = \frac{P_\beta(D,H^{(i)})}{P_\beta(D,H^{(i)})}P_\beta(D) = P_\beta(D), \tag{136}$$

qui ne dépend pas de H. Par conséquent, tous les termes seront égaux et la variance de (134) sera de 0. Malheureusement, nous ne connaissons pas la distribution requise. Nous ne sommes pas en mesure de générer d'histoires ancestrales à partir de celle-ci et nous ne pouvons certainement pas calculer la valeur que $Q^*_\beta(H^{(i)})$ prendra pour un quelconque $H^{(i)}$. Il nous faudra donc approximer $Q^*_\beta(H)$. Plus notre approximation ressemblera à la distribution optimale, plus elle sera précise.

Trouver la distribution optimale est malheureusement hors de portée. Or, on peut déjà obtenir une relativement bonne approximation en se concentrant uniquement sur les distributions de H compatibles avec l'échantillon D. Rappelons-nous que $P_\beta(D|H)$ ne peut prendre que les valeurs 1 ou 0. Ainsi, si l'on utilise une telle distribution, on peut réécrire (132) de la façon suivante :

$$L(\beta) = \int \frac{P_\beta(H)}{Q_\beta(H)} Q_\beta(H) dH. \tag{137}$$

Par conséquent, (134) prendra la valeur

$$L(\beta) \approx \frac{1}{M}\sum_{i=1}^{M} \frac{P_\beta(H^{(i)})}{Q_\beta(H^{(i)})}, \tag{138}$$

$H^{(i)}$ étant une réalisation de H.

Un bon choix pour $Q_\beta(H)$ serait la distribution d'une chaîne de Markov en temps inversé ayant comme état initial l'échantillon D. Considérons maintenant $H^{(i)}_j$, défini comme la configuration ancestrale après j transitions d'une histoire ancestrale i. Le rapport $\frac{P_\beta(H)}{Q_\beta(H)}$ est appelé *importance sampling weight*. Puisque la séquence $\{H^{(i)}_a\}$, $a = 1, 2, ..., m$, forme une chaîne de Markov, celui-ci peut être exprimé ainsi :

$$\frac{P_\beta(H^{(i)})}{Q_\beta(H^{(i)})} = P(H_m)\prod_{j=1}^{m} \frac{P_\beta(H^{(i)}_{j-1}|H^{(i)}_j)}{Q_\beta(H^{(i)}_j|H^{(i)}_{j-1})}, \tag{139}$$

où $H_0^{(i)}$ correspond à D, P correspond aux probabilités de transition de $H_j^{(i)}$, Q constitue les probabilités de transition en temps inversé de cette chaîne et $P(H_m)$ correspond à la distribution du type de l'ancêtre ultime [**6**].

La méthode présente malgré tout certaines faiblesses. Nous savons que même la fonction optimale $Q^*_\beta(H)$ dépend du paramètre β. Ainsi, il est normal d'assumer que Q^* ne puisse pas être optimal pour tous les β. En fait, l'efficacité de l'estimateur résultant risque de varier avec β. Nous nous verrons obligés d'utiliser une valeur de base, β_0, appelée «valeur proposée». On peut s'attendre à ce que la fonction d'*importance sampling* soit plus efficace si la valeur proposée est proche de la véritable valeur de β et de moins en moins efficace à mesure que l'on s'en éloigne [**22**]. Stephens et Donnelly [**23**] ont établi que cette implémentation de la méthode avait tendance à causer en pratique une sous-estimation de la vraisemblance pour les valeurs du paramètre loin de β_0. Ils ajoutent que cela pourrait avoir deux conséquences néfastes. D'une part, cela peut résulter en un maximum de la fonction de vraisemblance proche de β_0, même si le maximum devrait se trouver à une autre valeur. D'autre part, si la valeur véritable du paramètre se trouve près de β_0, la valeur de vraisemblance accordée à ce maximum risque d'être surestimée, ce qui pourrait rendre les intervalles de confiance autour de l'estimé trop étroits. Malheureusement, il est difficile de quantifier la magnitude des effets décrits précédemment.

Toutefois, on peut en partie remédier à ce problème en adaptant la valeur de β_0 après un certain nombre d'itérations. La méthode de Monte Carlo garantissant une convergence asymptotique, on peut espérer que la position estimée du maximum converge progressivement vers sa véritable valeur. On devrait théoriquement observer un ralentissement dans la variation de la position estimée du maximum après chaque ronde d'itérations à mesure que β_0 se rapproche de sa vraie position. Il n'existe pas de règles pour déterminer combien d'itérations seront nécessaires pour pouvoir profiter de cette progression. Il est donc conseillé d'en faire jusqu'à ce qu'on puisse clairement certifier qu'il y a bel et bien convergence.

Une autre approche, qu'on appelle par point («pointwise»), s'offre à nous. On peut tenter d'estimer la fonction de vraisemblance point par point en utilisant comme valeur proposée pour le paramètre β le point auquel on tente de calculer la vraisemblance. Autrement dit, si on cherche à calculer $L(\beta_i)$, on utilisera $Q_{\beta_i}(H)$ comme distribution proposée. Cette approche est toutefois plutôt inefficace. En effet, elle nécessite des échantillons générés à partir de chaque

distribution $Q_{\beta_i}(H)$ et chaque échantillon ne sert qu'à déterminer la valeur de vraisemblance à un point. Cette technique peut être améliorée. En utilisant une mixture de distributions, on peut assurément obtenir plus efficacement la courbe de vraisemblance. Dans cette situation, la distribution proposée prendra la forme

$$Q(H) = \frac{1}{R}\sum_{i=1}^{R} Q_{\beta_i}(H). \tag{140}$$

Stephens [**22**] affirme qu'utiliser (140) sera assurément plus efficace qu'utiliser l'approche par point décrite précédemment.

Dans le cas qui nous occupe, la distribution proposée sera donnée par (127), mais avec un paramètre ρ_m fixé à l'avance. Cette distance proposée, désignée par ρ_0, devra être choisie de façon à ce qu'elle soit le plus proche possible de la distance véritable. La méthode sélectionnée assure une convergence pour $0 < \rho_0 < \rho$. Cependant, choisir un ρ_0 très éloigné de la distance véritable aura comme impact d'augmenter significativement le nombre d'itérations nécessaires pour avoir un début de convergence. En fait, afin de fixer une valeur de départ raisonnable pour ρ_0, il est généralement recommandé de commencer avec une distance proposée de $\frac{\rho}{2}$, de réaliser quelques simulations successives plus courtes et d'utiliser la valeur associée au maximum de vraisemblance de la courbe produite après chaque ronde de simulations comme distance proposée pour la ronde suivante.

À l'aide de la formulation donnée à l'équation (129), on peut trouver une expression pour la valeur qui nous intéresse :

$$\begin{aligned}
Q(H_0) &= \sum_{H_1} N(H_0)P(H_1|H_0)Q(H_1) \\
&= \sum_{H_1} N(H_0)P(H_1|H_0)\sum_{H_2} N(H_1)P(H_2|H_1)Q(H_2) \\
&= \sum_{H_1}\sum_{H_2} N(H_0)N(H_1)P(H_1|H_0)P(H_2|H_1)Q(H_2) \\
&= \ldots \\
&= \sum_{H_0,H_1,\ldots,H_{\tau^*}} \prod_{\tau=0}^{\tau^*-1} N(H_\tau)P(H_{\tau+1}|H_\tau)Q(H_{\tau^*}),
\end{aligned} \tag{141}$$

ce qui donne

$$Q(H_0) = E_P \left[Q(H_{\tau^*}) \prod_{\tau=0}^{\tau^*-1} N(H_\tau) \right], \tag{142}$$

où τ^* correspond au premier temps auquel on aura trouvé un ancêtre commun aux loci considérés, à l'exception du locus sous sélection. Ici, $Q(H_{\tau^*})$ prendra la valeur 1 puisqu'il est assumé qu'à τ^*, tous les loci ancestraux sont monomorphes et non-mutants, sauf le locus sous sélection qui reste assurément polymorphe quand il y a sélection stabilisatrice. Remarquons aussi que (142) découle directement du théorème 2 (voir explication suivant (118)).

En incorporant maintenant la distribution proposée à l'équation, on obtient

$$Q(H_0) = E_{P_0} \left[\prod_{\tau=0}^{\tau^*-1} N(H_\tau) \frac{P(H_{\tau+1}|H_\tau)}{P_0(H_{\tau+1}|H_\tau)} \right], \tag{143}$$

où $P_0(.)$ correspond à la distribution proposée avec paramètre de distance ρ_0.

La méthode de Monte Carlo sera utilisée pour calculer la valeur de cette espérance. Ainsi, on a

$$\hat{L}(r_T) = \frac{1}{K} \sum_{k=1}^{K} \left[\prod_{\tau=0}^{\tau^*-1} N(H_\tau^{(k)}) \frac{P(H_{\tau+1}^{(k)}|H_\tau^{(k)})}{P_0(H_{\tau+1}^{(k)}|H_\tau^{(k)})} \right], \tag{144}$$

où $H_\tau^{(k)}$ correspond à l'état de l'échantillon au temps τ pour le k^e graphe généré. Notons que la fonction $\hat{L}(r_T)$ est définie pour $0 < \rho_T < \rho$ et que les graphes seront tous générés à partir de $P_0(.)$.

Chapitre 4

DESCRIPTION DE L'ÉCHANTILLON UTILISÉ ET RÉSULTATS

4.1. Considérations préliminaires

Les échantillons utilisés quand la sélection sera assumée stabilisatrice comporteront 50 séquences de la famille des cas et 50 séquences de la famille des témoins, tous générés à partir de l'algorithme décrit dans la section 3.2.1. Quand la sélection sera assumée génique, les échantillons seront composés de 30 cas et de 30 témoins et seront générés à l'aide de l'algorithme donné à la section 3.2.2. Le nombre de séquences a été réduit pour accélérer la vitesse de simulation. De toutes façons, la dépendance existant entre les séquences de l'échantillon réduit considérablement l'impact de cette diminution [**22**]. Remarquons de nouveau que la composition en partie arbitraire de nos échantillons pourrait introduire un biais. Après tout, l'algorithme proposé nécessite un échantillon aléatoire prélevé à partir de la population générale. Rappelons-nous que cette technique d'échantillonnage avait été choisie pour pallier à la rareté de l'allèle mutant sous sélection dans la population. Or, nous souhaitons voir si une telle technique permettra encore une estimation précise de la distance entre la mutation ciblée et le marqueur immédiatement à sa gauche quand celle-ci est soumise à un effet de sélection. En d'autres mots, nous souhaitons nous assurer que l'extension de ce type d'échantillonnage au cas où il y a sélection naturelle n'introduira pas de biais dans la détermination de la position la plus probable de la mutation.

Quand il y aura sélection stabilisatrice, à moins de mention contraire, les séquences considérées auront 13 loci. Dans le cas de la sélection génique, elles en auront 25. Ces nombres ont été choisis arbitrairement afin de balancer rapidité de simulation et précision de l'estimation. Or, plusieurs courbes de vraisemblance dérivées à partir de séquences plus courtes et plus longues seront aussi présentées.

Après que les échantillons requis auront été obtenus, la génération des graphes se fera différemment selon le type de sélection considéré. Quand il y a sélection stabilisatrice, la distribution utilisée sera donnée par l'équation (127). Les valeurs de vraisemblance seront mises à jour selon le principe décrit à la section 3.4. Quand il y a sélection génique, la génération d'événements sera malheureusement quelque peu plus compliquée. En effet, le changement temporel dans la fréquence des cas dans la population a pour effet de rendre nécessaire la génération au préalable d'un temps pour l'événement. Après avoir déterminé ce temps, la fréquence des cas et des témoins au moment de l'événement étant enfin déterminée, l'équation (127) pourra être utilisée pour déterminer quel événement aura lieu. Les taux employés pour déterminer ce temps seront les suivants :

Taux de coalescence pour deux séquences de type i :

$$\frac{n_i(n_i-1)}{2p_F(t)}, \tag{145}$$

où $p_F(t)$ correspond à la fréquence dans la population au moment t de la famille F, où $i \in F$ ($i,j \in C$ ou $i,j \in T$).

Taux de coalescence pour deux séquences compatibles de type i et j :

$$\frac{n_i n_j}{p_F(t)}. \tag{146}$$

Taux de recombinaison pour un intervalle m compris dans une séquence de type i quand la séquence parentale n'ayant pas transmis le gène sous sélection est de type F :

$$\frac{n_i \rho_m p_F(t)}{2} I\{m \in B_{(i)}\}, \tag{147}$$

où $B_{(i)}$ désigne l'ensemble des intervalles d'une séquence de type i compris entre deux loci ancestraux et où $I\{.\}$ est une fonction indicatrice.

Taux de mutation à un locus m ancestral :

$$\frac{\theta_m}{2} I\{\text{Nombre de séquences mutantes au locus ancestral } m = 1\} \tag{148}$$

où $\theta_m/2$ est le taux de mutation au locus m.

L'ancêtre ultime a, cette fois, obligatoirement comme configuration 111...111 .

Dans les simulations qui suivent, θ_m prendra la valeur $0,2$, une valeur choisie volontairement élevée afin d'accélerer la vitesse de simulation, et ρ sera égal à 5. Quand viendra le temps de générer la fréquence des séquences de la famille des cas dans la population entière, N sera assumé comme étant égal à 300. La fréquence des cas dans la population à chaque instant sera assumée connue. Quand la sélection sera stabilisatrice, cette fréquence sera maintenue constante à $0,1$ tandis que quand elle sera génique, cette fréquence prendra la valeur $0,1$ au temps 0 et atteindra 0 éventuellement. Enfin, le paramètre de sélection β prendra la valeur $1,5$. Les simulations auront comme but de déterminer si par la méthode proposée, on parvient à trouver la distance entre un marqueur et la mutation sous sélection et ce, malgré l'échantillonnage arbitraire de cas et de témoins dans la population.

4.2. Résultats

4.2.1. Sélection stabilisatrice

D'une ronde de simulations à l'autre, la distance proposée a été ajustée en fonction de la valeur de ρ_0 correspondant au maximum de vraisemblance dans la ronde de simulations précédente. Chaque courbe a été créée avec $K = 10^6$ dans l'équation (144). On peut les voir aux figures (14), (15) et (16). La distance véritable est de 3.

On remarque tout d'abord le niveau des courbes. En effet, il varie grandement d'une ronde de simulations à l'autre. Par exemple, quand on considère une distance proposée de $2,5$, le sommet se trouve environ au niveau $4e^{-58}$. Dans la ronde suivante, quand la distance proposée est de $2,7$, le sommet se trouve environ au niveau $1,5e^{-52}$. N'oublions pas qu'il n'y a qu'une seule authentique fonction de vraisemblance et que quand $K \to \infty$, les deux courbes devraient être superposées. Dans notre cas, même après 10^6 graphes simulés, on ne peut observer une

convergence certaine. Par contre, la position du maximum semble tendre vers 3.

Il doit aussi être mentionné que la vraisemblance accordée à chaque graphe varie énormément. La plupart auront une vraisemblance qui sera presque partout inférieure au niveau de précision d'une variable de type «double» dans C++, soit $1,7e^{-308}$. Certains autres auront un maximum de vraisemblance surpassant e^{-46}. Or, ces graphes qu'on dira à «vraisemblance élevée» sont très rarement générés, mais ont un impact significatif sur le calcul. Ainsi, si l'on tente de modéliser la valeur de vraisemblance au sommet de la courbe en fonction du nombre d'itérations, on observera une succession de progressions lentes et irrégulières séparées par des sauts provoqués par la génération d'un de ces graphes à vraisemblance élevée. Le phénomène est illustré à la figure (13). Ainsi, on peut déduire que la convergence ne se fait pas à un rythme régulier et ce rythme peut varier grandement d'une ronde de simulations à l'autre. Chacun de ces graphes à vraisemblance élevée nous rapproche de la véritable fonction de vraisemblance. Par conséquent, un bon moyen de vérifier s'il y a convergence est de générer un grand nombre de graphes et d'arrêter la génération quand les simulations ne semblent plus en mesure de fournir des graphes affectant de façon significative la forme de la courbe. Notons que plus la distance proposée est loin de la distance véritable, plus il faudra d'itérations avant qu'un de ces graphes à vraisemblance élevée fasse son apparition. Ceci explique l'«adhérence» qu'on peut observer entre la distance maximisant la fonction de vraisemblance et la valeur de la distance proposée. Par «adhérence», on entend la tendance qu'a le maximum à rester proche de la distance proposée. En fait, plus il est loin de la distance véritable, moins le maximum aura tendance à s'en éloigner.

Cette grande variabilité suggère au moins une amélioration d'ordre algorithmique. En effet, par exemple, assumant 10^6 itérations, à partir du moment où le maximum atteint e^{-52}, il est inutile de continuer à calculer la vraisemblance de graphes dont le maximum de vraisemblance s'élève à moins de e^{-59}. Après tout, ils n'auront pas un impact significatif sur la position du maximum. Dans un tel cas, il est mieux d'accorder au graphe en cours de génération une vraisemblance de 0 pour toutes les distances entre 0 et ρ et de passer à l'itération suivante. Puisque la probabilité de génération d'un graphe à vraisemblance élevée augmente avec K, cette technique a le potentiel d'augmenter l'efficacité du programme par un facteur de plus en plus grand à mesure que la valeur de K augmente. Cela est dû au fait que la génération de graphes de plus en plus significatifs augmentera la valeur du seuil nécessaire pour qu'un graphe puisse avoir une influence sur la position du maximum. Par conséquent, de plus en plus de rondes de simulations

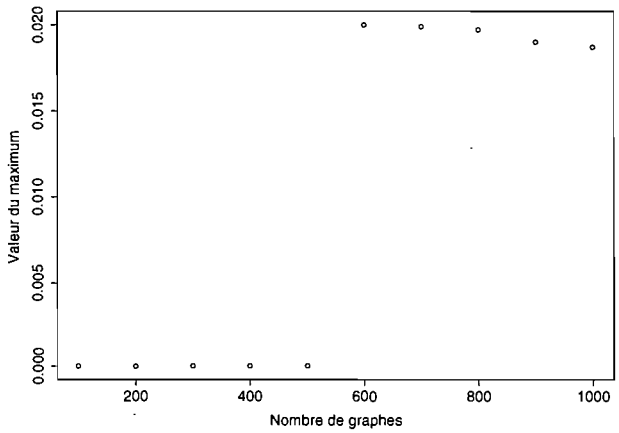

FIG. 13. Graphique illustrant la progression du niveau de vraisemblance en fonction du nombre d'itérations. Les valeurs présentées sont arbitraires.

seront interrompues de plus en plus rapidement, puisque la valeur de vraisemblance d'un graphe a tendance à diminuer avec le nombre d'événements qu'il comprend.

Remarquons que bien que l'expression «valeur de vraisemblance» soit utilisée, nous n'avons pas ici d'authentiques courbes de vraisemblance. En effet, une courbe de vraisemblance représente théoriquement une fonction de densité, ce qui n'est clairement pas le cas ici. Après tout, l'aire sous la courbe est loin d'égaler 1.

Des essais ont aussi été réalisés afin de déterminer l'effet du nombre de loci considérés sur la procédure d'estimation. Les résultats présentés aux figures (14), (15) et (16) ont été dérivés à partir d'échantillons de séquences comportant 13 loci. S'il est possible d'obtenir des résultats similaires en considérant moins de loci, afin d'augmenter la rapidité d'exécution de l'algorithme, il serait mieux d'en retrancher.

Ainsi, l'algorithme a aussi été exécuté successivement sur des échantillons de 100 séquences comportant 7 loci et 11 loci. À 7 loci, après 10^6 graphes générés, nous n'observons encore aucune véritable convergence vers 3, peu importe la distance proposée utilisée. On peut observer à la figure (17) la forme que prennent les courbes de vraisemblance générées successivement à partir de la distance proposée de $2,5$. On croit y observer convergence et stabilisation vers 2. Cependant, cette convergence ne peut être considérée comme authentique puisque quand on utilise $1,5$ comme distance proposée, on n'observe aucune tendance marquée vers 2 (voir figure

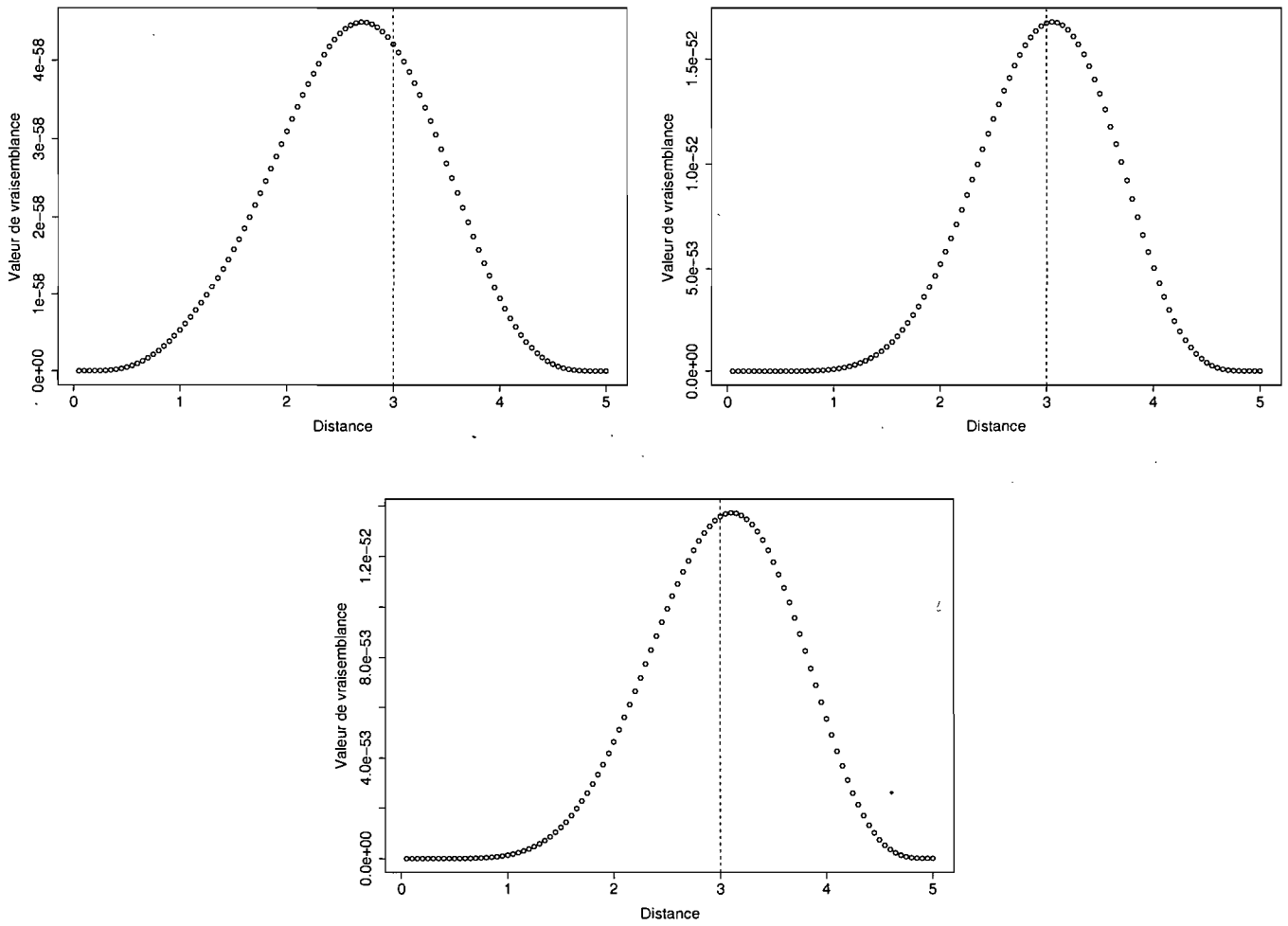

FIG. 14. Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le premier graphique a été généré à partir d'une distance proposée de 2,5, le deuxième à partir d'une distance proposée de 2,7 et le troisième à partir d'une distance proposée de 3.

(18)). À 11 loci, après 10^6 itérations, à partir d'une distance proposée de 2, on ne peut toujours pas remarquer de convergence vers la distance véritable de 3. L'algorithme rend encore 2 comme distance la plus probable (voir figure 19).

Toutefois, à 13 loci, une bonne convergence a été observée à partir du moment où la distance proposée était supérieure à 2. Or, tel que prévu, la vitesse de convergence diminue à mesure que s'accroît l'écart entre la distance proposée et la distance véritable. Quand la distance proposée est de 1, après 10^6 itérations, l'algorithme rend la distance proposée comme distance la plus commune, ce qui signifie qu'aucun graphe vraiment significatif n'a été généré. Toutefois, une

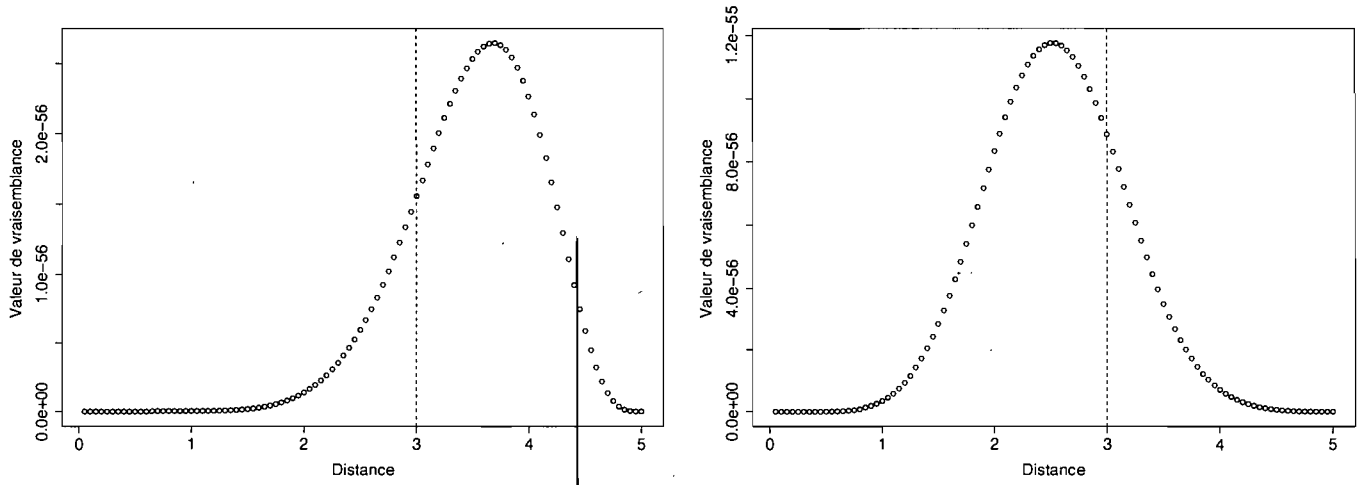

FIG. 15. Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le premier graphique a été généré à partir d'une distance proposée de 4, le deuxième à partir d'une distance proposée de 3,8. Le troisième devrait aussi être généré à partir d'une distance de 2,5. Puisqu'il est déjà affiché à l'illustration (14), il n'est pas reproduit ici.

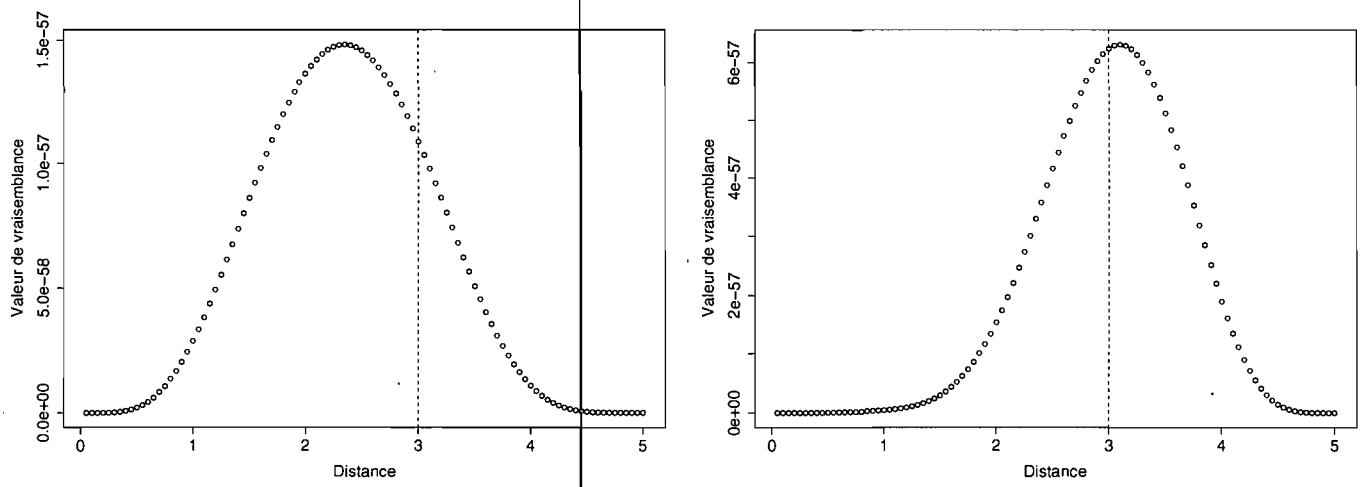

FIG. 16. Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le premier graphique a été généré à partir d'une distance proposée de 2, le deuxième à partir d'une distance proposée de 2,3. Le troisième devrait aussi être généré à partir d'une distance de 3. Puisqu'il est déjà affiché à l'illustration (14), il n'est pas reproduit ici.

augmentation du nombre de loci peut améliorer ce résultat. À 15 loci, la procédure produit encore un maximum à une distance très proche de 1 (voir courbe à gauche dans la figure (20)). À 17 loci, on remarque une nette amélioration : le maximum apparaît quand la distance se situe aux alentours de 2 (voir courbe à droite dans la figure (20)). Rendu à 19 loci, le point maximisant

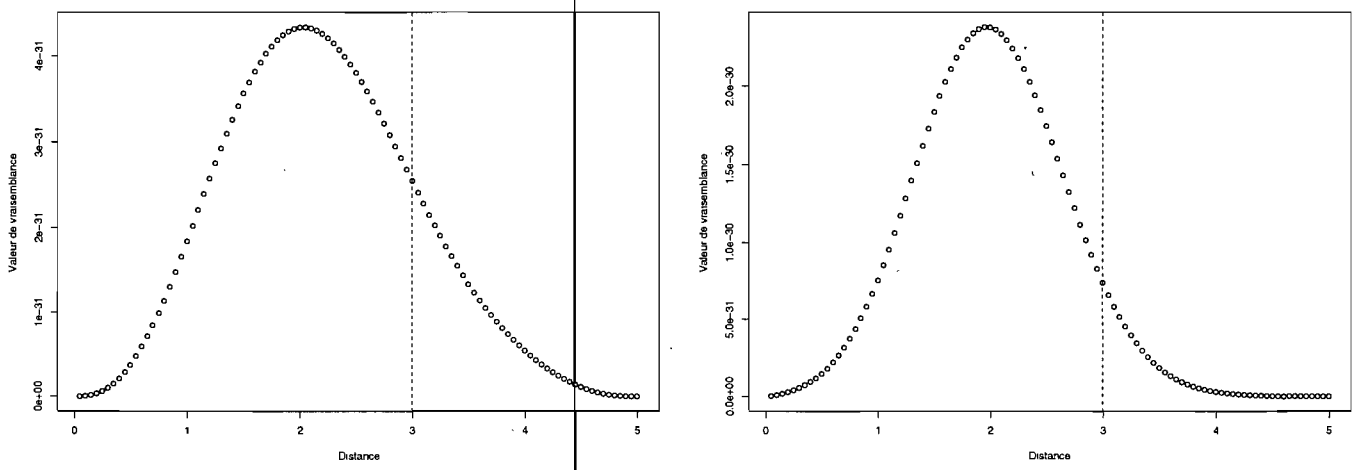

FIG. 17. Courbes de vraisemblance. Chaque courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le premier graphique a été généré à partir d'une distance proposée de 2,5, le deuxième à partir d'une distance proposée de 2. Les séquences de l'échantillon comportent 7 loci.

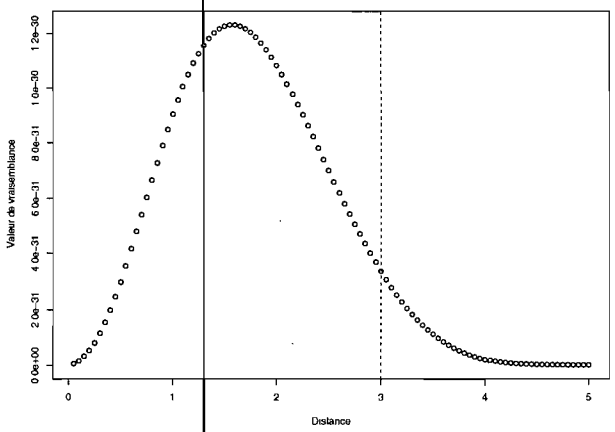

FIG. 18. Courbe de vraisemblance. La courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le graphique a été généré à partir d'une distance proposée de 1,5. Les séquences de l'échantillon comportent 7 loci.

la fonction de vraisemblance se déplace un peu vers la droite, mais reste tout de même proche de 2 (voir courbe à gauche dans la figure (21)). À 21 loci, la courbe prend une apparence bien irrégulière. Son maximum est atteint avant 1, mais son niveau est encore relativement élevé à 2 (voir courbe à droite dans la figure (21)). Un plus grand nombre d'itérations serait donc nécessaire. À 23 loci, on observe un maximum pour une distance proche de 3, mais la vraisemblance entre 1 et 2 reste non négligeable (voir courbe à gauche dans la figure (22)).

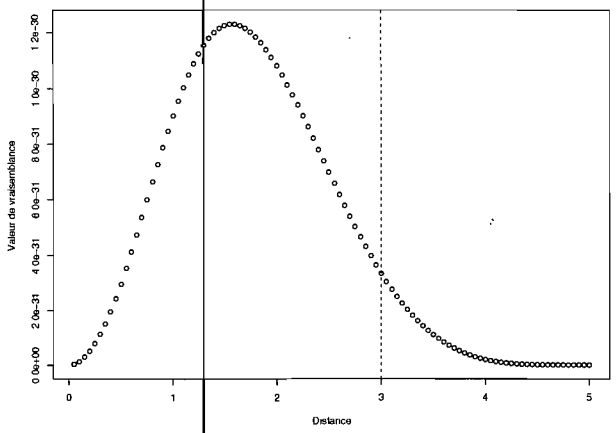

FIG. 19. Courbe de vraisemblance. La courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Le graphique a été généré à partir d'une distance proposée de 2. Les séquences de l'échantillon comportent 11 loci.

Enfin, à 25 loci, nous obtenons une courbe atteignant encore un maximum dans la région de 2 (voir courbe à droite dans la figure (22)). Cependant, elle est plus étroite que celle créée avec un échantillon de séquences à 17 loci, ce qui nous porte à croire que l'estimation qu'elle nous permet de faire est plus précise. En bout de ligne, on remarque un effet non négligeable du nombre de sites considérés sur la valeur rendue par la procédure d'estimation. L'amélioration de la vitesse de convergence n'est cependant pas garantie en pratique, ce qui est illustré amplement par la courbe obtenue quand la séquence comporte 21 loci. Malgré tout, l'augmentation du nombre de loci s'accompagne assurément d'une amélioration qu'on aurait tort d'ignorer.

4.2.2. Sélection génique

Les effets positifs de l'augmentation du nombre de loci sur le rythme de convergence, décelés dans la section précédente, de la méthode proposée nous ont incités à faire en sorte que les séquences composant l'échantillon utilisé dans cette section aient 25 loci. La distance véritable est maintenant de 4 et la distance proposée sera d'abord de $2,5$. Dans ces conditions, même après 3×10^6 itérations, l'algorithme n'indique toujours pas de convergence véritable vers 4. Si l'on part toutefois d'une distance proposée de $2,85$, après 10^6 itérations, on commence à remarquer une légère convergence vers la droite. En effet, la courbe de vraisemblance atteint maintenant un sommet à $3,05$ (voir figure 23). Ensuite, si l'on considère $3,05$ comme distance proposée, on obtient encore une légère tendance vers la droite. En effet, la courbe obtenue atteint son maximum à la distance $3,35$ (voir figure 23). Notons que la vraisemblance accordée à $3,35$ est

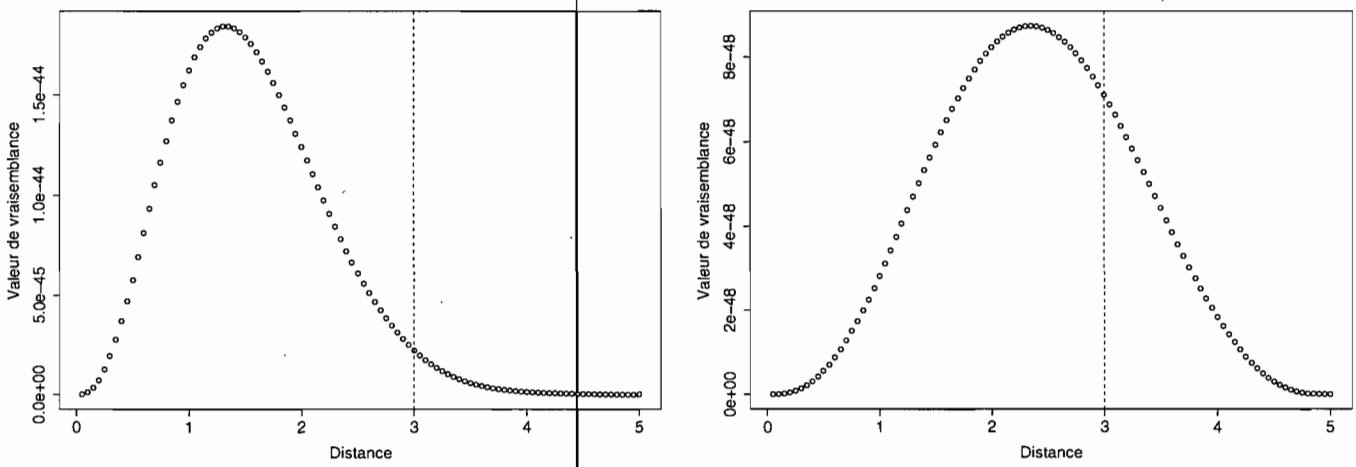

FIG. 20. Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Les deux courbes ont été générées à partir d'une distance proposée de 1. La courbe à gauche a été produite à partir d'un échantillon de séquences à 15 loci, tandis que celle à droite a été produite à partir d'un échantillon de séquences comportant 17 loci.

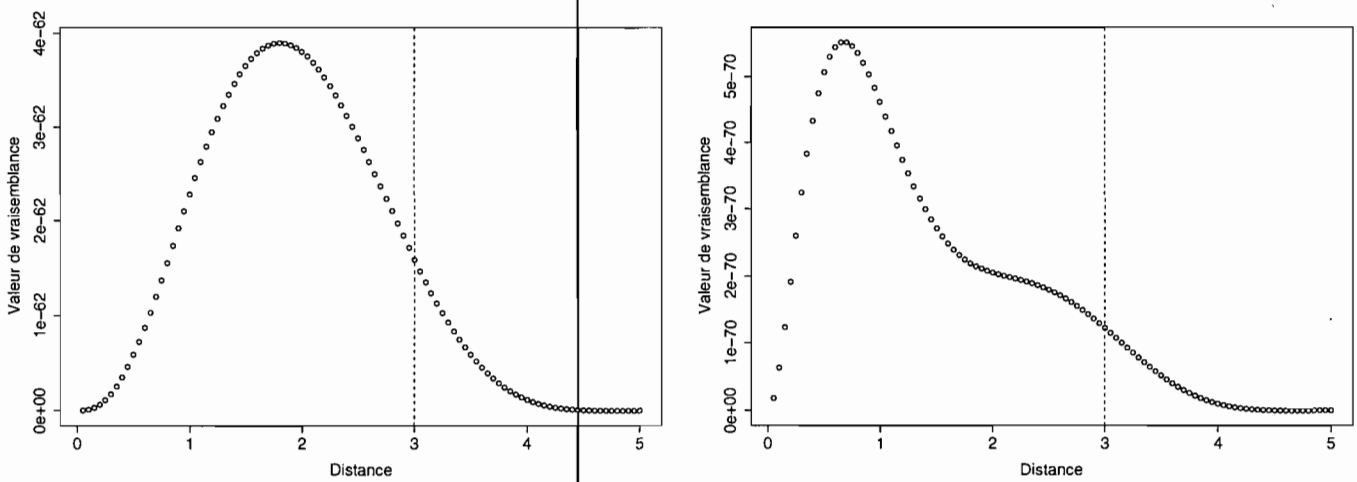

FIG. 21. Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Les deux courbes ont été générées à partir d'une distance proposée de 1. La courbe à gauche a été produite à partir d'un échantillon de séquences à 19 loci, tandis que celle à droite a été produite à partir d'un échantillon de séquences comportant 21 loci.

très proche de celle accordée à $3,05$ lors de la ronde de simulations précédente. On observe une véritable différence sur le plan des valeurs de vraisemblance à partir du moment où on utilise une distance proposée de $3,35$. En effet, maintenant, le sommet est atteint à $3,6$, mais la valeur de vraisemblance qui lui est accordée est de loin supérieure à celle donnée à $3,35$ ou $3,05$ dans

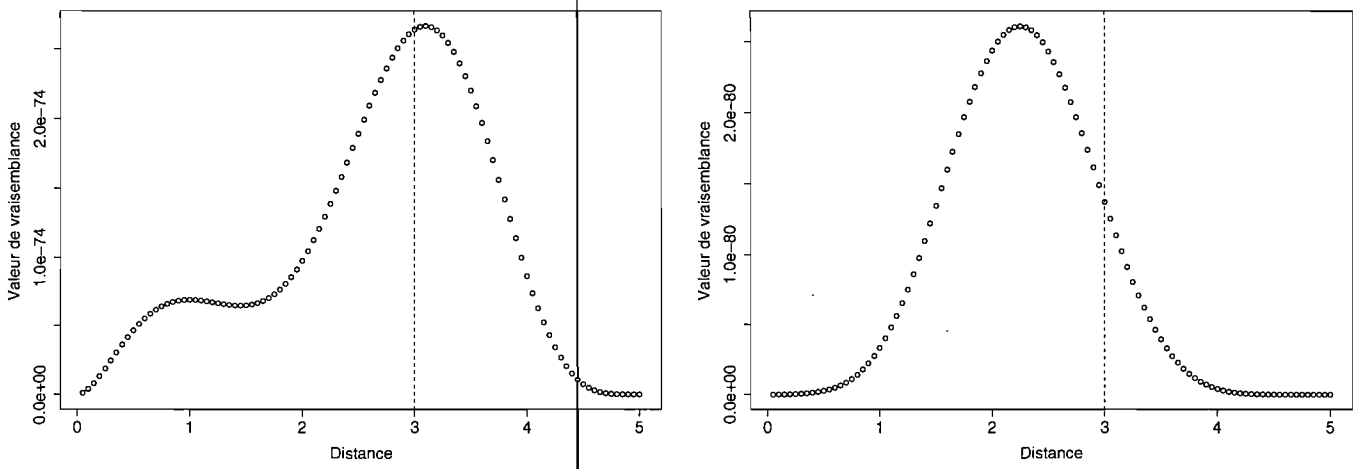

FIG. 22. Courbes de vraisemblance. Chaque courbe est générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. Les deux courbes ont été générées à partir d'une distance proposée de 1. La courbe à gauche a été produite à partir d'un échantillon de séquences à 23 loci, tandis que celle à droite a été produite à partir d'un échantillon de séquences comportant 25 loci.

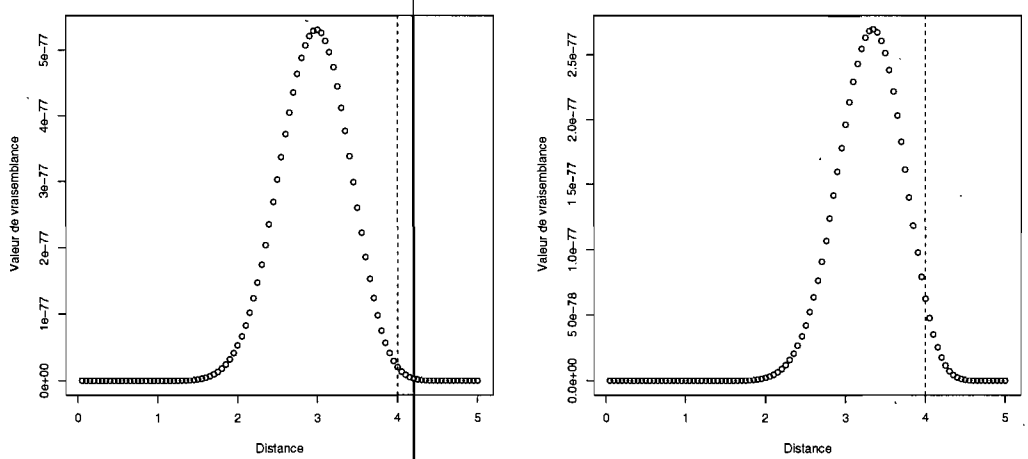

FIG. 23. Courbes de vraisemblance. Chaque courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. La courbe de gauche a été générée à partir d'une distance proposée de 2.85. La courbe à droite a été produite à partir d'une distance proposée de 3, 05.

les rondes de simulations précédentes (voir figure 24). Enfin, une nouvelle ronde de simulations, 3, 6 étant maintenant la distance proposée, produit 4 comme distance maximisant la fonction de vraisemblance (voir figure 24). Cette ronde de simulations produit aussi la plus haute valeur de vraisemblance obtenue jusqu'à maintenant.

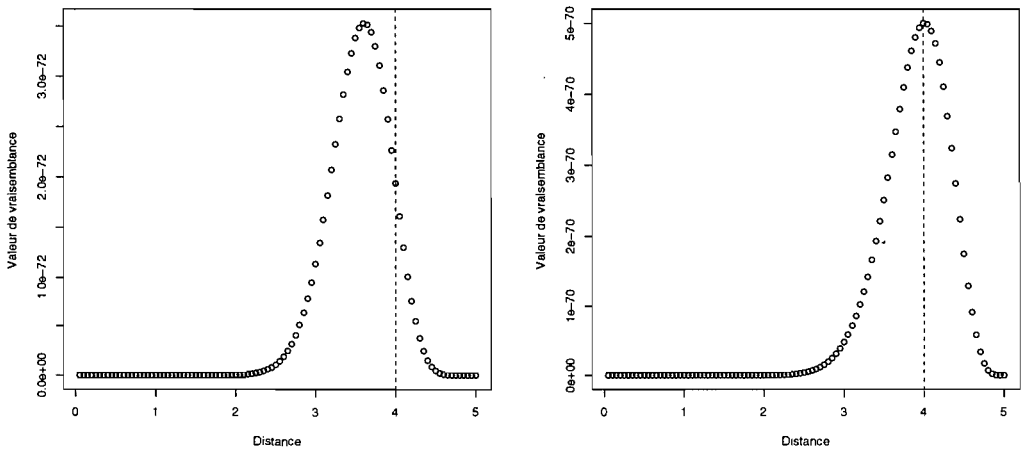

FIG. 24. Courbes de vraisemblance. Chaque courbe a été générée à partir de 1 000 000 de graphes. La ligne pointillée correspond à la distance véritable. La courbe de gauche a été générée à partir d'une distance proposée de 3.35. La courbe à droite a été produite à partir d'une distance proposée de $3,6$.

Il faut maintenant vérifier si la méthode rend encore 4 comme distance maximisant la vraisemblance quand le taux proposé est de 4. Même après 3×10^6 graphes générés, l'algorithme ne produit toujours pas 4 comme distance correspondant au maximum de la courbe (voir figure 25). Or, la courbe en question est asymétrique : on peut apercevoir sur celle-ci une bosse aux alentours de 4. On ne peut donc pas totalement écarter 4 comme distance véritable. Remarquons aussi que la valeur de vraisemblance maximale calculée dans la situation est de loin inférieure à celle obtenue avec une distance proposée de $3,6$, i.e. 4. Ainsi, on est porté à croire que 4 correspond bel et bien au taux ayant servi à générer l'échantillon.

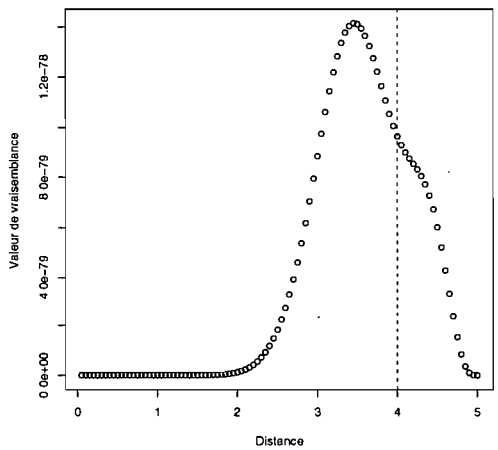

FIG. 25. Courbe de vraisemblance. La courbe a été générée à partir de 3 000 000 de graphes. La ligne pointillée correspond à la distance véritable. La distance proposée était de 4.

Chapitre 5

DISCUSSION

Le présent mémoire avait pour but de présenter dans un premier temps plusieurs concepts en génétique des populations et de les utiliser afin de dériver un modèle applicable en cartographie génétique. Dans un deuxième temps, nous cherchions à déterminer si l'échantillonage arbitraire de cas et de témoins, au coeur de ce modèle, empêchait la détermination précise de la position d'une mutation sous sélection entre deux marqueurs, celle-ci étant mesurée en fonction du taux de recombinaison. En effet, en l'absence d'interférence, comme dans le cas décrit dans le présent travail, le taux de recombinaison entre deux sites peut être utilisé comme unité de mesure de distance.

Or, il n'est pas commun d'observer une absence totale d'interférence sur un brin d'ADN suffisamment long. Par conséquent, la méthode développée ici ne peut être appliquée que dans les cas où une région très étroite a été ciblée sur le chromosome. En plus, seulement un type de recombinaison a été considéré, soit la recombinaison par *crossover*. Un modèle plus général devrait aussi inclure la possibilité de recombinaison par conversion (voir illustration 26 de l'annexe B). Cependant, l'introduction de la conversion viendrait complexifier le graphe de recombinaison ancestrale, ce qui rendrait la méthode plus difficile à appliquer.

La technique proposée présente aussi une faiblesse dans la mesure où elle ne permet pas l'estimation de la précision de la valeur trouvée pour la position de la mutation. Puisque nous n'avons aucune distribution décrivant la probabilité d'obtenir une certaine valeur pour cette position, déterminer les bornes d'un intervalle de confiance devient difficile. En effet, la dépendance existant dans l'échantillon fait en sorte que nous ne puissions pas affirmer que l'estimateur du maximum de vraisemblance est distribué normalement. En plus, la fonction obtenue par la méthode proposée n'est pas une véritable fonction de vraisemblance, mais plutôt une fonction

qui lui est analogue.

D'un autre côté, même après un grand nombre d'itérations, nous ne parvenons pas à obtenir de convergence parfaite. Le niveau de la courbe de vraisemblance varie toujours énormément d'une ronde de simulations à l'autre. Heureusement, la position du maximum semble rester dans la même région. Malgré cela, sans convergence parfaite, il est difficile d'affirmer que la méthode nous donne une estimation fiable de la position de la mutation. Cependant, les analyses de sensibilité de l'estimateur par rapport au nombre de sites retenus ont révélé qu'une augmentation du nombre de sites avait un impact positif sur la précision de l'estimation. Ainsi, il est fort possible que retenir un nombre de loci supérieur à 30 puisse diminuer de façon significative le nombre d'itérations nécessaires avant d'observer une stabilisation de la valeur de l'estimateur du maximum de vraisemblance.

D'autre part, l'utilisation de la méthode d'«importance sampling» améliore de beaucoup l'efficacité de l'algorithme pour générer des graphes de recombinaison. Or, il n'est pas rare, particulièrement quand la sélection n'est plus assumée stabilisatrice, de générer des arbres dont la vraisemblance est inférieure à la limite minimum permise par le langage C++. En fait, une augmentation notable de la vitesse de traitement a été observée quand une condition, selon laquelle la simulation d'événements devait être arrêtée quand la vraisemblance s'approchait trop de 0, a été implémentée, ce qui prouve que l'algorithme a une forte tendance à générer ces arbres. En fait, comme l'ont déjà fait remarquer Larribe et Lessard **[14]**, l'algorithme a tendance à générer seulement une très faible proportion d'arbres qui, en bout de ligne, auront un véritable impact sur la forme de la fonction de vraisemblance. Il pourrait être intéressant de déterminer les caractéristiques de ces arbres dans le but d'améliorer l'algorithme.

Il faut également noter que la distribution proposée n'est pas optimale. Une meilleure distribution pourrait sans doute être utilisée. Dans le cadre de l'estimation d'un taux de recombinaison, Stephens **[22]** a noté que le choix d'un taux proposé aura pour effet d'exagérer la vraisemblance des points proches de ce taux. La présente méthode, ayant toutefois comme but d'estimer la position d'une mutation, a tenté de compenser ce biais en adaptant la position proposée après chaque ronde de simulations. Or, l'utilisation d'une mixture comme (140) pourrait être intéressante, d'autant plus que la puissance des ordinateurs le permet maintenant aisément. Aussi, adapter une méthode éprouvée, comme celle proposée par Stephens et Donnelly **[23]** pour

le taux de recombinaison, au phénomène de sélection naturelle pourrait être une bonne idée.

Dans le cas où nous avions sélection génique, la fréquence de l'allèle au site sous sélection était assumée connue en tout temps. Ce genre d'information peut parfois être difficile à obtenir. Deux avenues seraient intéressantes à explorer. D'une part, on peut avant chaque ronde de simulation générer un processus ancestral possible. Or, la forme du processus ancestral pouvant varier énormément, employer une telle méthode augmentera probablement le nombre d'itérations nécessaires pour obtenir un résultat intéressant. D'autre part, on pourrait s'inspirer des travaux de Rannala et Slatkin [**19**] pour déterminer approximativement le temps d'apparition de la mutation. À partir de là, on pourrait générer la fréquence allélique selon un processus déterministe. Ceci aurait pour effet de rendre les simulations beaucoup plus rapides.

La disparition obligatoire de la mutation au site sous sélection dans l'échantillon au temps où elle disparaît dans la population pose une difficulté. En effet, puisqu'elle est assumée unique, il doit obligatoirement ne rester qu'une séquence mutante dans l'échantillon au moment où le temps de disparition de l'allèle mutant est atteint. Cependant, seules des séquences compatibles peuvent coalescer. Or, pour que deux séquences deviennent compatibles, certains événements de mutation sont parfois nécessaires et puisque le taux de mutation ne dépend pas de la fréquence du mutant au site sous sélection, il arrive qu'au moment où la fréquence du mutant dans la population atteigne 0, il reste encore plus d'une séquence mutante dans l'échantillon. Nous avons donc une incohérence. Afin de résoudre ce problème, au lieu de fixer à 0 la fréquence du mutant au moment de sa disparition, cette fréquence a plutôt été mise à un niveau très bas. Ainsi, puisque le taux de coalescence dans la famille des cas est inversement proportionnel à la fréquence de l'allèle mutant au site sous sélection dans la population, les événements de coalescence nécessaires pour permettre la disparition de l'allèle mutant se produiront dès que possible. Or, il faut parfois attendre longtemps avant que ce ne soit possible, ce qui induit assurément un biais. Toutefois, si l'ordre de disparition des allèles aux sites marqueurs et au site sous sélection était connu, ce problème ne se poserait plus. Nous saurions en effet quels sites sont encore polymorphes au moment de la disparition de l'allèle mutant au site sous sélection et il y aurait nécessairement compatibilité. Encore une fois, ce genre d'information peut être difficile à obtenir[1]. Il existe heureusement des méthodes d'estimation de l'âge des mutations qui pourraient nous aider à déterminer leur ordre de disparition [**7**]. D'autre part, si nous savons

[1]Notons toutefois que Griffiths et Tavaré [**6**] avaient assumé connu l'ordre de disparition des mutations.

que la mutation au site sous sélection est la plus ancienne de la séquence, ce type de problème n'apparaîtra pas non plus. Le processus de mutation pourrait être modifié pour que tous les événements de mutation aux sites marqueurs se produisent dans un intervalle de temps prédéfini.

Enfin, il est prouvé que la méthode donnera des résultats satisfaisants en autant que la taille de la population soit suffisamment grande. En pratique, il semble bien difficile de dire quand il est raisonnable d'assumer que $N \to \infty$. Une valeur de N trop petite enlève au modèle de nombreux avantages découlant de l'approche par la coalescence, notamment son efficacité.

La méthode présente aussi de nombreux avantages. D'une part, elle peut aisément s'adapter à l'inclusion du phénomène de migration. En fait, pour inclure ce phénomène, il suffirait de modifier la fonction de récurrence en classifiant les individus à l'aide de deux indicateurs, l'un pour l'îlot auquel ils appartiennent et l'autre pour l'allèle qu'ils comportent au site sous sélection. Le taux de coalescence dans un îlot i quelconque dans la famille F (cas ou témoin) sera désormais donné par $\frac{n_F^{(i)}(n_F^{(i)}-1)}{2p_F^{(i)}}$, $n_F^{(i)}$ désignant le nombre d'individus de la famille F dans l'échantillon dans l'îlot i et $p_F^{(i)}$ correspondant à la fréquence des allèles de la famille F sur l'îlôt i. Le taux de migration d'un îlot i à un îlot j, $j \neq i$ sera donné par $n^{(i)}M(i \to j)/2$, $n^{(i)}$ étant le nombre total d'individus sur l'îlôt i et $M(i \to j)/2$ étant le taux de migration pour un individu de l'échantillon ($\lim_{N\to\infty} Nm(i \to j) = M(i \to j)/2$, $m(.)$ étant le taux de migration par individu). Il faudra donc d'une part ajouter un événement possible à l'équation de récurrence (124) et d'autre part recalculer le nombre de permutations des individus de l'échantillon après chaque événement possible, ce qui nous forcera à modifier les valeurs de toutes les probabilités de transition données par (127).

Elle s'adapte facilement aussi à un changement du type de sélection, en autant que la sélection reste faible, i.e. la limite quand N tend vers l'infini de $Ns_{2N}(j)$ existe. L'ajout d'événements ayant eu une influence déterminée sur la configuration de la population à un certain instant ne semble pas problématique non plus, en autant que cet événement n'ait pas affecté de façon significative le paramètre de sélection. De plus, par l'utilisation d'une technique de vraisemblance composite, telle celle présentée par Larribe et Lessard **[14]**, délier les sites marqueurs pourrait se faire sans difficultés majeures.

D'autre part, la disponibilité d'ordinateurs toujours plus puissants rend de plus en plus attrayantes les méthodes de vraisemblance complète, comparativement aux méthodes basées sur des statistiques descriptives qui reposent la plupart du temps sur plusieurs hypothèses simplificatrices et qui sont rarement robustes à un changement d'hypothèses. Dans un même ordre d'idées, la programmation parallèle, couplée à une amélioration des algorithmes d'optimisation, semble offrir un avenir prometteur aux méthodes numériques comme celles développées dans le cadre de ce mémoire. En effet, puisqu'avec la méthode de Monte Carlo, les rondes de simulation sont indépendantes, on peut aisément répartir le travail entre un nombre arbitraire de processeurs. Un superordinateur dont l'architecture est de type «cluster» permet donc d'augmenter sans limite théorique le débit d'information traitée.

Enfin, le résultat des simulations nous permet de conclure que l'échantillonnage arbitraire de cas et de témoins ne constitue pas un obstacle à l'estimation non biaisée de la distance recherchée. Au contraire, elle permet de retrouver, en fonction du type de sélection considéré, plus ou moins efficacement la distance du site mutant sous sélection, ce qui pourrait être impossible si l'on échantillonnait totalement au hasard et si on se retrouvait de fait même avec un échantillon ne comportant aucune séquence ne portant la mutation qui nous intéresse dans les circonstances. Nous sommes confiants que les avantages du modèle qui s'est vu validé par les résultats des simulations réalisées compensent largement ses inconvénients, inconvénients qui pourront certainement être amoindris par des travaux de recherché ultérieurs.

ANNEXE

Annexe A

LE CONCEPT DE DÉSÉQUILIBRE D'APPARIEMENT (*LINKAGE DISEQUILIBRIUM*)

On ne peut aborder le concept de recombinaison sans expliquer tout d'abord ce qu'on entend par «déséquilibre d'appariement». On dira qu'il y a déséquilibre d'appariement si l'on ne peut obtenir la probabilité d'observer simultanément une paire d'allèles sur un même chromosome en simplement multipliant la probabilité d'observer chaque allèle à son locus respectif. Or, plusieurs mesures formelles du taux d'appariement ont été proposées. Considérons une séquence génétique formée de deux loci numérotés 1 et 2, comportant respectivement soit l'allèle A ou a, soit l'allèle B ou b. Définissons p_{ij} comme la probabilité d'observer simultanément au locus 1 l'allèle i et au locus 2 l'allèle j et p_i comme la probabilité d'observer l'allèle i au locus correspondant. Deux formules se démarquent comme étant particulièrement populaires pour déterminer le taux d'association entre différents marqueurs. D'une part, on peut calculer le taux d'association entre deux loci à l'aide de

$$R^2 = \frac{(p_{AB}p_{ab} - p_{Ab}p_{aB})^2}{p_A p_a p_B p_b}. \tag{149}$$

S'il y a équilibre d'appariement, la configuration à chaque locus est indépendante et R^2 prend la valeur 0. Une valeur de 1 en revanche indique que seulement deux combinaisons sont possibles. D'autre part, Lewontin **[16]** a proposé d'utiliser plutôt :

$$|D'| = \frac{|\delta|}{\delta_{max}}, \tag{150}$$

où

$$\delta = p_{AB}p_{ab} - p_{Ab}p_{aB} \tag{151}$$

et

$$\delta_{max} = \begin{cases} \min(p_A p_b, p_a p_B) & , \text{ si } \delta > 0, \\ \min(p_A p_B, p_a p_b) & , \text{ si } \delta < 0. \end{cases} \quad (152)$$

Notons que $0 \leq |D'| \leq 1$ et que 1 correspond à un appariement parfait tandis que 0 correspond à la situation d'équilibre d'appariement. La mesure introduite par Lewontin a l'avantage d'être liée directement à la distance entre deux loci. On peut voir ceci en supposant d'abord qu'une seule mutation est à l'origine d'un certain trait phénotypique. On place cette mutation au temps 0. Au moment où elle est apparue, $|\delta|$ était à son maximum. Après tout, puisqu'il n'y a pas encore eu d'événement de recombinaison pour briser le lien unissant le nouvelle mutation aux sites voisins, il est normal d'assumer que notre mesure devrait donner un taux d'appariement de 1. À mesure que le temps passe, les événements de recombinaison et de mutation viennent briser cette forte association. Cependant, plus un locus est proche du locus comportant l'allèle mutant à l'origine du caractère étudié, plus le taux d'appariement sera élevé. En fait, on observera un déclin dans le taux d'appariement suivant la loi $\delta_t = (1-\rho)^\tau \delta_0$, où ρ correspond à la distance entre les deux loci, exprimée en unités du taux de recombinaison et τ, au nombre de générations depuis l'apparition de la mutation et δ_0, au taux d'appariement initial **[16]**. On remarque donc que quand $t \to \infty$, $|D'| \to 0$, ce qui signifie que le recombinaison a pour effet d'éliminer la dépendance existant entre les deux loci.

Malheureusement, bien qu'il existe bel et bien un lien entre le déséquilibre d'appariement et la distance physique séparant deux loci sur un chromosome, la grande variabilité en pratique dans le calcul de ce déséquilibre rend très incertaines les estimations réalisées à l'aide de ces outils. En plus, différentes mesures de déséquilibre d'appariement donneront des résultats souvent très différents **[24]**. Malgré tout, de nombreux efforts ont été faits afin de développer des techniques de cartographie génétique basées sur le déséquilibre d'appariement. L'utilisation du taux de recombinaison, comme dans la méthode dont fait l'objet ce projet, est plus récente et présente aussi certaines faiblesses, telle la difficulté à transformer un taux de recombinaison en distance sur de longues séquences.

Annexe B

LE PHÉNOMÈNE DE RECOMBINAISON

La recombinaison est un phénomène courant dans la reproduction sexuée. Au moment de la méiose, il arrive qu'il y ait formation d'enjambements (aussi appelés «chiasmes») et échange de matériel génétique (voir figure 26). Par le fait même, la recombinaison accroît le nombre de combinaisons phénotypiques possibles à l'intérieur d'une même population.

Dans le cadre du projet entrepris, il est assumé qu'un seul événement de recombinaison peut se produire à la fois et que l'enjambement créé sera compris entre les deux loci périphériques. On assume qu'il ne peut y avoir que des événements de «crossover» (voir scénario 1 dans l'illustration 26). Or, cette hypothèse n'est valide que si l'on se restreint à une très courte région du génôme. Autrement, il est probable que plusieurs chiasmes se formeront à la fois.

Karlin et Liberman [**10**] se sont attardés au problème de la recombinaison et ont dérivé une distribution pour modéliser le nombre de chiasmes et leur position respective sur un chromosome. Ils notent avant de commencer leur développement que, tel qu'énoncé par Geiringer [**3**] et Schnell [**20**], la dynamique de la recombinaison se décrit mieux à l'aide de ce qu'on appelle les valeurs de *linkage*, définies dans la prochaine sous-section.

B.0.3. La distribution de recombinaison-ségrégation[1]

La distribution de recombinaison-ségrégation modélise la fréquence des différents génomes résultant de la recombinaison et de la ségrégation lors de la multiplication des gamètes. Assumons qu'un trait diploïde est défini par n loci et dénotons l'allèle au locus k ($k = 1, 2, ..., n$) par $A^{(k)}_{\epsilon_k}$, $\epsilon_k = 0$ si l'allèle est hérité du père et $\epsilon_k = 1$ si l'allèle est hérité de la mère. Il y aura donc

[1]Le développement présenté dans cette section et dans la suivante est tiré de Karlin et Liberman 1979 [**10**].

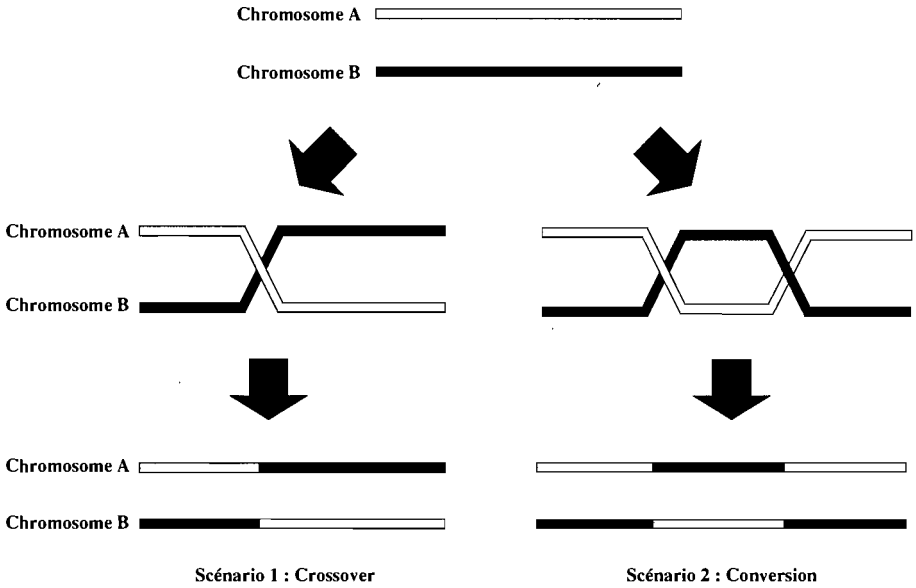

FIG. 26. Deux types de recombinaison. Dans le scénario 1, un seul chiasme est formé. Nous observons donc ce qu'on appelle un événement de «crossover». Dans le scénario 2, deux chiasmes sont formés. Nous sommes donc en présence de conversion génétique.

2^n combinaisons possibles.

Soit $R(\epsilon)$, la distribution dite de *recombinaison-ségrégation*. Cette distribution donne en fait la probabilité de chaque combinaison $\{A_{\epsilon_k}^{(k)}\}$, $k = 1, 2, ..., n$. Elle sera dénotée

$$R(\epsilon) = R(\epsilon_1, \epsilon_2, ..., \epsilon_n), \tag{153}$$

où $\epsilon_k = 0$ ou 1.

Puique chaque parent contribue également à la formation des gamètes (i.e. aucun des deux n'a plus de chance de voir ses gènes représentés dans les gamètes produits) et puisque $R(\epsilon)$ est une distribution, on peut déduire que

$$R(\epsilon_1, \epsilon_2, ..., \epsilon_n) = R(1 - \epsilon_1, 1 - \epsilon_2, ..., 1 - \epsilon_n) \tag{154}$$

et que

$$\sum_{\epsilon} R(\epsilon) = 1. \tag{155}$$

Les événements désignés $\mathbf{0} = (0, 0, ..., 0)$ et $\mathbf{1} = (1, 1, ..., 1)$ signifient que le gamète résultant est identique à celui d'un de ses parents. En d'autres mots, il n'y a pas eu de recombinaison

qu'on puisse détecter.

Quelques cas particuliers méritent d'être mentionnés.

- $n = 2$: Si r est la fréquence de recombinaison entre les deux loci, alors $R(0,0) = R(1,1) = \frac{1-r}{2}$ et $R(1,0) = R(0,1) = \frac{r}{2}$. La division en deux est une conséquence directe de (154).
- $n = 3$: Soit r, la fréquence des événements de recombinaison entre les loci 1 et 2 sans recombinaison entre les loci 2 et 3 et soit s, la fréquence des événements de recombinaison entre les loci 2 et 3 sans recombinaison entre les loci 1 et 2. Enfin, soit t, la fréquence des événements de recombinaison simultanés entre les loci 1 et 2 et les loci 2 et 3. Nous obtenons que $R(0,0,0) = R(1,1,1) = \frac{1-r-s-t}{2}$, $R(0,0,1) = R(1,1,0) = \frac{s}{2}$, $R(0,1,1) = R(1,0,0) = \frac{r}{2}$ et $R(0,1,0) = R(1,0,1) = \frac{t}{2}$.

Dans certains cas, il y aura *non-interférence*. Cela signifie que les événements de recombinaison se produisant dans des intervalles disjoints sont indépendants, ce qui implique que $t = (r+t)(s+t)$. Dans la situation où un événement de recombinaison dans un intervalle empêche la recombinaison dans le prochain intervalle, ce qu'on appelle « interférence complète », alors t prend la valeur 0.

B.0.4. Les valeurs de *linkage*

On peut obtenir une formulation alternative de la distribution de recombinaison-ségrégation en calculant sa transformée de Fourier.

Définition Soit $\boldsymbol{R} = \{R(\boldsymbol{\epsilon})\}$, la distribution de recombinaison-ségrégation. Les valeurs de *linkage* $\Gamma = \{\gamma(\boldsymbol{\delta})\}$ associées à $\boldsymbol{R}$ sont définies pour tous les vecteurs $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_n)$, où $\delta_k = 0$ ou $\delta_k = 1$ comme

$$\gamma(\boldsymbol{\delta}) = \sum_{\boldsymbol{\epsilon}} R(\boldsymbol{\epsilon})(-1)^{\sum_{k=1}^{n} \epsilon_k \delta_k}. \tag{156}$$

Les propriétés de $R(\boldsymbol{\epsilon})$ énoncées précédemment nous permettent de déduire que

- $\gamma(\mathbf{0}) = 1$, $|\gamma(\boldsymbol{\delta})| \leq 1$,
- $\gamma(\boldsymbol{\delta}) = 0$ si $|\boldsymbol{\delta}| = \sum_{k=1}^{n} \delta_k$ est impair.

Le premier item de la liste précédente est trivial. Le deuxième item l'est moins. On peut le prouver ainsi.

Puisque $R(\boldsymbol{\epsilon}) = R(\mathbf{1} - \boldsymbol{\epsilon})$, alors

$$\begin{aligned} \gamma(\boldsymbol{\delta}) = \sum_{\boldsymbol{\epsilon}} R(\boldsymbol{\epsilon})(-1)^{\sum_{k=1}^{n} \epsilon_k \delta_k} &= \gamma(\boldsymbol{\delta}) = \sum_{\boldsymbol{\epsilon}} R(\mathbf{1} - \boldsymbol{\epsilon})(-1)^{\sum_{k=1}^{n} (1-\epsilon_k)\delta_k} \\ &= \sum_{\boldsymbol{\epsilon}} R(\mathbf{1} - \boldsymbol{\epsilon})(-1)^{\sum_{k=1}^{n} \delta_k}(-1)^{-\sum_{k=1}^{n} \epsilon_k \delta_k} \\ &= \sum_{\boldsymbol{\epsilon}} R(\boldsymbol{\epsilon})(-1)^{\sum_{k=1}^{n} \epsilon_k \delta_k}(-1)^{\sum_{k=1}^{n} \delta_k}, \end{aligned} \tag{157}$$

ce qui implique que $\gamma(\boldsymbol{\delta}) = 0$ si $\sum_{k=1}^{n} \delta_k$ est impair.

Si $|\boldsymbol{\delta}|$ est pair, nous pouvons déduire de (156) que $\gamma(\boldsymbol{\delta})$ sera déterminé par la distribution marginale de $\boldsymbol{R}$, $\boldsymbol{R_I}$, où $I = I(\boldsymbol{\delta}) = \{i : \delta_i = 1\}$.

En inversant la transformée de Fourier, nous obtenons

$$R(\boldsymbol{\epsilon}) = \frac{1}{2^n} \sum_{\boldsymbol{\delta}} \gamma(\boldsymbol{\delta})(-1)^{\sum_{k=1}^{n} \delta_k \epsilon_k}. \tag{158}$$

La relation bijective existant entre une fonction et sa transformée de Fourier fait en sorte que $\gamma(\boldsymbol{\delta})$ puisse être associée sans ambivalence à $R(\boldsymbol{\epsilon})$.

Or, cette nouvelle formulation recèle un avantage. Tel qu'indiqué par Schnell **[20]**, les valeurs de *linkage* ont une interprétation biologique. En effet, par exemple, si $|\boldsymbol{\delta}| = 2$ avec $\delta_i = \delta_j = 1$ alors $r(\boldsymbol{\delta}) = r_{ij}$ correspond au taux de recombinaison entre les loci i et j. En outre, si $|\boldsymbol{\delta}| = \sum_{k=1}^{n} \delta_k$ est pair, alors $\gamma(\boldsymbol{\delta}) = 1 - 2r(\boldsymbol{\delta})$ et on appelle $r(\boldsymbol{\delta})$ une «valeur de recombinaison généralisée».

Afin de calculer cette valeur, on définit d'abord un ensemble de segments disjoints formés par m paires de loci, chaque loci étant identifié par i_k, $k = 1, 2, ..., 2m$, $2m \leq n$, et chaque intervalle par $\Delta_{i_{2k-1}, i_{2k}}$, $k = 1; 2, ..., m$. Maintenant $\boldsymbol{\delta} = (\delta_1, \delta_2, ...\delta_n)$ correspond au vecteur d'incidence de $\{i_k,\ k = 1, 2, ..., 2m\}$, c'est-à-dire

- $\delta_k = 1$ si le locus k est compris dans $\{i_j\}$, $j = 1, 2, ..., 2m$,

- $\delta_k = 0$ autrement.

Dans ce cas, $r(\boldsymbol{\delta})$ correspond à la probabilité qu'un nombre impair d'enjambements se produisent sur la région du génôme $\bigcup_{k=1}^{m} \Delta_{i_{2k-1},i_{2k}}$ et prendra la valeur

$$r(\boldsymbol{\delta}) = \sum_{<\boldsymbol{\epsilon},\boldsymbol{\delta}> \text{ impair}} R(\boldsymbol{\epsilon}), \tag{159}$$

où la condition sur la somme correspond à tous les $\boldsymbol{\epsilon}$ satisfaisant $< \boldsymbol{\epsilon}, \boldsymbol{\delta} >= \sum \epsilon_k \delta_k = \sum_{k=1}^{2m} \epsilon_{i_k} =$ *nombre impair.*

On peut maintenant noter le lien unissant les valeurs de *linkage* aux taux de recombinaison généralisés :

$$\gamma(\boldsymbol{\delta}) = \begin{cases} 1, \text{ si } \boldsymbol{\delta} = \mathbf{0}, \\ 1 - 2r(\boldsymbol{\delta}) \text{ si } |\boldsymbol{\delta}| = \sum_{i=1}^{n} \delta_i \text{ est pair et supérieur à } 0, \\ 0 \text{ si } |\boldsymbol{\delta}| \text{ est impair.} \end{cases} \tag{160}$$

Les cas où $\boldsymbol{\delta} = \mathbf{0}$ et où $|\boldsymbol{\delta}|$ est impair ont déjà été couverts précédemment. Le cas où $|\boldsymbol{\delta}|$ est pair mérite une preuve.

$$\begin{aligned} \gamma(\boldsymbol{\delta}) &= \sum_{\boldsymbol{\epsilon}} R(\boldsymbol{\epsilon})(-1)^{\sum_{k=1}^{n} \epsilon_k \delta_k} \\ &= -\sum_{<\boldsymbol{\epsilon},\boldsymbol{\delta}> \text{ impair}} R(\boldsymbol{\epsilon}) + \sum_{<\boldsymbol{\epsilon},\boldsymbol{\delta}> \text{ pair}} R(\boldsymbol{\epsilon}) \\ &= -r(\boldsymbol{\delta}) + (1 - r(\boldsymbol{\delta})) = 1 - 2r(\boldsymbol{\delta}) \end{aligned}$$

On constate donc que $r(\boldsymbol{\delta})$ nous donne plus d'information que les simples taux de recombinaison $\{r_{ij}\}$.

Attardons-nous enfin au cas de non-interférence, puisqu'il s'agit de l'hypothèse qui sera employée dans les sections subséquentes. Tel qu'énoncé auparavant, l'hypothèse de non-interférence signifie que les événements d'enjambement se produisent indépendamment sur des intervalles distincts du génome. Commençons par noter que si un événement de recombinaison produit une séquence $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$ et si $\epsilon_k - \epsilon_{k+1} = 0$, alors on peut affirmer qu'il n'y a pas eu de recombinaison entre les loci k et $k+1$. En revanche, si $|\epsilon_{k+1} - \epsilon_k| = 1$, on peut conclure qu'un événement de recombinaison s'est produit entre les loci k et $k+1$. Si r_k désigne le taux

de recombinaison entre ces loci, en assumant qu'il y a non-interférence, la probabilité $R(\epsilon)$ est donnée par

$$R(\epsilon) = \frac{1}{2}\prod_{k=1}^{n-1} r_k^{|\epsilon_{k+1}-\epsilon_k|}(1-r_k)^{1-|\epsilon_{k+1}-\epsilon_k|}. \tag{161}$$

De plus, les valeurs de *linkage* associées avec l'hypothèse de non-interférence ont aussi une forme multiplicative :

$$\gamma(\boldsymbol{\delta}) = \prod_{v=1}^{2m}(1-2r_{i_{2v-1},i_{2v}}), \tag{162}$$

$|\boldsymbol{\delta}| = 2m$, $\delta_{i_1} = \delta_{i_2} = ... = \delta_{i_{2m}} = 1$ et $r_{i,j}$ correspond au taux de recombinaison entre les loci i et j. Les $r_{i,j}$, utilisés dans l'équation (162), peuvent être calculés à partir des taux $r_1, r_2, ..., r_{n-1}$ à l'aide de la formule de récurrence

$$r_{i,j} = r_i(1-r_{i+1,j}) + (1-r_i)r_{i+1,j}, \tag{163}$$

où $r_i = r_{i,i+1}$ et $1 \leq i \leq j \leq n$.

On remarque enfin, à partir de l'équation précédente, que si $0 \leq r_i \leq \frac{1}{2}$, alors $r_{i,j}$ sera une fonction monotone non-décroissante de r_i, $i = 1, 2, ..., n-1$.

Annexe C

MÉTHODES D'ESTIMATION BAYÉSIENNES ET NON-BAYÉSIENNES DE PARAMÈTRES GÉNÉTIQUES

Bien que les méthodes d'inférence employées dans le cadre de ce projet de recherche soient non-bayésiennes, plusieurs spécialistes préfèrent l'approche bayésienne. Les deux approches pour estimer des paramètres génétiques qui dépendent de généalogies sous-jacentes inconnues ont des caractéristiques qui se doivent d'être précisées[1].

Supposons que l'on tente d'estimer la valeur d'un quelconque paramètre θ lié à une population donnée. Le meilleur moyen d'y parvenir est de prélever un échantillon représentatif de cette population et de calculer la valeur de vraisemblance de l'échantillon en fonction de ce paramètre. Pour ce faire, on a besoin d'une fonction de vraisemblance

$$L(\theta) = P(D|\theta), \tag{164}$$

où D correspond à l'information contenue dans notre échantillon et θ au paramètre que l'on tente d'estimer. Dans un problème de génétique des populations, D correspond la plupart du temps à un échantillon aléatoire de segments chromosomiques et θ, à une variété de paramètres, dont les plus communs sont le taux de mutation, le taux de recombinaison, la taille effective de la population ou, pour un modèle avec populations cloisonnées, le taux de migration.

Presque toujours, même dans les cas les plus simples, il est impossible d'obtenir une formulation explicite pour la vraisemblance d'un échantillon de séquences génétiques. Ce problème provient entre autres de la dépendance entre les séquences échantillonées. Cette dépendance résulte des liens ancestraux les unissant. Par exemple, si l'on assume le modèle à une infinité de

[1] La plupart des affirmations de cette section sont tirées de Stephens 2001 [**22**].

sites, on sait que la variante mutante d'un allèle n'a qu'un unique ancêtre. Ainsi, si l'on observe cette variante plusieurs fois dans l'échantillon, à cause de la dépendance existant aussi entre les configurations à différents loci, on peut déduire que ces séquences partagent des ancêtres communs. Cet état de choses nous oblige à faire appel à des approximations. Deux approches se sont imposées : l'approche du maximum de vraisemblance et l'approche bayesienne.

L'approche du maximum de vraisemblance consiste simplement à bâtir une fonction de vraisemblance et à choisir comme estimé la valeur $\hat{\theta}$ maximisant cette fonction. La précision de l'estimation est généralement évaluée à l'aide d'un intervalle de confiance bâti autour de la valeur du paramètre. En théorie, si les membres de l'échantillon sont indépendants, la valeur de l'estimé sera asymptotiquement distribuée normalement avec moyenne prenant la valeur véritable du paramètre qu'on tente d'évaluer. D'autre part, la statistique du rapport de log-vraisemblance

$$\Lambda = -2log\frac{L(\theta_0)}{L(\hat{\theta})} \tag{165}$$

sera asymptotiquement décrite par une distribution χ^2 si θ_0 est la véritable valeur du paramètre. Ceci vient faciliter la détermination de la grandeur des intervalles de confiance. Malheureusement, la dépendance entre les membres de l'échantillon rend le comportement asymptotique de l'estimateur difficilement prévisible. Ainsi, il est difficile de dire quand ces propriétés avantageuses sont applicables.

L'approche bayésienne fait aussi appel au concept de vraisemblance pour estimer la valeur du paramètre recherché. Or, pour l'appliquer, il faut faire une hypothèse par rapport à la distribution du paramètre. On l'appelle la distribution *a priori*. Notre but ultime est de trouver la valeur de $P(\theta|D)$, qu'on appelle la distribution *a posteriori*. On peut y parvenir en appliquant la relation

$$P(\theta|D) = L(\theta)\frac{P(\theta)}{P(D)}. \tag{166}$$

Cette équation nous permet de réaliser que $P(\theta|D)$ est élevée quand la valeur de θ est très compatible avec les données, ce qui se traduit par une valeur élevée pour $L(\theta)$ et quand la valeur du paramètre est probable, ce qui signifie que $P(\theta)$ est assez haut.

L'approche bayesienne présente deux bénéfices majeurs. D'une part, contrairement à l'approche du maximum de vraisemblance, elle ne dépend pas du comportement asymptotique d'un

estimateur. Ainsi, elle peut nous être utile dans les cas où la méthode du maximum de vraisemblance ne nous permet pas de tirer de conclusion. D'autre part, elle nous permet d'obtenir des intervalles de crédibilité. Un intervalle de crédibilité de $(1-\alpha)100\%$ contiendra avec probabilité $(1-\alpha)$ la véritable valeur du paramètre. Un intervalle de confiance n'a malheureusement pas cette caractéristique. En revanche, cette approche présente aussi un désavantage majeur. En effet, très souvent, la valeur de la distribution *a posteriori* sera fortement influencée par la valeur de la distribution *a priori*. Par conséquent, deux chercheurs ayant une perception totalement différente de la nature de la distribution *a priori* risquent d'arriver à des conclusions très différentes. Par conséquent, on peut dire que cette méthode s'accompagne d'une certaine dose de subjectivité. Or, tel qu'énoncé précédemment, la nature corrélée des échantillons de segments chromosomiques élimine les avantages découlant des caractéristiques asymptotiques de la méthode du maximum de vraisemblance. Ainsi, une approche bayesienne peut malgré ses défauts se révéler utile dans un tel contexte.

Notons au passage que la méthode d'inférence de Griffiths et Marjoram **[5]**, celle de Larribe, Lessard et Schork **[15]** ainsi que celle de Rannala et Slatkin **[19]** font partie de la famille des méthodes non-bayesiennes. En revanche, Nielsen **[17]** a plutôt opté pour une approche bayesienne.

BIBLIOGRAPHIE

[1] W. J. Ewens. Conditional diffusion processes in population genetics. *Theoretical Population Biology*, 4 :21–30, 1973.

[2] W. J. Ewens. In *Mathematical Population Genetics*. Springer-Verlag, New York, N.Y., 2004.

[3] H. Geiringer. On the probability theory of linkage in Mendelian heredity. *Ann. Math. Statistics*, 15 :25–57, 1944.

[4] R. C. Griffiths. The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoretical Population Biology*, 64 :241–251, 2003.

[5] R. C. Griffiths and P. Marjoram. Ancestral inference from samples of dna sequences with recombination. *J. Comput. Biol.*, 3 :479–502, 1996.

[6] R. C. Griffiths and Simon Tavaré. Ancestral inference in population genetics. *Statist. Sci.*, 9(3) :307–319, 1994.

[7] R.C. Griffiths and S. Lessard. Ewens' sampling formula and related formulae, combinatorial proofs : extensions to variable population size and applications to ages of alleles. *Theoretical Population Biology*, 68 :167–177, 2005.

[8] R. R. Hudson and N. L. Kaplan. The coalescent process in models with selection and recombination. *Genetics*, 120 :831–840, 1988.

[9] N. L. Kaplan, T. Darden, and R. R. Hudson. The coalescent process in models with selection. *Genetics*, 120 :819–829, 1988.

[10] S. Karlin and U. Liberman. A natural class of multilocus recombination processes and related measures of crossover interference. *Adv. in Appl. Probab.*, 11(3) :479–501, 1979.

[11] S. Karlin and H. M. Taylor. In *A Second Course in Stochastic Processes*, volume 2. Wiley, New York, N.Y., 1980.

[12] J. F. C. Kingman. The coalescent. *Stoch. Proc. Appl.*, 13 :235–248, 1982.

[13] S. M. Krone and C. Neuhauser. Ancestral processes with selection. *Theoretical Population Biology*, 51 :210–237, 1997.

[14] F. Larribe and S. Lessard. A composite-conditional likelihood approach for gene mapping based on linkage disequilibrium in windows or marker loci. *Statistical Applications in Genetics and Molecular Biology*, To appear 2008.

[15] F. Larribe, S. Lessard, and N. J. Schork. Gene mapping via the ancestral recombination graph. *Theoretical Population Biology*, 62 :215–229, 2002.

[16] R. Lewontin. On measures of gametic disequilibrium. *Genetics*, 120 :849–852, 1988.

[17] R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154 :931–942, 2000.

[18] M. Nordborg. Coalescent theory. In *Handbook of Statistical Genetics*, pages 179–209. Wiley, 2001.

[19] B. Rannala and M. Slatkin. Likelihood analysis of disequilibrium mapping, and related problems. *American Journal of Human Genetics*, 62 :459–473, 1998.

[20] F. W. Schnell. Some general formulations of linkage effects in inbreeding. *Genetics*, 46 :947–957, 1961.

[21] P. Sonigo and J. J. Kupiec. *Ni Dieu ni gène*. Seuil, 2001.

[22] M. Stephens. Inference under the coalescent. In *Handbook of Statistical Genetics*, pages 213–238. Wiley, 2001.

[23] M. Stephens and P. Donnelly. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(4) :605–655, 2000. With discussion and a reply by the authors.

[24] S. Tavaré. Ancestral inference in population genetics. In *Lectures on Probability Theory and Statistics*, volume 1837 of *Lecture Notes in Mathematics*, pages 1–188. Springer, Berlin, 2004.

[25] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7 :256–276, 1975.