Université de Montréal

Probabilité de fixation dans des modèles génétiques de populations à plusieurs allèles

par

Philippe Lahaie

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 en Mathématiques appliquées

juin 2008

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Probabilité de fixation dans des modèles génétiques de populations à plusieurs allèles

présenté par

Philippe Lahaie

a été évalué par un jury composé des personnes suivantes :

.....

président-rapporteur

.....

directeur de recherche

.....

membre du jury

Mémoire accepté le .....

Résumé

Pour des modèles génétiques de populations finies haploïdes ou diploïdes à plusieurs allèles, on dégage une équation donnant la probabilité de fixation d'un allèle particulier sous l'action d'une faible sélection naturelle à partir de l'espérance du changement de fréquence de cet allèle en une génération. On montre ensuite comment calculer des espérances de produits de fréquences alléliques à l'aide de la théorie de la coalescence, ce qui permet d'exprimer la probabilité de fixation en fonction des fréquences initiales des allèles. Cette méthode est finalement appliquée à différents modèles de populations, variantes des modèles de Wright-Fisher, Moran et Cannings, pour lesquels on obtient cette probabilité de fixation. En analysant ces résultats, on établit des conditions sous lesquelles un allèle est avantagé ou désavantagé par une faible sélection.

Mots clés :

Modèle de Wright-Fisher, modèle de Moran, modèle de Cannings, probabilité de fixation, théorie de la coalescence.

Abstract

For genetic models of finite haploid or diploid populations with multiple alleles, we deduce an equation giving the probability of fixation of an allele under the influence of a weak natural selection based on the expectation of the change in frequency of this allele in one generation. Then we show how to compute the expectation of products of allelic frequencies in any generation using the coalescent theory. This allows us to give the fixation probability as a function of the initial allelic frequencies. Finally, this method is applied to different population genetic models, namely the Wright-Fisher, Moran and Cannings models. The results serve to establish conditions under which an allele is favored or disfavored by weak selection.

Keywords :

Wright-Fisher Model, Moran model, Cannings model, probability of fixation, coalescent theory.

Table des matières

Introduction

La société actuelle utilise le vocable de *Théorie de l'évolution* pour désigner la conception scientifique moderne du mécanisme temporel par lequel une population d'êtres vivants se modifie et crée de nouvelles espèces. C'est plus précisément aux idées de la *Théorie synthétique de l'évolution* qu'on veut référer, celle-ci étant généralement acceptée des biologistes comme l'explication la plus étoffée de l'évolution des espèces.

La genèse de cette théorie remonte à la publication en 1859 de *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life* par le naturaliste Charles Darwin. Ce titre, en soit, est l'idée principale du darwinisme : les individus dont les caractéristiques favorisent la survie ou influencent positivement tout autre aspect qui améliore leurs perspectives de reproduction dans une population où il y a compétition pour les ressources auront de meilleures chances de créer une progéniture, progéniture qui conservera leurs caractères. Par ce processus, ces caractères favorables seront transmis dans le temps, c'est-à-dire conservés par sélection naturelle, tandis que les caractères défavorables disparaîtront progressivement. Bien que ce principe soit exact et capital, Darwin est arrivé à cette conclusion par observation et logique, sans pouvoir en expliquer le fonctionnement, la mécanique. Entre autres, il n'avait qu'une vague idée de ce qui pouvait être transmis lors de la reproduction (caractères héréditaires) et encore moins de quelle façon s'effectuait cette transmission.

À la même époque, le botaniste Gregor Mendel réalisait ses expériences sur les petits pois, maintenant légendaires, qui furent déterminantes dans l'énonciation des trois lois de Mendel, qui sont la clé de la transmission des caractères héréditaires. En croisant des pois jaunes et des pois verts provenant de lignées pures, Mendel n'obtenait que des pois jaunes. La couleur des pois de Mendel est un exemple de dominance chez un gène diploïde avec deux allèles possibles. Un gène est une unité d'information génétique, en fait une séquence d'ADN, se retrouvant à un emplacement précis d'un chromosome, emplacement dit locus ou loci au pluriel. Ici, les pois sont diploïdes, car chacun de leurs gènes est formé de deux allèles, par contraste avec un individu haploïde, dont les gènes ne comportent qu'un seul allèle. Les allèles sont les variantes possibles pour les séquences d'ADN pouvant apparaître à un certain locus. Si on dénote A l'allèle donnant la couleur jaune aux pois et a l'allèle responsable de la couleur verte, AA, Aa

et aa sont les génotypes possibles pour les pois de Mendel, un génotype étant la composition allélique des gènes d'un organisme. Les génotypes AA et aa sont homozygotes puisqu'ils sont composés du même allèle, tandis que Aa est dit hétérozygote. Quand un nouvel individu est formé, il emprunte un allèle à chacun de ses parents. Le botaniste utilisait des lignées pures de pois jaunes et de pois verts, c'est-à-dire des pois qui, croisés avec eux-mêmes, ne produisent que des pois jaunes et verts respectivement. Ceci est le résultat de l'homozygotie. Ses pois jaunes étaient donc de génotype AA et ses pois verts de génotype aa. En croisant les deux ensemble, il obtenait uniquement des individus Aa, génotype qui résultait invariablement en un pois jaune. L'expression d'un génotype est appelé le phénotype. Ici, comme l'allèle a dans le génotype Aa ne s'exprime pas dans le phénotype de l'individu, on dit qu'il est récessif, et que l'allèle A est dominant. C'est un seul des multiples cas d'interdépendance possibles entre les allèles. Mendel n'utilisait pas un tel vocabulaire pour décrire ses expériences. Ces termes ont été définis par les botanistes qui ont redécouvert ses travaux au début du vingtième siècle et qui lui ont reconnu la paternité de la science qu'on appelerait désormais la génétique.

Dans toutes les expériences de Gregor Mendel, les caractères étudiés étaient discrets. Par exemple, couleur jaune ou verte, peau lisse ou ridée, forme longue ou courte, etc. La théorie mendélienne n'expliquait pas de quelle façon, chez beaucoup d'espèces, un caractère pouvait être exprimé à plusieurs degrés différents : couleur d'un ton entre le jaune et le vert, longueur de la queue entre dix et quinze centimètres, etc. Ceci semblait en contradiction avec Darwin qui endossait l'axiome grec *natura non facit saltus* signifiant *la nature ne fait pas de saut.* Pour cette raison, les lois mendéliennes furent d'abord perçues comme un argument contre le darwinisme, et ce jusqu'en 1918. À cette date, le généticien et statisticien Ronald Fisher termine *The Correlation Between Relatives on the Supposition of Mendelian Inheritance* dans lequel il explique comment une variation continue peut être le résultat de l'interdépendance de plusieurs loci, conciliant du coup les idées de Darwin et de Mendel. Ce texte est le premier soubresaut de la génétique de population, qui prendra son envol dans les années suivantes grâce aux contributions répétées de Ronald Fisher, John Burdon Sanderson Haldane et Sewall Wright. Cette synthèse mathématique est la pierre angulaire d'un ensemble de contributions réunies pour créer la *théorie synthétique de l'évolution*, qui gagna son nom dans les années mille neuf cent quarante. Cette thèse, bien que développée, complétée et raffinée, demeure à ce jour ce qu'on appelle la *théorie de l'évolution.*

Ce mémoire relève de la génétique de population puisqu'on y étudiera l'évolution temporelle de populations de taille constante finie d'individus haploïdes ou diploïdes avec de multiples allèles sous l'action de la sélection naturelle et du hasard. On considérera trois cadres principaux pour modéliser nos populations, soit les modèles de Wright-Fisher, de Moran et de Cannings, qui sont introduits dans le livre [2]. Le modèle de Wright et de Fisher, tout deux mentionnés

ci-dessus, se veut le modèle le plus simple qu'on puisse imaginer. Un groupe d'individus crée un autre groupe d'individus en pigeant dans son bagage génétique commun. Le modèle de Moran considère qu'un seul individu dans le groupe se reproduit, son rejeton prenant la place d'un individu du groupe. Pour le modèle de Cannings, l'approche est tout autre. On considère le nombre d'enfants qu'aura chaque individu du groupe comme une variable aléatoire. Dans le premier modèle, la sélection s'exprime en modifiant les probabilités de survie des individus. Dans le deuxième, elle influe sur les probabilités des individus d'être remplacés par le nouveau né. Pour le dernier modèle, la sélection apparaît dans les moyennes et les variances des variables aléatoires, un individu favorisé par la sélection ayant en moyenne plus de rejetons qu'un autre moins favorisé. Ces modèles existent sous une foule de variantes, parmi lesquelles on sélectionnera quelques exemples.

Si on considère un allèle particulier dans une population d'individus, sa fréquence peut augmenter ou diminuer dans le temps. Il peut éventuellement disparaître totalement ou, à l'opposé, constituer l'intégralité de la population. On dit alors qu'il y a fixation de l'allèle. Sous certaines conditions, relativement laxistes, on peut montrer que, si on attend suffisamment longtemps, un de ces deux événements se produira certainement. Autrement dit, les populations ici étudiées seront éventuellement composées d'un unique allèle. Il serait intéressant de savoir avec quelle probabilité chaque allèle peut attendre cet état de domination, ce qui tracerait le portrait du destin de la population. C'est l'objectif de ce mémoire.

Dans l'article [5] publié en 2007, Sabin Lessard et Véronique Ladret s'intéressent à la probabilité de fixation dans le modèle de Cannings. Ils montrent comment on peut calculer l'effet de premier ordre d'une faible sélection sur la probabilité de fixation d'un nouvel allèle plongé à un unique exemplaire dans une population homogène de taille finie. Les résultats qu'ils obtiennent sont cohérents avec ceux donnés par des approximations par processus de diffusion lorsque la taille de la population tend vers l'infini.

Nous emprunterons ici la même démarche, mais généraliserons le cadre de départ en considérant des gènes polymorphes, c'est-à-dire des gènes comportant plus de deux allèles différents. À partir des fréquences initiales de chaque allèle, il nous sera possible, pour une sélection faible et sous certaines conditions de continuité, de donner une formule explicite pour la probabilité de fixation de chacun d'entre-eux en négligeant les termes d'ordre du carré de l'intensité de sélection ou supérieur. Ceci est possible pour les modèles présentés et pour beaucoup d'autres. Ce résultat est d'autant plus intéressant que, dans le cas de gènes polymorphes, l'approche par processus de diffusion est difficile à emprunter.

Chapitre 1

Méthode générale

Dans ce premier chapitre sont regroupés des éléments qui seront utiles au calcul de la probabilité de fixation d'un allèle dans tous les modèles de population que nous présenterons par la suite. On y trouve une formule donnant la probabilité de fixation d'un allèle en fonction de l'espérance conditionnelle d'une variable aléatoire particulière. On développe par la suite les outils nécessaires au calcul de cette espérance, ce qui termine notre préparation à l'étude des différents modèles.

1.1 Généralités sur les modèles

Les différents modèles étudiés seront tous des modèles de populations haploïdes ou diploïdes à générations discrètes. Nous considérerons des populations finies constituées invariablement de $2N$ gènes, donc de $2N$ individus si elles sont haploïdes ou de N individus si elles sont diploïdes.

Dans tous les cas, il y aura n allèles différents que l'on représentera par $A_1, A_2, \ldots, A_n$ et dont on notera les fréquences à la génération t par les variables aléatoires $x_1(t), x_2(t), \ldots, x_n(t)$. On considérera $x_1(0), x_2(0), \ldots, x_n(0)$ comme des constantes connues; c'est l'état initial de la population.

Nous supposerons la présence de sélection naturelle, mais l'absence de mutations. L'intensité de la sélection sera représentée par le paramètre positif s qu'on considérera beaucoup plus petit que 1, ce qui correspond à une sélection faible. Ce paramètre représente la magnitude de l'écart entre les différents phénotypes des individus de la population pour ce qui est de leur aptitude à survivre puis produire une descendance. Le cas $s = 0$ correspond à l'absence de sélection. On se dira alors dans une situation de neutralité, où aucun individu n'est avantagé par rapport aux autres et où la population évolue selon les forces uniques du hasard. Plus de détails se trouvent à l'annexe B qui pousse l'étude de la neutralité dans une problématique propre à la section 1.7.

1.2 Un peu de vocabulaire

À l'aide d'un exemple, nous allons définir quelques termes que nous utiliserons librement par la suite.

Un gène quelconque choisi à la génération t (appelons-le α) porte un allèle particulier. Cet allèle est la copie de celui d'un gène de la génération $t-1$ (appelons-le β), qui est lui-même une copie d'un allèle de la génération $t-2$ (appelons-le γ). On peut ainsi reconstruire la lignée ancestrale de cet allèle jusqu'à un gène (appelons-le δ) d'une génération antérieure quelconque.

Les gènes β, γ et δ seront appelés les *ancêtres* de α. De façon plus précise, γ est l'ancêtre de α à la génération $t-2$. De plus, β sera appelé le *géniteur* de α, car il est son plus proche ancêtre. Définissons aussi les concepts réciproques de *descendant* et de *rejeton*. Ainsi, α est le descendant de β, γ et δ. Il est aussi le rejeton de β.

On peut considérer les ancêtres de plusieurs gènes choisis à la génération t. Il y aura parfois recoupement entre les ancêtres des différents gènes et le résultat sera un arbre généalogique que nous appellerons l'*arbre ancestral*. L'ensemble des ancêtres à une génération donnée est appelé le *matériel ancestral* à cette génération. Cette approche est celle de la théorie de la coalescence. On dit qu'il y a *coalescence* entre deux ou plusieurs gènes s'ils ont le même géniteur. On dit de façon équivalente qu'ils *coalescent*.

1.3 Probabilité de fixation d'un allèle

Notre but est d'obtenir une expression pour la probabilité qu'un allèle forme éventuellement l'intégralité de la population. On dit alors qu'il y a fixation de l'allèle. Il sera donc utile de définir l'événement

$$F_i : \text{Fixation éventuelle de l'allèle } A_i, \tag{1.1}$$

et de dénoter par $P_s(F_i)$ la probabilité de fixation éventuelle de l'allèle A_i lorsque l'intensité de la sélection est s. La probabilité de fixation de l'allèle A_i peut être exprimée à l'aide d'espérances de changements de fréquences d'allèles par le raisonnement suivant tiré de [5].

Tous les modèles de reproduction que nous considérerons respectent la propriété markovienne, c'est-à-dire que les fréquences des différents allèles à une génération dépendent uniquement de l'état de la population à la génération précédente. De façon plus formelle, si l'état de la population à la génération t est noté $\vec{x}(t) = (x_1(t), \ldots, x_n(t))$, alors

$$P_s\left[\vec{x}(t) = \vec{v}(t) \middle| \forall\, u \in \{0,1,\ldots,t-1\}, \vec{x}(u) = \vec{v}(u)\right] = P_s\left[\vec{x}(t) = \vec{v}(t) \middle| \vec{x}(t-1) = \vec{v}(t-1)\right], \tag{1.2}$$

pour tout $t \geqslant 1$ et pour toute suite de vecteurs $\left\{\overrightarrow{v}(r)\right\}_{r\in\mathbf{Z}^+}$ appartenant à $\{0, 1/2N, 2/2N, \ldots, 1\}^n$ tels que $\sum_{i=1}^{n} v_i(r) = 1$ pour tout $r \in \mathbf{Z}^+$. La probabilité est encore une fois indicée par s parce qu'elle dépend de l'intensité de la sélection. En fait, non seulement la propriété markovienne est respectée, mais les probabilités de transition ne changent pas avec le temps. On peut donc renforcer la dernière proposition en disant que

$$P_s\left[\overrightarrow{x}(t) = \overrightarrow{v}(t) \middle| \forall\, u \in \{0, 1, \ldots, t-1\}, \overrightarrow{x}(u) = \overrightarrow{v}(u)\right] = P_s\left[\overrightarrow{x}(1) = \overrightarrow{v}(t) \middle| \overrightarrow{x}(0) = \overrightarrow{v}(t-1)\right]. \tag{1.3}$$

Ainsi, la suite des vecteurs aléatoires $\overrightarrow{x}(t)$, pour $t \geqslant 0$, forme une chaîne de Markov homogène, dont $\mathbf{P}_s$ sera la matrice de transition sous une intensité de sélection s. La matrice $\mathbf{P}_s^{(t)}$ sera, quant à elle, la matrice de transition en t pas. Ainsi, l'entrée de la matrice $\mathbf{P}_s\left(\overrightarrow{v}, \overrightarrow{v'}\right)$ est la probabilité que cette population passe de l'état $\overrightarrow{v}$ à l'état $\overrightarrow{v'}$ en une génération, tandis que l'entrée de la matrice $\mathbf{P}_s^{(t)}\left(\overrightarrow{v}, \overrightarrow{v'}\right)$ est la probabilité du même changement, mais en t générations. Notons que la matrice de transition en t pas est la t-ieme puissance de la matrice de transition, c'est-à-dire que

$$\mathbf{P}_s^{(t)} = \mathbf{P}_s^t. \tag{1.4}$$

Ces matrices carrées sont de dimension

$$\binom{2N + n - 1}{n - 1}. \tag{1.5}$$

Cette valeur est un nombre de combinaisons avec répétitions (défini dans [3, p.27-29]) qui donne le nombre de façons de choisir $2N$ allèles parmi n exemplaires en ayant droit de choisir plusieurs fois le même, soit le nombre d'états possibles pour la population à n'importe quelle génération. Parmi eux, les états de fixation $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, ..., $(0, \ldots, 0, 1)$ seront absorbants, c'est-à-dire qu'une fois que la chaîne atteint un de ces états, elle ne peut le quitter. Ceci est vrai, car nous supposons qu'il n'y a pas de mutations. On supposera que les autres états sont transitoires, ce qui fait qu'il est possible, à partir de n'importe lequel d'entre eux de se rendre à un état de fixation. Ce sera le cas de tous les modèles considérés.

On a alors que $x_i(t)$ converge en probabilité vers la variable aléatoire

$$x_i(\infty) := \begin{cases} 1 & \text{si } \exists\, t \geqslant 0 \text{ t.q. } x_i(t) = 1 \\ 0 & \text{sinon} \end{cases}. \tag{1.6}$$

En effet, pour tout $t \geqslant 0$ et pour tout $\epsilon > 0$, on a

$$P_s\left(\left|x_i(t) - x_i(\infty)\right| > \epsilon\right) \leqslant P_s\left(0 < x_i(t) < 1\right) \leqslant \sum_{\vec{v} \text{ transitoire}} \mathbf{P}_s^{(t)}\left(\vec{x}(0), \vec{v}\right), \tag{1.7}$$

où la somme est effectuée sur tous les vecteurs appartenant à $\{0, 1/2N, 2/2N, \ldots, (2N-1)/2N\}^n$ tels que $\sum_{i=1}^n v_i = 1$, donc sur les états où aucun allèle n'est fixé, c'est-à-dire sur tous les états transitoires. Or, le théorème ergodique stipule que pour tout état transitoire $\vec{v}$, la probabilité $\mathbf{P}_s^{(t)}\left(\vec{x}(0), \vec{v}\right)$ tend vers zéro quand t tend vers l'infini (voir [1, p.48]). Comme la somme est sur un nombre fini d'états, ceci entraîne que

$$\lim_{t\to\infty} P_s\left(\left|x_i(t) - x_i(\infty)\right| > \epsilon\right) = 0, \tag{1.8}$$

ce qui prouve la convergence en probabilité.

Notons que

$$x_i(\infty) = \begin{cases} 1 & \text{avec probabilité } P_s(F_i) \\ 0 & \text{avec probabilité } 1 - P_s(F_i) \end{cases}, \tag{1.9}$$

ce qui a comme conséquence directe que

$$E_s\left[x_i(\infty)\right] = P_s(F_i), \tag{1.10}$$

où $E_s\left[x_i(\infty)\right]$ est l'espérance de $x_i(\infty)$ quand l'intensité de sélection est s. Ainsi, nous avons transformé le calcul de la probabilité de fixation éventuelle de l'allèle A_i en un calcul d'espérance. Nous utilisons ensuite le théorème de convergence dominée tel qu'énoncé dans [8, p.175]. Comme on a la convergence en probabilité (1.8) et que $\left|x_i(t)\right| \leqslant 1$, le théorème nous indique que

$$E_s\left[x_i(\infty)\right] = \lim_{t\to\infty} E_s\left[x_i(t)\right]. \tag{1.11}$$

De plus, si on décompose $x_i(t)$ en somme télescopique, on obtient que

$$\begin{aligned} x_i(t) &= x_i(0) + \left[-x_i(0) + x_i(1)\right] + \left[-x_i(1) + x_i(2)\right] + \ldots + \left[-x_i(t-1) + x_i(t)\right] \\ &= x_i(0) + \sum_{u=0}^{t-1} \Delta x_i(u), \end{aligned} \tag{1.12}$$

où $\Delta x_i(t) := x_i(t+1) - x_i(t)$. En combinant les trois dernières équations, on déduit que

$$P_s(F_i) = E_s\left[x_i(\infty)\right] = \lim_{t\to\infty} E_s\left(x_i(0) + \sum_{u=0}^{t} \Delta x_i(u)\right). \tag{1.13}$$

Comme l'espérance d'une somme finie de variables aléatoires d'espérances finies est la somme de ces espérances, on a

$$P_s(F_i) = \lim_{t\to\infty}\left(x_i(0) + \sum_{u=0}^{t} E_s\left[\Delta x_i(u)\right]\right), \tag{1.14}$$

ce qu'on écrira plus joliment par

$$P_s(F_i) = x_i(0) + \sum_{t\geqslant 0} E_s\left[\Delta x_i(t)\right]. \tag{1.15}$$

1.4 Approximation sous sélection faible

Comme la sélection est faible (s petit et positif), nous allons développer l'équation (1.15) en puissances de s pour ensuite considérer les termes de deuxième ordre comme négligeables.

On récrit donc $P_s(F_i)$ en un polynôme de Taylor de premier degré en s autour de 0, ce qui donne

$$P_s(F_i) = P_0(F_i) + sP_0'(F_i) + o(s), \tag{1.16}$$

où $P_0'(F_i)$ est la dérivée par rapport à s de $P_s(F_i)$ évaluée à $s = 0$ et $o(s)$ est une fonction petit ordre de s, c'est-à-dire telle que $o(s)/s$ tend vers 0 lorsque s tend vers 0. Quant à $P_0(F_i)$, la probabilité de fixation éventuelle de l'allèle A_i en absence de sélection, sa valeur est connue. La population sera éventuellement composée de copies d'un unique gène. Sous neutralité, aucun gène n'a d'avantage, donc la probabilité que le gène fixé soit d'allèle A_i est la fréquence de cet allèle à la génération initiale, $x_i(0)$ (voir [5, p.729]). On reformule donc (1.16) par

$$P_s(F_i) = x_i(0) + sP_0'(F_i) + o(s). \tag{1.17}$$

La dérivée $P_0'(F_i)$ donne l'effet de premier ordre de la sélection. En dérivant de chaque côté de (1.15), on obtient que

$$P_s'(F_i) = \frac{\partial}{\partial s}\sum_{t\geqslant 0} E_s\left[\Delta x_i(t)\right]. \tag{1.18}$$

On veut maintenant affirmer que la dérivée de la série est la série des dérivées, ce qui est vrai si cette dernière converge uniformément dans un voisinage de $s = 0$. Sous l'hypothèse que la matrice de transition $\mathbf{P}_s$ de la chaîne de Markov décrite au début de la section 1.3 ainsi que sa dérivée par rapport à s sont continues en $s = 0$, on fournit une preuve de cette convergence à l'annexe A. Ces hypothèses ne sont pas nécessaires, mais elles sont bien suffisamment générales pour l'étude de modèles de populations génétiques, dans ce sens qu'elles ne sont alors pas contraignantes.

Pour le moment, nous procédons immédiatement pour obtenir

$$P_s'(F_i) = \sum_{t \geqslant 0} E_s' \left[\Delta x_i(t)\right], \tag{1.19}$$

qui, évalué à $s = 0$, donne

$$P_0'(F_i) = \sum_{t \geqslant 0} E_0' \left[\Delta x_i(t)\right]. \tag{1.20}$$

L'espérance $E_s\left[\Delta x_i(t)\right]$ serait difficile à calculer, mais nous ne désirons connaître que sa dérivée évaluée à $s = 0$. Par la propriété de la tour, $E_s\left[\Delta x_i(t)\right] = E_s\left[E_s\left[\Delta x_i(t) \middle| \overrightarrow{x}(t)\right]\right]$, ce qui signifie que

$$E_s\left[\Delta x_i(t)\right] = \sum_{\overrightarrow{v}} P_s\left[\overrightarrow{x}(t) = \overrightarrow{v}\right] E_s\left[\Delta x_i(t) \middle| \overrightarrow{x}(t) = \overrightarrow{v}\right], \tag{1.21}$$

où la somme est effectuée sur tous les vecteurs $\overrightarrow{v} = (v_1, v_2, \ldots, v_n)$ tels que v_i appartient à $\{0, 1/2N, 2/2N, \ldots, 1\}$ pour tout $i \in \{1, \ldots, n\}$ et tels que $\sum_{i=1}^{n} v_i = 1$. On obtient ainsi la dérivée

$$\begin{aligned} E_s'\left[\Delta x_i(t)\right] = \sum_{\overrightarrow{v}} \Big(& P_s'\left[\overrightarrow{x}(t) = \overrightarrow{v}\right] E_s\left[\Delta x_i(t) \middle| \overrightarrow{x}(t) = \overrightarrow{v}\right] \\ & + P_s\left[\overrightarrow{x}(t) = \overrightarrow{v}\right] E_s'\left[\Delta x_i(t) \middle| \overrightarrow{x}(t) = \overrightarrow{v}\right] \Big), \end{aligned} \tag{1.22}$$

que l'on peut évaluer à $s = 0$ pour obtenir

$$\begin{aligned} E_0'\left[\Delta x_i(t)\right] = \sum_{\overrightarrow{v}} \Big(& P_0'\left[\overrightarrow{x}(t) = \overrightarrow{v}\right] E_0\left[\Delta x_i(t) \middle| \overrightarrow{x}(t) = \overrightarrow{v}\right] \\ & + P_0\left[\overrightarrow{x}(t) = \overrightarrow{v}\right] E_0'\left[\Delta x_i(t) \middle| \overrightarrow{x}(t) = \overrightarrow{v}\right] \Big). \end{aligned} \tag{1.23}$$

Or, $E_0\left[\Delta x_i(t) \middle| \overrightarrow{x}(t) = \overrightarrow{v}\right]$ sera toujours nulle, car aucun allèle n'est avantagé dans un modèle neutre. Ceci nous permet de simplifier (1.23) par

$$E_0'\left[\Delta x_i(t)\right] = \sum_{\overrightarrow{v}} P_0\left[\overrightarrow{x}(t) = \overrightarrow{v}\right] E_0'\left[\Delta x_i(t) \middle| \overrightarrow{x}(t) = \overrightarrow{v}\right]. \tag{1.24}$$

Inversement à ce qu'on a fait précédemment, on peut retransformer la partie droite de l'équation, ce qui donne

$$E_0'\left[\Delta x_i(t)\right] = E_0\left[E_0'\left[\Delta x_i(t) \middle| \overrightarrow{x}(t)\right]\right]. \tag{1.25}$$

Remplaçons cette valeur dans (1.20), puis utilisons le résultat dans (1.17) pour obtenir

$$P_s(F_i) = x_i(0) + s\sum_{t\geqslant 0} E_0\left[E_0'\left[\Delta x_i(t)\middle|\vec{x}(t)\right]\right] + o(s). \tag{1.26}$$

L'espérance $E_s\left[\Delta x_i(t)\middle|\vec{x}(t)\right]$ est relativement simple à calculer pour chacun des modèles, car elle ne concerne que deux générations consécutives. On pourra ensuite la dériver et l'évaluer à $s = 0$. Pour un modèle donné, cette variable, jumelée des probabilités de non coalescence (voir section 1.6.2), constituent la totalité de l'information nécessaire au calcul de la probabilité de fixation d'un allele. Ceci nous mène au premier résultat.

Résultat 1.1

Considérant un modèle de population finie à générations discrètes sans chevauchement avec sélection faible, la probabilité de fixation d'un allèle A_i est donnée par

$$P_s(F_i) = x_i(0) + s\sum_{t\geqslant 0} E_0\left[\phi_i(t)\right] + o(s), \tag{1.27}$$

où $\phi_i(t)$ est la dérivée de l'espérance du changement de fréquence de l'allèle A_i de la génération t à la génération $t+1$ par rapport à l'intensité de la sélection s évaluée à $s = 0$, i.e.

$$\phi_i(t) = E_0'\left[\Delta x_i(t)\middle|\vec{x}(t)\right]. \tag{1.28}$$

Ceci est conditionnel à ce que la matrice de transition $\mathbf{P}_s$ de la chaîne de Markov représentant la population et sa dérivée par rapport à s soient continues en $s = 0$.

Dans tous les modèles que nous considérerons dans ce texte, $\phi_i(t)$ est un polynôme multivarié en $\{x_1(t), x_2(t), \ldots, x_n(t)\}$ et la complexité du calcul de son espérance dépend du degré de ce polynôme, qui sera ici de degré inférieur ou égal à 3. Ainsi, calculer $E_0\left[\phi_i(t)\right]$ revient à calculer l'espérance de chacun des termes possibles de $\phi_i(t)$, des termes de degré inférieur ou égal à 3 en $\{x_1(t), x_2(t), \ldots, x_n(t)\}$. La section suivante explique comment la théorie de la coalescence peut nous aider dans cette tâche.

1.5 Espérance d'un produit de fréquences alléliques

Montrons comment on peut calculer $E_0[x_i(t)x_j(t)x_k(t)]$ pour un triplet (i,j,k) quelconque de $\{1,2,\ldots,n\}^3$. On peut écrire cette espérance comme

$$E_0[x_i(t)x_j(t)x_k(t)] = \sum_{\vec{v}} P_0\left(\vec{x}(t) = \vec{v}\right) v_i v_j v_k, \tag{1.29}$$

où la somme est effectuée sur tous les vecteurs $\vec{v} = (v_1, v_2, \ldots, v_n)$ appartenant à l'ensemble $\{0, 1/2N, 2/2N, \ldots, 1\}^n$ tels que $\sum_{i=1}^n v_i = 1$. Comme v_l est la probabilité en absence de sélection de tirer au hasard un gène d'allèle A_l à la génération t sachant que $\vec{x}(t) = \vec{v}$, probabilité que nous noterons $P_0\left(A_l(t)\middle|\vec{x}(t) = \vec{v}\right)$, on a

$$\begin{aligned} &E_0[x_i(t)x_j(t)x_k(t)] \\ &= \sum_{\vec{v}} P_0(\vec{x}(t) = \vec{v}) P_0\left(A_i(t)\middle|\vec{x}(t) = \vec{v}\right) P_0\left(A_j(t)\middle|\vec{x}(t) = \vec{v}\right) P_0\left(A_k(t)\middle|\vec{x}(t) = \vec{v}\right). \end{aligned} \tag{1.30}$$

Comme les tirages sont conditionnellement indépendants, ceci est équivalent à

$$E_0[x_i(t)x_j(t)x_k(t)] = \sum_{\vec{v}} P_0\left(\vec{x}(t) = \vec{v}\right) P_0\left(A_iA_jA_k(t)\middle|\vec{x}(t) = \vec{v}\right), \tag{1.31}$$

où $P_0\left(A_iA_jA_k(t)\middle|\vec{x}(t) = \vec{v}\right)$ est la probabilité sous neutralité, lors d'un tirage avec remise de trois gènes à la génération t, de piger dans l'ordre des gènes d'allèles A_i, A_j, puis A_k, sachant que $\vec{x}(t) = \vec{v}$. On remarque que ceci est un conditionnement sur $\vec{x}(t)$ de la probabilité de l'événement en question, et donc

$$E_0[x_i(t)x_j(t)x_k(t)] = P_0(A_iA_jA_k(t)). \tag{1.32}$$

Nous allons maintenant conditionner sur le nombre de gènes distincts parmi les trois choisis au hasard et sur leur nombre d'ancêtres distincts à la génération 0.

Considérons l'expérience qui consiste à *choisir au hasard avec remise c gènes à la génération t*, et définissons l'événement

$$H_{ba}^c(t) : \text{Obtenir } b \text{ gènes distincts ayant } a \text{ ancêtres distincts à la génération 0,}$$

pour $1 \leqslant a \leqslant b \leqslant c = 3$. Les probabilités d'occurence de ces événements dépendent de la sélection, mais l'utilisation que l'on en fait dans ce texte est restreinte au modèle neutre. Pour cette raison,

l'intensité de sélection s n'est pas mentionnée dans la notation. Nous aurons alors

$$E_0\left[x_i(t)x_j(t)x_k(t)\right] = \sum_{1\leqslant a\leqslant b\leqslant 3} P_0\left(H_{ba}^3(t)\right) P_0\left(A_iA_jA_k(t)\middle|H_{ba}^3(t)\right). \tag{1.33}$$

L'événement *piger trois gènes d'allèles respectifs* A_i, A_j *et* A_k a lieu si, et seulement si, les ancêtres respectifs des gènes sont d'allèles A_i, A_j et A_k. Sous neutralité, chaque gène présent à la génération 0 a la même chance d'être l'ancêtre d'un gène choisi au hasard à la génération t. Sachant le nombre d'ancêtres à la génération 0, le calcul de la probabilité de l'événement ci-dessus se fait facilement.

Le raisonnement effectué de (1.29) à (1.33) pourrait être fait pour des termes de n'importe quel ordre. Entre autres, pour des termes de degré 2, on poserait $c = 2$ pour avoir

$$E_0\left[x_i(t)x_j(t)\right] = \sum_{1\leqslant a\leqslant b\leqslant 2} P_0\left(H_{ba}^2(t)\right) P_0\left(A_iA_j(t)\middle|H_{ba}^2(t)\right). \tag{1.34}$$

Pour des termes de degré 1, on poserait $c = 1$ et on trouverait

$$E_0\left[x_i(t)\right] = P_0\left(H_{11}^1(t)\right) P_0\left(A_i(t)\middle|H_{11}^1(t)\right). \tag{1.35}$$

Afin d'utiliser ces formules pour le calcul de l'espérance d'un produit de fréquences alléliques, nous devons calculer les probabilités des événements $H_{ba}^c(t)$, ce que nous ferons à la section suivante.

1.6 Probabilités pour le nombre d'ancêtres d'un échantillon de gènes

Définissons q_{cb} comme la probabilité, lors d'un choix au hasard avec remise de c gènes d'une même génération, de piger b gènes distincts. Définissons aussi $p_{ba}^{(t)}$, la probabilité sous neutralité, que b gènes distincts pigés au hasard à la génération t aient a ancêtres à la génération 0. En utilisant ceci, on récrit

$$P_0\left(H_{ba}^c(t)\right) = q_{cb}p_{ba}^{(t)}. \tag{1.36}$$

Nous montrons ci-dessous comment déterminer chacune des probabilités présentes à droite de cette équation.

1.6.1 Probabilités pour le nombre de gènes distincts dans un échantillon

Nous calculons la probabilité q_{cb}, c'est-à-dire la probabilité, quand on choisit c gènes au hasard avec remise, d'en piger b distincts. Ce calcul combinatoire ne dépend que de la taille de la population, qui est de $2N$. Donnons d'abord les résultats :

$$q_{11} = 1, \tag{1.37}$$

$$q_{22} = \frac{2N-1}{2N}, \tag{1.38}$$

$$q_{21} = \frac{1}{2N}, \tag{1.39}$$

$$q_{33} = \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right), \tag{1.40}$$

$$q_{32} = 3\left(\frac{1}{2N}\right)\left(\frac{2N-1}{2N}\right), \tag{1.41}$$

$$q_{31} = \left(\frac{1}{2N}\right)^2. \tag{1.42}$$

La valeur de q_{11} est 1, car un unique gène est nécessairement distinct.

Passons aux cas où $c = 2$. Pour piger deux gènes distincts dans un double tirage avec remise, il faut simplement que le deuxième soit différent du premier. Donc, q_{22} prend comme valeur $(2N-1)/2N$. D'autre part, q_{21} dénote la probabilité complémentaire de q_{22}. La somme de leurs valeurs est donc de 1, ce qui implique que $q_{21} = 1/2N$.

Finalement, considérons le tirage de trois gènes, donc lorsque $c = 3$. Pour piger trois gènes distincts dans un tirage avec remise, il faut simplement que le deuxième soit différent du premier, puis que le troisième soit différent des deux premiers. On obtient donc que

$$q_{33} = \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right). \tag{1.43}$$

La probabilité q_{32} est celle de piger deux gènes distincts en trois tirages consécutifs avec remise. Il y a trois possibilités pour l'ordre des deux gènes identiques : les deux premiers, le premier et le dernier, ou les deux derniers. Ces deux gènes sont les mêmes avec probabilité $1/2N$ et l'autre en est différent avec probabilité $(2N-1)/2N$. On a donc

$$q_{32} = 3\left(\frac{1}{2N}\right)\left(\frac{2N-1}{2N}\right). \tag{1.44}$$

Comme les trois gènes sont identiques si et seulement si le deuxième et le troisième sont identiques au premier, on sait que $q_{31} = (1/2N)^2$.

1.6.2 Probabilités pour le nombre d'ancêtres de gènes distincts

Rappelons que $p_{ba}^{(t)}$ est la probabilité, sous neutralité, que b gènes distincts pigés au hasard en t aient a ancêtres à la génération 0. Pour simplifier la notation, on utilisera p_{ba} au lieu de $p_{ba}^{(1)}$, c'est-à-dire la probabilité que b gènes aient a géniteurs à la génération précédente. Les probabilités p_{21}, p_{31} et p_{32} seront appelées les *probabilités de coalescence*, tandis que p_{22} et p_{33} seront les *probabilités de non coalescence*.

Il est clair que, pour tout $t \geqslant 1$, on a

$$p_{11}^{(t)} = 1. \tag{1.45}$$

De plus, remarquons que $p_{22}^{(t)} = p_{22}^{t}$ et $p_{33}^{(t)} = p_{33}^{t}$ pour la raison suivante. Pour que m gènes aient m ancêtres t générations en arrière, ils doivent avoir m géniteurs, qui doivent eux-mêmes avoir m géniteurs, et ainsi de suite pour t générations successives, ce qui représente t événements indépendants, chacun de probabilité p_{mm}. Les deux notations seront utilisées selon le contexte.

Les probabilités définies ci-dessus ont de toute évidence plusieurs liens entre elles. Voici les deux premières relations auxquelles on peut penser :

$$p_{21}^{(t)} + p_{22}^{(t)} = 1, \tag{1.46}$$

$$p_{31}^{(t)} + p_{32}^{(t)} + p_{33}^{(t)} = 1. \tag{1.47}$$

On n'a fait qu'énumérer le nombre d'ancêtres possibles de deux gènes, 1 ou 2, et de trois gènes, 1, 2 ou 3. Le matériel ancestral à la génération t sera dans un de ces états avec probabilité 1.

Nous allons déduire une troisième relation à partir du principe d'inclusion-exclusion. En considérant l'arbre ancestral de trois gènes distincts pigés au hasard à la génération t et numérotés 1, 2, 3, on définit les événements suivants :

$E_1(t)$: Les gènes 1 et 2 ont un ancêtre commun à la génération 0;

$E_2(t)$: Les gènes 1 et 3 ont un ancêtre commun à la génération 0;

$E_3(t)$: Les gènes 2 et 3 ont un ancêtre commun à la génération 0.

En appliquant le principe d'inclusion-exclusion, on obtient que

$$\begin{aligned} &1 - P\left[E_1^C(t) \cap E_2^C(t) \cap E_3^C(t)\right] \\ &\quad = \sum_{1 \leqslant i \leqslant 3} P\left[E_i(t)\right] - \sum_{1 \leqslant i < j \leqslant 3} P\left[E_i(t) \cap E_j(t)\right] + P\left[E_1(t) \cap E_2(t) \cap E_3(t)\right]. \end{aligned} \tag{1.48}$$

Or, $P\left[E_1^C(t) \cap E_2^C(t) \cap E_3^C(t)\right]$ est la probabilité qu'il n'y ait aucune coalescence en t généra-

tions, c'est-à-dire $p_{33}^{(t)}$. De plus, pour $1 \leqslant i \leqslant 3$, $P[E_i(t)]$ est par définition $p_{21}^{(t)}$. Finalement, pour tous $1 \leqslant i < j \leqslant 3$, $P[E_i(t) \cap E_j(t)]$ est la probabilité que les trois gènes coalescent en t générations, tout comme $P[E_1(t) \cap E_2(t) \cap E_3(t)]$, ce qu'on a déjà défini comme $p_{31}^{(t)}$. Ainsi, on obtient que

$$1 - p_{33}^{(t)} = 3p_{21}^{(t)} - 2p_{31}^{(t)}. \tag{1.49}$$

Pour un $t \geqslant 1$ quelconque, on a trois relations entre les cinq probabilités $p_{21}^{(t)}$, $p_{22}^{(t)}$, $p_{31}^{(t)}$, $p_{32}^{(t)}$ et $p_{33}^{(t)}$. Ceci nous permet d'exprimer trois d'entre elles en fonction des deux autres. Il est logique de garder les deux plus simples à calculer. Comme les identités $p_{22}^{(t)} = p_{22}^t$ et $p_{33}^{(t)} = p_{33}^t$ donnent les valeurs de $p_{22}^{(t)}$ et $p_{33}^{(t)}$ à partir des probabilités de non coalescence p_{22} et p_{33}, il est intéressant de conserver uniquement ces deux dernières. Par (1.46), on a que

$$p_{21}^{(t)} = 1 - p_{22}^t. \tag{1.50}$$

Par (1.49) et (1.50), on peut dire que

$$\begin{aligned} p_{31}^{(t)} &= \frac{3p_{21}^{(t)} + p_{33}^{(t)} - 1}{2}, \\ &= \frac{3\left(1 - p_{22}^{(t)}\right) + p_{33}^{(t)} - 1}{2}, \\ &= 1 - \tfrac{3}{2}p_{22}^t + \tfrac{1}{2}p_{33}^t. \end{aligned} \tag{1.51}$$

Puis, par (1.47) et (1.51), on sait que

$$\begin{aligned} p_{32}^{(t)} &= 1 - p_{33}^{(t)} - p_{31}^{(t)}, \\ &= 1 - p_{33}^{(t)} - \left(1 - \tfrac{3}{2}p_{22}^t + \tfrac{1}{2}p_{33}^t\right), \\ &= \tfrac{3}{2}p_{22}^t - \tfrac{3}{2}p_{33}^t. \end{aligned} \tag{1.52}$$

Remarquons que les équations (1.50), (1.51) et (1.52) réduisent le calcul des probabilités des différents états ancestraux, probabilités concernant plusieurs générations, à deux probabilités ne concernant que deux générations successives. Les valeurs de p_{22} et p_{33} dépendent du modèle considéré et devront être calculées pour chacun d'entre eux.

Enfin, énonçons une dernière identité qui permettra parfois d'accélérer les calculs. Par (1.49)

et (1.47), on obtient que

$$\begin{aligned} 3\left(p_{21}^{(t)} - p_{31}^{(t)}\right) &= \left(3p_{21}^{(t)} - 2p_{31}^{(t)}\right) - p_{31}^{(t)}, \\ &= 1 - p_{33}^{(t)} - p_{31}^{(t)}, \\ &= p_{32}^{(t)}. \end{aligned} \tag{1.53}$$

À toutes fins pratiques, nous savons maintenant calculer les probabilités $P_0(H_{ba}^c(t))$ par l'équation (1.36). Nous ne les expliciterons pas immédiatement, car elles sont plus faciles à manipuler sous cette forme.

1.7 Calcul de l'espérance d'un produit de fréquences alléliques

À partir de (1.35), (1.34) et (1.33), nous pouvons calculer l'espérance sous neutralité des produits de fréquences de trois allèles ou moins à la génération $t \geqslant 0$, avec répétitions permises. Sous les restrictions présentées à la fin de la section 1.4, ceci couvre tous les termes possibles de la variable aléatoire $\phi_i(t)$.

Pour tous les calculs, nous utiliserons un même argument : sous neutralité, sachant qu'un échantillon de b gènes choisis au hasard à la génération t ont a ancêtres à la génération 0, tout groupe de gènes de cette taille de la génération 0 a chance égale d'être le groupe d'ancêtres de ces b gènes. Cette proproposition est démontrée à l'annexe B et on y réfèrera parfois dans les pages qui suivent par le terme d'*égalité des chances.*

Pour référence ultérieure, nous énonçons d'abord le résultat 1.2.

Résultat 1.2

L'espérance sous neutralité des produits de fréquences alléliques de degré inférieur ou égal à trois à la génération $t \geqslant 0$ en fonction des fréquences alléliques initiales et des probabilités de non coalescence est donnée par les expressions suivantes :

$$E_0\left[x_i(t)\right] = x_i(0), \tag{1.54}$$

$$E_0\left[x_i^2(t)\right] = x_i(0)\left(1 - p_{22}^t\left(1 - x_i(0)\right)\right), \tag{1.55}$$

$$E_0\left[x_i(t)x_j(t)\right] = p_{22}^t x_i(0)x_j(0), \text{ pour } i \neq j, \tag{1.56}$$

$$E_0\left[x_i^3(t)\right] = x_i(0)\left[1 - \frac{1 - x_i(0)}{2}\left(3p_{22}^t - p_{33}^t\left(1 - 2x_i(0)\right)\right)\right], \tag{1.57}$$

$$E_0\left[x_i^2(t)x_j(t)\right] = \frac{x_i(0)x_j(0)}{2}\left(p_{22}^t - p_{33}^t\left(1 - 2x_i(0)\right)\right), \text{ pour } i \neq j, \tag{1.58}$$

$$E_0\left[x_i(t)x_j(t)x_k(t)\right] = p_{33}^t x_i(0)x_j(0)x_k(0)h_j h_k, \text{ pour } i \neq j \neq k \neq i. \tag{1.59}$$

Les fréquence alléliques à la génération initiale sont des constantes connues, tandis que p_{22} et p_{33} sont les probabilités, sous neutralité, que deux et trois gènes distincts pris au hasard dans une génération aient des géniteurs distincts.

Déduisons maintenant ces expressions dans l'ordre. Reprenons d'abord (1.35) qui dit que

$$E_0\left[x_i(t)\right] = P_0\left(H_{11}^1(t)\right)P_0\left(A_i(t)\middle|H_{11}^1(t)\right). \tag{1.60}$$

Par définition, $P_0\left(H_{11}^1(t)\right)$ a une valeur de 1. D'autre part $P_0\left(A_i\middle|H_{11}^1(t)\right)$ est la probabilité, sous neutralité, de choisir au hasard à la génération t un gène A_i. Cette probabilité est égale à celle que l'ancêtre à la génération 0 du gène pigé en t soit d'allèle A_i. Comme la probabilité est évaluée en absence de sélection, les gènes présents à la génération initiale ont des chances égales d'être l'ancêtre du gène pigé. Ainsi, on a simplement

$$E_0\left[x_i(t)\right] = x_i(0). \tag{1.61}$$

Ceci n'est pas surprenant. Sous neutralité, on ne s'attend en moyenne ni a une augmentation, ni a une diminution de la fréquence de l'allèle A_i dans la population.

Pour calculer l'espérance de $x_i^2(t)$, nous invoquons (1.34) qui stipule que

$$E_0\left[x_i(t)x_j(t)\right] = \sum_{1\leqslant a\leqslant b\leqslant 2} P_0\left(H_{ba}^2(t)\right) P_0\left(A_iA_j(t)\middle|H_{ba}^2(t)\right). \tag{1.62}$$

Dans le cas présent, on considère $i = j$. Si on développe la somme ci-dessus, on trouve que

$$\begin{aligned} E_0\left[x_i^2(t)\right] &= P_0\left(H_{11}^2(t)\right) P_0\left(A_iA_i(t)\middle|H_{11}^2(t)\right) \\ &\quad + P_0\left(H_{21}^2(t)\right) P_0\left(A_iA_i(t)\middle|H_{21}^2(t)\right) + P_0\left(H_{22}^2(t)\right) P_0\left(A_iA_i(t)\middle|H_{22}^2(t)\right). \end{aligned} \tag{1.63}$$

Le deuxième facteur du premier terme est la probabilité que l'ancêtre de l'unique gène pigé doublement soit d'allèle A_i. Par égalité des chances, c'est $x_i(0)$. De son côté, $P_0\left(A_iA_i(t)\middle|H_{21}^2(t)\right)$ est la probabilité que l'unique ancêtre des deux gènes distincts pigés soit d'allèle A_i, ce qui a la même valeur. Finalement, $P_0\left(A_iA_i(t)\middle|H_{22}^2(t)\right)$ est la probabilité que les deux ancêtres des deux gènes pigés soient d'allèle A_i sachant qu'ils sont distincts. Toujours par égalité des chances, sa valeur est

$$\frac{\binom{2Nx_i(0)}{2}}{\binom{2N}{2}} = x_i(0)\frac{2Nx_i(0)-1}{2N-1}. \tag{1.64}$$

En remplaçant ces valeurs ainsi que celles trouvées à la section 1.6, on obtient que

$$E_0\left[x_i^2(t)\right] = \frac{1}{2N}p_{11}^{(t)}x_i(0) + \frac{2N-1}{2N}p_{21}^{(t)}x_i(0) + \frac{2N-1}{2N}p_{22}^{(t)}x_i(0)\frac{2Nx_i(0)-1}{2N-1}. \tag{1.65}$$

En remplaçant $p_{11}^{(t)}$ par 1, $p_{22}^{(t)}$ par p_{22}^t et $p_{21}^{(t)}$ par (1.50), on a

$$E_0\left[x_i^2(t)\right] = \frac{1}{2N}\left[x_i(0) + \left(1-p_{22}^t\right)x_i(0)\left(2N-1\right) + p_{22}^t x_i(0)\left(2Nx_i(0)-1\right)\right], \tag{1.66}$$

qui se simplifie pour donner le résultat

$$E_0\left[x_i^2(t)\right] = x_i(0)\left(1-p_{22}^t\left(1-x_i(0)\right)\right). \tag{1.67}$$

Passons maintenant à l'espérance de $x_i(t)x_j(t)$, où $i \neq j$. Toujours en développant (1.34), on obtient que

$$\begin{aligned} E_0\left[x_i(t)x_j(t)\right] &= P_0\left(H_{11}^2(t)\right) P_0\left(A_iA_j(t)\middle|H_{11}^2(t)\right) + P_0\left(H_{21}^2(t)\right) P_0\left(A_iA_j(t)\middle|H_{21}^2(t)\right) \\ &\quad + P_0\left(H_{22}^2(t)\right) P_0\left(A_iA_j(t)\middle|H_{22}^2(t)\right). \end{aligned} \tag{1.68}$$

On remarque ici que les probabilités $P_0\left(A_iA_j(t)\middle|H_{11}^2(t)\right)$ et $P_0\left(A_iA_j(t)\middle|H_{21}^2(t)\right)$ sont nulles. En effet, si les gènes pigés ont un ancêtre commun, ils sont nécessairement du même allèle. La

probabilité $P_0\left(A_iA_j(t)\middle|H_{22}^2(t)\right)$, quant à elle, est celle avec laquelle les deux ancêtres distincts des deux gènes pigés à la génération 0 sont d'allèles respectifs A_i et A_j. Toujours en exploitant l'égalité des chances due à la neutralité, cette probabilité vaut

$$x_i(0)\frac{2Nx_j(0)}{(2N-1)}. \tag{1.69}$$

En remplaçant ces trois valeurs ainsi que celles trouvées à la section précédente, nous trouvons que

$$E_0\left[x_i(t)x_j(t)\right] = \left(\frac{2N-1}{2N}\right)p_{22}^{(t)}x_i(0)x_j(0)\left(\frac{2N}{2N-1}\right), \tag{1.70}$$

que nous écrivons plus simplement comme

$$E_0\left[x_i(t)x_j(t)\right] = p_{22}^t x_i(0)x_j(0). \tag{1.71}$$

Trouvons maintenant $E_0\left[x_i(t)x_j(t)x_k(t)\right]$. On a montré à la section 1.5 que

$$E_0\left[x_i^3(t)\right] = \sum_{1\leqslant a\leqslant b\leqslant 3} P_0\left(H_{ba}^3(t)\right)P_0\left(A_iA_jA_k(t)\middle|H_{ba}^3(t)\right). \tag{1.72}$$

Dans le cas présent, i, j et k sont tous égaux. Remplaçons j et k par i puis développons la somme ci-dessus. On obtient alors

$$\begin{aligned}E_0\left[x_i^3(t)\right] = {} & P_0\left(H_{11}^3(t)\right)P_0\left(A_iA_iA_i(t)\middle|H_{11}^3(t)\right) + P_0\left(H_{21}^3(t)\right)P_0\left(A_iA_iA_i(t)\middle|H_{21}^3(t)\right)\\ & + P_0\left(H_{22}^3(t)\right)P_0\left(A_iA_iA_i(t)\middle|H_{22}^3(t)\right) + P_0\left(H_{31}^3(t)\right)P_0\left(A_iA_iA_i(t)\middle|H_{31}^3(t)\right)\\ & + P_0\left(H_{32}^3(t)\right)P_0\left(A_iA_iA_i(t)\middle|H_{32}^3(t)\right) + P_0\left(H_{33}^3(t)\right)P_0\left(A_iA_iA_i(t)\middle|H_{33}^3(t)\right).\end{aligned} \tag{1.73}$$

L'événement qui nous intéresse ici n'a lieu que si les ancêtres à la génération 0 des gènes pigés à la génération t sont tous d'allèle A_i. Étant donné $H_{11}^3(t)$, $H_{21}^3(t)$ ou $H_{31}^3(t)$, sa probabilité vaut $x_i(0)$, puisque, alors, les gènes pigés ont tous le même ancêtre à la génération 0. Étant donné $H_{22}^3(t)$ ou $H_{32}^3(t)$, il y a deux ancêtres distincts à la génération 0 et la probabilité conditionnelle est donc

$$\frac{\binom{2Nx_i(0)}{2}}{\binom{2N}{2}} = x_i(0)\left(\frac{2Nx_i(0)-1}{2N-1}\right). \tag{1.74}$$

Selon la même logique, la dernière probabilité, sous l'événement conditionnel $H_{33}^3(t)$ qui assure que les gènes ont trois ancêtres distincts à la génération 0, est

$$\frac{\binom{2Nx_i(0)}{3}}{\binom{2N}{3}} = x_i(0)\left(\frac{2Nx_i(0)-1}{2N-1}\right)\left(\frac{2Nx_i(0)-2}{2N-2}\right). \tag{1.75}$$

Dans tous les cas, on fait appel à l'égalité des chances dans le modèle neutre. En utilisant les valeurs de la section précédente, on trouve

$$\begin{aligned}
E_0\left[x_i^3(t)\right] = &\left(\frac{1}{2N}\right)\left(\frac{1}{2N}\right)p_{11}^{(t)}x_i(0) \\
&+3\left(\frac{1}{2N}\right)\left(\frac{2N-1}{2N}\right)p_{21}^{(t)}x_i(0) \\
&+3\left(\frac{1}{2N}\right)\left(\frac{2N-1}{2N}\right)p_{22}^{(t)}x_i(0)\left(\frac{2Nx_i(0)-1}{2N-1}\right) \\
&+\left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)p_{31}^{(t)}x_i(0) \\
&+\left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)p_{32}^{(t)}x_i(0)\left(\frac{2Nx_i(0)-1}{2N-1}\right) \\
&+\left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)p_{33}^{(t)}x_i(0)\left(\frac{2Nx_i(0)-1}{2N-1}\right)\left(\frac{2Nx_i(0)-2}{2N-2}\right).
\end{aligned} \tag{1.76}$$

Par une mise en évidence, on obtient

$$\begin{aligned}
E_0\left[x_i^3(t)\right] = \frac{x_i(0)}{(2N)^2}\Big[&p_{11}^{(t)} + 3(2N-1)p_{21}^{(t)} + 3p_{22}^{(t)}(2Nx_i(0)-1) \\
&+(2N-1)(2N-2)p_{31}^{(t)} + (2N-2)p_{32}^{(t)}(2Nx_i(0)-1) \\
&+p_{33}^{(t)}(2Nx_i(0)-1)(2Nx_i(0)-2)\Big].
\end{aligned} \tag{1.77}$$

En utilisant les identités (1.45), (1.50), (1.51) et (1.52), la dernière équation devient

$$\begin{aligned}
E_0\left[x_i^3(t)\right] = \frac{x_i(0)}{(2N)^2}\Big[&1+3\left(1-p_{22}^t\right)(2N-1)+\left(1-\tfrac{3}{2}p_{22}^t+\tfrac{1}{2}p_{33}^t\right)(2N-1)(2N-2) \\
&+3p_{22}^t(2Nx_i(0)-1)+\tfrac{3}{2}\left(p_{22}^t-p_{33}^t\right)(2Nx_i(0)-1)(2N-2) \\
&+p_{33}^t(2Nx_i(0)-1)(2Nx_i(0)-2)\Big],
\end{aligned} \tag{1.78}$$

ce qui se simplifie pour donner

$$E_0\left[x_i^3(t)\right] = x_i(0)\left[1-\frac{1-x_i(0)}{2}\left(3p_{22}^t-p_{33}^t\left(1-2x_i(0)\right)\right)\right]. \tag{1.79}$$

Reprenons encore une fois (1.33), c'est-à-dire

$$E_0\left[x_i(t)x_j(t)x_k(t)\right] = \sum_{1\leqslant a\leqslant b\leqslant 3} P_0\left(H_{ba}^3(t)\right) P_0\left(A_iA_jA_k(t)\middle|H_{ba}^3(t)\right). \tag{1.80}$$

En posant $k = i$, l'équation ci-dessus nous donne

$$\begin{aligned} E_0\left[x_i^2(t)x_j(t)\right] = {} & P_0\left(H_{11}^3(t)\right) P_0\left(A_iA_jA_i(t)\middle|H_{11}^3(t)\right) + P_0\left(H_{21}^3(t)\right) P_0\left(A_iA_jA_i(t)\middle|H_{21}^3(t)\right) \\ & + P_0\left(H_{22}^3(t)\right) P_0\left(A_iA_jA_i(t)\middle|H_{22}^3(t)\right) + P_0\left(H_{31}^3(t)\right) P_0\left(A_iA_jA_i(t)\middle|H_{31}^3(t)\right) \\ & + P_0\left(H_{32}^3(t)\right) P_0\left(A_iA_jA_i(t)\middle|H_{32}^3(t)\right) + P_0\left(H_{33}^3(t)\right) P_0\left(A_iA_jA_i(t)\middle|H_{33}^3(t)\right). \end{aligned} \tag{1.81}$$

Pour que les gènes pigés à la génération t soient d'allèles A_i, A_j et A_i respectivement, il faut qu'ils aient au moins deux ancêtres t générations en arrière. Pour cette raison, les premier, deuxième et quatrième termes du membre de droite de la dernière équation sont nuls. Étant donné $H_{22}^3(t)$, c'est-à-dire sachant qu'on n'a pigé que deux gènes différents à la génération t, qui ont deux ancêtres distincts à la génération 0, il faut absolument que les gènes identiques soient le premier et le dernier, ce qui est de probabilité 1/3. Ainsi, on sait que

$$P_0\left(A_iA_jA_i(t)\middle|H_{22}^3(t)\right) = \frac{1}{3}P_0\left(A_iA_j(t)\middle|H_{22}^2(t)\right), \tag{1.82}$$

mais on a déjà calculé $P_0\left(A_iA_j(t)\middle|H_{22}^2(t)\right)$ en (1.70), ce qui nous permet de dire que

$$P_0\left(A_iA_jA_i(t)\middle|H_{22}^3(t)\right) = \frac{1}{3}x_i(0)x_j(0)\left(\frac{2N}{N-1}\right). \tag{1.83}$$

La valeur de $P_0\left(A_iA_jA_i(t)\middle|H_{32}^3(t)\right)$ est la même. Sachant qu'il y a eu coalescence entre deux des trois gènes distincts pigés en t avant la génération 0, il faut que la coalescence ait lieu entre le premier et le troisième gène pour que l'événement considéré puisse avoir lieu. Finalement, $P_0\left(A_iA_jA_i(t)\middle|H_{33}^3(t)\right)$ est la probabilité de l'événement, sachant que les ancêtres des trois gènes à la génération 0 sont distincts. Comme la probabilité est calculée en absence de sélection, la probabilité que l'ancêtre du premier gène soit d'allèle A_i est $x_i(0)$, la probabilité que l'ancêtre du deuxième gène soit d'allèle A_j sachant qu'il est distinct du premier est $2Nx_j(0)/(2N-1)$, et la probabilité que l'ancêtre du troisième gène soit d'allèle A_i sachant qu'il est distinct des deux premiers est $(2Nx_i(0)-1)/(2N-2)$. Ainsi, on trouve

$$P_0\left(A_iA_jA_i(t)\middle|H_{33}^3(t)\right) = x_i(0)\left(\frac{2Nx_j(0)}{2N-1}\right)\left(\frac{2Nx_i(0)-1}{2N-2}\right). \tag{1.84}$$

On peut maintenant récrire (1.81) comme

$$\begin{aligned} E_0\left[x_i^2(t)x_j(t)\right] &= 3\left(\frac{1}{2N}\right)\left(\frac{2N-1}{2N}\right)p_{22}^{(t)}\left(\frac{1}{3}\right)x_i(0)x_j(0)\left(\frac{2N}{2N-1}\right) \\ &+ \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)p_{32}^{(t)}\left(\frac{1}{3}\right)x_i(0)x_j(0)\left(\frac{2N}{2N-1}\right) \\ &+ \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)p_{33}^{(t)}x_i(0)\left(\frac{2Nx_j(0)}{2N-1}\right)\left(\frac{2Nx_i(0)-1}{2N-2}\right). \end{aligned} \tag{1.85}$$

Mettons en évidence $x_i(0)x_j(0)/(2N)$, ce qui nous donne

$$E_0\left[x_i^2(t)x_j(t)\right] = \frac{x_i(0)x_j(0)}{2N}\left[p_{22}^{(t)} + \frac{p_{32}^{(t)}}{3}(2N-2) + p_{33}^{(t)}(2Nx_i(0)-1)\right]. \tag{1.86}$$

En utilisant (1.52), on a

$$E_0\left[x_i^2(t)x_j(t)\right] = \frac{x_i(0)x_j(0)}{2N}\left[p_{22}^{t} + \frac{p_{22}^{t}-p_{33}^{t}}{2}(2N-2) + p_{33}^{t}(2Nx_i(0)-1)\right], \tag{1.87}$$

ce qui se réduit à

$$E_0\left[x_i^2(t)x_j(t)\right] = \frac{x_i(0)x_j(0)}{2}\left(p_{22}^{t} - p_{33}^{t}(1-2x_i(0))\right). \tag{1.88}$$

Il ne reste que l'espérance de $x_i(t)x_j(t)x_k(t)$ quand i, j et k sont tous différents. Cette fois, on utilise directement l'équation (1.33) sous sa forme générale, soit

$$\begin{aligned} E_0\left[x_i(t)x_j(t)x_k(t)\right] &= P_0\left(H_{11}^3(t)\right)P_0\left(A_iA_jA_k(t)\middle|H_{11}^3(t)\right) \\ &+ P_0\left(H_{21}^3(t)\right)P_0\left(A_iA_jA_k(t)\middle|H_{21}^3(t)\right) \\ &+ P_0\left(H_{22}^3(t)\right)P_0\left(A_iA_jA_k(t)\middle|H_{22}^3(t)\right) \\ &+ P_0\left(H_{31}^3(t)\right)P_0\left(A_iA_jA_k(t)\middle|H_{31}^3(t)\right) \\ &+ P_0\left(H_{32}^3(t)\right)P_0\left(A_iA_jA_k(t)\middle|H_{32}^3(t)\right) \\ &+ P_0\left(H_{33}^3(t)\right)P_0\left(A_iA_jA_k(t)\middle|H_{33}^3(t)\right). \end{aligned} \tag{1.89}$$

Parmi les six termes de la somme à droite de l'équation, un seul est non nul. En effet, les trois gènes pigés ne peuvent être de trois allèles différents que s'ils ont trois ancêtres distincts, ce qui est le cas seulement pour le dernier terme. On a donc

$$E_0\left[x_i(t)x_j(t)x_k(t)\right] = P_0\left(H_{33}^3(t)\right)P_0\left(A_iA_jA_k(t)\middle|H_{33}^3(t)\right). \tag{1.90}$$

La valeur du premier facteur à droite de l'équation ci-dessus peut être trouvée à la section 1.6. Pour le deuxième facteur, on procède comme précédemment par égalité des chances à la

génération 0 sous neutralité, ce qui nous donne

$$P_0\left(A_iA_iA_i(t)\middle|H_{33}^3(t)\right) = x_i(0)\left(\frac{2Nx_j(0)}{2N-1}\right)\left(\frac{2Nx_k(0)}{2N-2}\right). \tag{1.91}$$

Remplaçons ceci dans l'équation précédente pour obtenir

$$E_0\left[x_i(t)x_j(t)x_k(t)\right] = \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)p_{33}^{(t)}x_i(0)\left(\frac{2Nx_j(0)}{2N-1}\right)\left(\frac{2Nx_k(0)}{2N-2}\right). \tag{1.92}$$

Ceci se simplifie pour donner

$$E_0\left[x_i(t)x_j(t)x_k(t)\right] = p_{33}^t x_i(0)x_j(0)x_k(0). \tag{1.93}$$

1.8 Étapes pour l'étude d'un modèle particulier

Nous voulions dans ce chapitre donner une méthode générale pour calculer la probabilité de fixation d'un allèle A_i dans un modèle de population à plusieurs allèles. Nous pouvons maintenant le faire ainsi :

1. Trouver l'expression pour l'espérance du changement de la fréquence de l'allèle A_i dans le modèle considéré avec intensité de sélection s.
2. Dériver cette espérance par rapport à s et évaluer à $s = 0$ pour obtenir la variable aléatoire $\phi_i(t)$.
3. À l'aide du résultat 1.2, déterminer l'espérance de $\phi_i(t)$ sous neutralité.
4. Calculer la somme

$$P_s(F_i) = x_i(0) + \sum_{t\geqslant 0} E_0\left[\phi_i(t)\right] + o(s).$$

5. Simplifier l'expression obtenue en calculant les probabilités de non coalescence qui s'y rattachent.

Ce travail peut être plus ou moins fastidieux dépendant de la complexité du modèle. Dans les chapitres suivants, nous utiliserons ce cheminement pour trouver cette probabilité de fixation dans différents modèles, puis nous analyserons les résultats.

1.9 Analyse de la probabilité de fixation

Bien que le résultat de cette marche à suivre, une expression explicite pour la probabilité de fixation de l'allèle A_i, soit un résultat complet et une finalité en lui-même, sa forme sera souvent

trop complexe pour que l'esprit en saisisse toutes les significations par simple observation. Par contre, on peut l'utiliser pour définir les conditions sous lesquelles un individu sera avantagé par la sélection naturelle dans un modèle précis. Cette analyse sera faite pour chacun des modèles (sauf celui de Cannings) selon les critères suivants.

On dira que la fixation d'un allèle A_i est avantagée par la sélection si

$$P_s(F_i) > P_0(F_i), \tag{1.94}$$

c'est-à-dire si sa probabilité de fixation dans le modèle avec intensité de sélection s est strictement supérieure à celle dans le modèle neutre. L'expression de $P_s(F_i)$ est donnée par (1.27), ce qui permet de dire que cette condition est équivalente à

$$\sum_{t\geqslant 0} E_0\left[\phi_i(t)\right] > 0, \tag{1.95}$$

pour s suffisamment petite. Cette condition sera nécessairement plus simple que la probabilité brute et facilitera son interprétation.

Chapitre 2

Modèle de Wright-Fisher haploïde

2.1 Description du modèle

Nous étudierons dans ce chapitre une population haploïde évoluant selon le modèle de Wright-Fisher avec sélection [2].

Notre population contient $2N$ individus portant chacun un des n allèles A_1, A_2, ..., A_n à un locus donné. On note les fréquences des allèles à la génération t par les variables $x_1(t)$, $x_2(t)$, ..., $x_n(t)$.

À chaque génération, tous les gènes de la population sont remplacés par le mécanisme habituel du modèle de Wright-Fisher. C'est-à-dire que les $2N$ individus de la population produisent une infinité de zygotes porteurs des mêmes allèles que leurs géniteurs. Ensuite, la sélection joue son rôle sur ces zygotes, modifiant les fréquences alléliques. Les coefficients de sélection v_j indiquent la viabilité d'un individu d'allèle A_j. Pour $j \in \{1, \ldots, n\}$, la fréquence de A_j passe à

$$\frac{x_j(t) f_j}{\overline{f}(t),} \tag{2.1}$$

où $f_j := 1 + sv_j$ est la valeur sélective de A_j et $\overline{f}(t)$ est la valeur sélective moyenne à la génération t. Parmi cette infinité de zygotes sont pigés $2N$ zygotes qui formeront l'effectif de la génération suivante.

La valeur sélective moyenne à la génération t est donnée par le produit scalaire du vecteur des fréquences alléliques et de celui des valeurs sélectives, c'est-à-dire que

$$\overline{f}(t) = \sum_{j=1}^{n} x_j(t) f_j. \tag{2.2}$$

Ceci implique que

$$\overline{f}(t) = \sum_{j=1}^{n} x_j(t)\,(1 + sv_j)\,, \tag{2.3}$$

et donc que

$$\overline{f}(t) = 1 + s\overline{v}(t), \tag{2.4}$$

où le coefficient de sélection moyen $\overline{v}(t)$ est le produit scalaire du vecteur des coefficients de sélection et de celui des fréquences alléliques à la génération t.

2.2 Espérance du changement de la fréquence allélique

Pour déterminer la probabilité de fixation d'un allèle A_i, il nous faut d'abord extraire la fonction $\phi_i(t)$ telle que décrite à la section 1.4. Il nous faut donc calculer $E_s\left[\Delta x_i(t)\middle|\overrightarrow{x}(t)\right]$.

En partant de

$$E_s\left[\Delta x_i(t)\middle|\overrightarrow{x}(t)\right] = E_s\left[x_i(t+1)\middle|\overrightarrow{x}(t)\right] - x_i(t), \tag{2.5}$$

on obtient que

$$E_s\left[\Delta x_i(t)\middle|\overrightarrow{x}(t)\right] = x_i(t)\frac{f_i}{\overline{f}(t)} - x_i(t). \tag{2.6}$$

En utilisant les expressions des valeurs sélective, on a

$$E_s\left[\Delta x_i(t)\middle|\overrightarrow{x}(t)\right] = x_i(t)\left(\frac{1+sv_i}{1+s\overline{v}(t)} - 1\right), \tag{2.7}$$

ce qui peut s'écrire sous la forme

$$E_s\left[\Delta x_i(t)\middle|\overrightarrow{x}(t)\right] = \frac{sx_i(t)\,(v_i - \overline{v}(t))}{1+s\overline{v}(t)}. \tag{2.8}$$

On voit que l'espérance du changement de la fréquence de A_i est positive si le coefficient de sélection de A_i est plus grand que le coefficient de sélection moyen de la population.

Il faut maintenant dériver par rapport à s, ce qui donne

$$E_s'\left[\Delta x_i(t)\middle|\overrightarrow{x}(t)\right] = \frac{x_i(t)\,(v_i - \overline{v}(t))}{(1+s\overline{v}(t))^2}, \tag{2.9}$$

puis évaluer à $s = 0$ pour obtenir $\phi_i(t)$. On a donc

$$\phi_i(t) = E_0'\left[\Delta x_i(t)\middle|\overrightarrow{x}(t)\right] = x_i(t)(v_i - \overline{v}(t)). \tag{2.10}$$

2.3 Calcul de la probabilité de fixation

Calculons d'abord

$$E_0\left[\phi_i(t)\right] = E_0\left[x_i(t)(v_i - \overline{v}(t))\right]. \tag{2.11}$$

Nous devons d'abord exprimer $\overline{v}(t)$ en fonction des fréquences alléliques, ce qui donne

$$E_0\left[\phi_i(t)\right] = E_0\left[x_i(t)\left(v_i - \sum_{j=1}^{n} v_j x_j(t)\right)\right], \tag{2.12}$$

puis réorganiser les termes pour obtenir

$$E_0\left[\phi_i(t)\right] = E_0\left[v_i x_i(t) - v_i x_i^2(t) - \sum_{\substack{j=1\\ j\neq i}}^{n} v_j x_i(t) x_j(t)\right]. \tag{2.13}$$

Nous pouvons maintenant utiliser les résultats (1.54), (1.55) et (1.56), ce qui mène à

$$E_0\left[\phi_i(t)\right] = v_i x_i(0) - v_i x_i(0)\left(1 - p_{22}^t\left(1 - x_i(0)\right)\right) - \sum_{\substack{j=1\\ j\neq i}}^{n} v_j p_{22}^t x_i(0) x_j(0), \tag{2.14}$$

ou encore,

$$E_0\left[\phi_i(t)\right] = v_i x_i(0) p_{22}^t - p_{22}^t x_i(0) \sum_{j=1}^{n} v_j x_j(0), \tag{2.15}$$

c'est-à-dire,

$$E_0\left[\phi_i(t)\right] = p_{22}^t x_i(0)\left(v_i - \overline{v}(0)\right). \tag{2.16}$$

On peut, à partir de là, obtenir la probabilité de fixation éventuelle de l'allèle A_i. On sait par (1.27) que

$$P_s(F_i) = x_i(0) + s\sum_{t\geq 0} E_0\left[\phi_i(t)\right] + o(s). \tag{2.17}$$

En utilisant l'expression trouvée en (2.16), on obtient que

$$P_s(F_i) = x_i(0) + s\sum_{t\geqslant 0} p_{22}^t x_i(0)\,(v_i - \overline{v}(0)) + o(s). \tag{2.18}$$

En évaluant la série géométrique de raison $0 < p_{22} < 1$, on trouve finalement

$$P_s(F_i) = x_i(0) + sx_i(0)\,(v_i - \overline{v}(0))\,\frac{1}{1-p_{22}} + o(s), \tag{2.19}$$

c'est-à-dire,

$$P_s(F_i) = x_i(0)\left(1 + \frac{s}{p_{21}}\,(v_i - \overline{v}(0))\right) + o(s). \tag{2.20}$$

Tout ce qu'il nous reste à faire est de calculer p_{21} qui est, rappelons-nous, la probabilité sous neutralité que deux gènes distincts choisis au hasard dans la population aient le même géniteur. Dans le modèle neutre de Wright-Fisher, cette probabilité est celle que le géniteur du deuxième gène pigé soit le même que celui du premier, et donc

$$p_{21} = \frac{1}{2N}. \tag{2.21}$$

En remplaçant la probabilité p_{21} par sa valeur dans (2.20), on obtient la probabilité de fixation

$$P_s(F_i) = x_i(0)\,(1 + 2Ns\,(v_i - \overline{v}(0))) + o(s), \tag{2.22}$$

pour l'allèle A_i. On peut voir que cette fixation est avantagée par la sélection (voir (1.94)) si

$$v_i > \overline{v}(0), \tag{2.23}$$

pour s suffisamment petite. Résumons ces résultats.

Résultat 2.1

Dans ce modèle de Wright-Fisher haploïde, la probabilité de fixation de l'allèle A_i *est*

$$P_s(F_i) = x_i(0)\,(1 + 2Ns\,(v_i - \overline{v}(0))) + o(s), \tag{2.24}$$

ce qui est explicite en connaissant l'état initial de la population. La fixation de cet allèle est avantagée par la sélection si

$$v_i > \overline{v}(0), \tag{2.25}$$

donc si son coefficient de sélection est plus grand que le coefficient moyen initial de la population.

Chapitre 3

Modèle de Moran haploïde

3.1 Description du modèle

Nous pouvons appliquer la méthode décrite au chapitre 1 pour étudier la probabilité de fixation dans le modèle de Moran. Le modèle étudié ici est différent du modèle classique décrit dans [2] du fait qu'il y a présence de multiples allèles différents, mais les fondements du modèle demeurent intacts. Le modèle de Moran est un modèle avec chevauchement de générations, c'est-à-dire que les individus ne sont pas tous renouvelés en même temps. Le temps demeure discret, mais est compté en nombre d'événements de naissance-mort (voir plus bas) qui correspondent aux générations des autres modèles.

On considère une population haploïde de taille $2N$ comportant n allèles à un locus donné que nous dénotons par A_1, A_2, ..., A_n. La fréquence de l'allèle A_j à l'instant discret $t \geqslant 0$ est représentée par la variable aléatoire $x_j(t)$. Les allèles se manifestent dans les phénotypes de leurs porteurs et influent sur leurs chances de survie. Ainsi, à chaque allèle A_j est associée un coefficient de sélection m_j et une valeur sélective associée $g_j := 1 + sm_j$, où s est est l'intensité de la sélection. Ces coefficients sont des paramètres de mortalité, comparativement aux paramètres de viabilité du modèle précédent. Une faible valeur sélective est ici bénéfique à un individu en raison du mécanisme qui suit.

À chaque instant discret d'un modèle de Moran est associé un unique événement de naissance-mort. Un géniteur est pigé totalement au hasard dans la population de $2N$ gènes et crée une copie de lui-même. Le rejeton créé prendra la place d'un des individus de la population, possiblement du géniteur. On dit alors que cet individu meurt, d'où l'appellation d'événement de naissance-mort. La probabilité qu'un gène meurt est proportionnelle à sa valeur sélective.

On définit

$$\overline{g}(t) := \sum_{j=1}^{n} g_j x_j(t), \tag{3.1}$$

la valeur sélective moyenne à l'instant discret $t \geqslant 0$, qu'on peut aussi écrire en fonction de s par l'expression

$$\overline{g}(t) = 1 + s\overline{m}(t), \tag{3.2}$$

où

$$\overline{m}(t) := \sum_{j=1}^{n} m_j x_j(t) \tag{3.3}$$

est le coefficient de sélection moyen à cet instant discret.

3.2 Espérance du changement de la fréquence allélique

Trouvons tout d'abord l'expression de $E_s\left[\Delta x_i(t)\middle|\vec{x}(t)\right]$. La variable aléatoire $\Delta x_i(t)$ peut prendre trois valeurs : $-1/2N$, 0 ou $1/2N$. Elle prend comme valeur $1/2N$ quand l'individu qui se reproduit n'est pas d'allèle A_i, mais que celui qui meurt l'est. Dans le cas inverse, elle vaut $-1/2N$. Il est aisé d'exprimer l'espérance en conditionnant sur le choix du géniteur, ce qui donne

$$E_s\left[\Delta x_i(t)\middle|\vec{x}(t)\right] = \frac{x_i(t)}{2N}\left(1 - x_i(t)\frac{g_i}{\overline{g}(t)}\right) - (1 - x_i(t))\left(\frac{x_i(t)g_i}{2N\overline{g}(t)}\right). \tag{3.4}$$

Avec quelques manipulations, on obtient

$$E_s\left[\Delta x_i(t)\middle|\vec{x}(t)\right] = \frac{x_i(t)}{2N}\left(1 - \frac{g_i}{\overline{g}(t)}\right). \tag{3.5}$$

Écrivons les valeurs sélectives en fonction de s avant de dériver. On a que

$$E_s\left[\Delta x_i(t)\middle|\vec{x}(t)\right] = \frac{x_i(t)}{2N}\left(1 - \frac{1 + sm_i}{1 + s\overline{m}(t)}\right), \tag{3.6}$$

qu'on dérive pour obtenir

$$E_s'\left[\Delta x_i(t)\middle|\vec{x}(t)\right] = \frac{-x_i(t)\left[m_i(1 + s\overline{m}(t)) - (1 + sm_i)\overline{m}(t)\right]}{2N(1 + s\overline{m}(t))^2}. \tag{3.7}$$

En évaluant à $s = 0$, on trouve

$$\phi_i(t) = E_0'\left[\Delta x_i(t) \middle| \overrightarrow{x}(t)\right] = \frac{x_i(t)(\overline{m}(t) - m_i)}{2N}. \tag{3.8}$$

3.3 Calcul de la probabilité de fixation

Avant de calculer l'espérance de $\phi_i(t)$, on développe $\overline{m}(t)$. On obtient que

$$\phi_i(t) = \frac{x_i(t)}{2N}\left(\sum_{j=1}^{n} m_j x_j(t) - m_i\right), \tag{3.9}$$

et donc que

$$\phi_i(t) = \frac{1}{2N}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} m_j x_j(t) x_i(t) + \left(x_i^2(t) - x_i(t)\right) m_i\right). \tag{3.10}$$

En utilisant les résultats (1.54), (1.55) et (1.56), on trouve comme espérance

$$E_0\left[\phi_i(t)\right] = \frac{1}{2N}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} m_j p_{22}^t x_j(0) x_i(0) + \left(x_i(t)\left(1 - p_{22}^t(1 - x_i(0))\right) - x_i(0)\right) m_i\right), \tag{3.11}$$

ou, après simplification,

$$E_0\left[\phi_i(t)\right] = \frac{1}{2N}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} m_j p_{22}^t x_j(0) x_i(0) + \left(x_i(0) p_{22}^t (x_i(0) - 1)\right) m_i\right), \tag{3.12}$$

c'est-à-dire,

$$E_0\left[\phi_i(t)\right] = \frac{1}{2N}\left(\sum_{j=1}^{n} m_j p_{22}^t x_j(0) x_i(0) - x_i(0) p_{22}^t m_i\right). \tag{3.13}$$

Ceci peut s'écrire sous la forme

$$E_0\left[\phi_i(t)\right] = \frac{x_i(0) p_{22}^t}{2N}\left(\overline{m}(0) - m_i\right). \tag{3.14}$$

En intégrant cette expression à (1.27), on obtient que la probabilité de fixation de A_i est

$$P_s(F_i) = x_i(0) + s \sum_{t\geqslant 0} \frac{x_i(0) p_{22}^t}{2N}\left(\overline{m}(0) - m_i\right) + o(s). \tag{3.15}$$

En calculant la série géométrique, on peut écrire que

$$P_s(F_i) = x_i(0) + \frac{x_i(0)}{2N(1-p_{22})}\left(\overline{m}(0) - m_i\right) + o(s). \tag{3.16}$$

Or, on sait que $1 - p_{22}$ est la valeur de p_{21}, c'est-à-dire la probabilité sous neutralité que deux gènes distincts à un instant donné proviennent d'un même gène à l'instant précédent. Il existe aucune ou une seule paire de gènes liés de cette façon, dépendant de la survie ou de la mort du gène qui a été copié. Le gène copié n'a pas été remplacé par son rejeton avec probabilité $(2N-1)/2N$. La paire de gènes ayant le même géniteur sera ensuite sélectionnée au hasard avec probabilité

$$\frac{1}{\binom{2N}{2}}. \tag{3.17}$$

On a donc que

$$p_{21} = \left(\frac{2N-1}{2N}\right)\frac{1}{\binom{2N}{2}}, \tag{3.18}$$

ce qui vaut

$$p_{21} = \frac{1}{2N^2}. \tag{3.19}$$

Insérons ceci dans (3.16), ce qui nous donne

$$P_s(F_i) = x_i(0) + \frac{x_i(0)}{N}\left(\overline{m}(0) - m_i\right) + o(s). \tag{3.20}$$

Symétriquement au modèle de Wright-Fisher haploïde, la fixation de l'allèle A_i sera avantagée par la sélection si

$$m_i < \overline{m}(0), \tag{3.21}$$

pour s suffisament petite. Ceci nous permet d'énoncer les résultats pour ce modèle.

Résultat 3.1

Dans un modèle de Moran haploïde à n allèles, la probabilité de fixation de l'allèle A_i est

$$P_s(F_i) = x_i(0)\left(1 + \frac{\overline{m}(0) - m_i}{N}\right) + o(s), \tag{3.22}$$

où $\overline{m}(0)$ est le coefficient de sélection moyen de la population initiale. Pour s petite, la fixation de cet allèle est avantagée par la sélection dans le cas où

$$m_i < \overline{m}(0). \tag{3.23}$$

Chapitre 4

Modèle de Wright-Fisher haploïde pour la détermination du sexe

4.1 Description du modèle

Dans ce chapitre, nous retournons au modèle de Wright-Fisher haploïde pour y ajouter une variante intéressante décrite brièvement à la page 24 de [6], ou plus en détails dans [4] et [7].

Notre population comporte $2N$ individus haploïdes présentant les allèles A_1, A_2, ..., A_n à un unique locus. Comme toujours, on représente les fréquences des allèles à une certaine génération t par les variables aléatoires $x_1(t)$, $x_2(t)$, ..., $x_n(t)$.

La particularité du modèle est que les individus sont hermaphrodites ou, du moins, peuvent jouer les rôles de mère ou de père avec certaines probabilités dépendant de leur génotype. De façon plus générale, on pourrait considérer un modèle avec autant de femelles que de mâles, mais avec des valeurs sélectives différentes chez les deux sexes.

On débute une génération avec N couples mère/père qui produisent une infinité de rejetons par la copie du matériel génétique d'un des deux parents, et ce de façon équiprobable. La probabilité qu'un rejeton agisse comme mâle ou femelle dépend de l'allèle qu'il porte. Un individu d'allèle A_j ainsi créé jouera le rôle d'un mâle avec probabilité r_j et le rôle d'une femelle avec probabilité complémentaire $1-r_j$. Ces individus forment des couples femelle/mâle parmi lesquels N seront pigés au hasard pour débuter la prochaine génération.

La sélection se présente dans la forme que prend r_j qui représente la probabilité de la fonction mâle chez l'allèle A_i. On suppose une probabilité de la forme $r_j := r + sd_j$. Le paramètre r est en quelque sorte la probabilité neutre et la probabilité associée à chaque allèle s'en éloigne proportionnellement à une intensité de sélection s et à un coefficient spécifique d_j. La valeur de r doit être comprise entre 0 et 1 exclusivement, sans quoi, sous neutralité, un des deux sexes

serait absent de la population, ce qui résulterait en son extinction totale.

4.2 Espérance du changement de la fréquence allélique

Pour trouver la probabilité de fixation d'un allèle A_i dans ce modèle, nous appliquons la procédure expliquée à la section 1.8. Il faut d'abord trouver l'espérance de la variation de la fréquence allélique de l'allèle. On part du fait que

$$E_s\left[\Delta x(t)\middle|\vec{x}(t)\right] = E_s\left[x(t+1)\middle|\vec{x}(t)\right] - x_i(t), \tag{4.1}$$

où la dernière espérance est la probabilité qu'un individu pigé au hasard à la génération $(t+1)$ soit d'allèle A_i. Comme le matériel génétique provient de façon équiprobable du père ou de la mère, cette probabilité est la somme de la moitié de la probabilité que son père soit d'allèle A_i et de la moitié de la probabilité que sa mère soit d'allèle A_i. Ceci nous donne

$$E_s\left[\Delta x(t)\middle|\vec{x}(t)\right] = \frac{1}{2}\left(x_i(t)\frac{r_i}{\bar{r}(t)} + x_i(t)\frac{1-r_i}{1-\bar{r}(t)}\right) - x_i(t), \tag{4.2}$$

où

$$\bar{r}(t) = \sum_{j=1}^{n} r_j x_j(t). \tag{4.3}$$

En définissant

$$\bar{d}(t) = \sum_{j=1}^{n} d_j x_j(t), \tag{4.4}$$

on peut récrire

$$E_s\left[\Delta x(t)\middle|\vec{x}(t)\right] = \frac{x_i(t)}{2}\left(\frac{r+sd_i}{r+s\bar{d}(t)} + \frac{1-r-sd_i}{1-r-s\bar{d}(t)}\right) - x_i(t). \tag{4.5}$$

Dérivons par rapport à s pour trouver

$$E_s'\left[\Delta x(t)\middle|\vec{x}(t)\right] = \frac{x_i(t)}{2}\left(\frac{d_i\left(r+s\bar{d}(t)\right) - \bar{d}(t)\left(r+sd_i\right)}{\left(r+s\bar{d}(t)\right)^2} + \frac{-d_i\left(1-r-s\bar{d}(t)\right) + \bar{d}(t)\left(1-r-sd_i\right)}{\left(1-r-s\bar{d}(t)\right)^2}\right), \tag{4.6}$$

qui, évalué à $s = 0$, donne

$$E_0'\left[\Delta x(t)\middle|\vec{x}(t)\right] = \frac{x_i(t)}{2}\left(\frac{d_i r - \bar{d}(t)r}{r^2} + \frac{-d_i(1-r) + \bar{d}(t)(1-r)}{(1-r)^2}\right). \tag{4.7}$$

Après quelques simplifications et manipulations, ceci est équivalent à

$$E_0'\left[\Delta x(t)\middle|\vec{x}(t)\right] = \frac{x_i(t)}{2}\left(\frac{\left(d_i - \bar{d}(t)\right)(1-r) - \left(d_i - \bar{d}(t)\right)r}{r(1-r)}\right), \tag{4.8}$$

ce qui entraîne que

$$\phi_i(t) = \frac{x_i(t)\left(d_i - \bar{d}(t)\right)(1-2r)}{2r(1-r)}. \tag{4.9}$$

Si $r = 1/2$, alors $\phi_i(t) = 0$, ce qui veut dire que l'effet de premier ordre de la sélection est nul. Ceci implique que la probabilité de fixation de l'allèle A_i est sa fréquence initiale additionnée d'une fonction petit ordre de s. Il faudrait alors étudier l'effet de deuxième ordre, mais ceci requerrait des justifications supplémentaires pour le calcul de la dérivée seconde, semblables à celles réalisées pour passer de (1.18) à (1.19). Ceci n'est pas l'objet de ce texte. Dans ce qui suit, nous traiterons le cas où $r \neq 1/2$.

4.3 Calcul de la probabilité de fixation

Il nous faut en deuxième lieu calculer l'espérance, sous neutralité, de la variable aléatoire $\phi_i(t)$. En reprenant le dernier résultat, on a

$$E_0\left[\phi_i(t)\right] = E_0\left[\frac{x_i(t)\left(d_i - \bar{d}(t)\right)(1-2r)}{2r(1-r)}\right]. \tag{4.10}$$

Les r ne sont que des constantes. C'est dans la moyenne $\bar{d}(t)$ que se trouvent plusieurs variables aléatoires. Explicitons-les pour obtenir

$$E_0\left[\phi_i(t)\right] = \frac{1-2r}{2r(1-r)}E_0\left[d_i x_i(t) - \sum_{\substack{j=1 \\ j\neq n}}^{n} d_j x_j(t) x_i(t) - d_i x_i^2(t)\right]. \tag{4.11}$$

À l'aide de (1.54), (1.56) et (1.55), il est facile de calculer

$$E_0\left[\phi_i(t)\right] = \frac{1-2r}{2r(1-r)}\left(d_i x_i(0) - \sum_{\substack{j=1 \\ j\neq n}}^{n} d_j p_{22}^t x_j(0) x_i(0) - d_i x_i(0)\left(1 + p_{22}^t\left(x_i(0) - 1\right)\right)\right). \tag{4.12}$$

En annulant une paire de termes opposés par le signe, en mettant en évidence p_{22}^t et en complétant la somme, ceci prend la forme

$$E_0\left[\phi_i(t)\right] = \frac{(1-2r)p_{22}^t}{2r(1-r)}\left(-\sum_{j=1}^{n} d_j x_j(0)x_i(0) + d_i x_i^2(0) - d_i x_i(0)\left(x_i(0)-1\right)\right). \tag{4.13}$$

La somme est une moyenne. De plus, il y a encore simplification évidente, simplification qui nous laisse avec l'expression

$$E_0\left[\phi_i(t)\right] = \frac{(1-2r)p_{22}^t}{2r(1-r)}\left(-x_i(0)\overline{d}(0) + d_i x_i(0)\right). \tag{4.14}$$

En remplaçant cette espérance dans (1.27), on a une expression pour la probabilité de fixation de l'allèle A_i, soit

$$P_s(F_i) = x_i(0) + s\sum_{t\geqslant 0}\frac{(1-2r)p_{22}^t x_i(0)\left(d_i - \overline{d}(0)\right)}{2r(1-r)} + o(s). \tag{4.15}$$

Le seul facteur dépendant de t est la probabilité de non coalescence. En calculant le résultat de la série géométrique, on trouve

$$P_s(F_i) = x_i(0) + \frac{s x_i(0)(1-2r)\left(d_i - \overline{d}(0)\right)}{2r(1-r)\left(1-p_{22}\right)} + o(s). \tag{4.16}$$

On sait bien que $(1-p_{22})$ est égale à p_{21}. Il nous faut calculer cette probabilité sous neutralité, c'est-à-dire dans le cas où la probabilité de la fonction mâle est la même pour tous les individus. Ceci implique que les fréquences des allèles sont les mêmes chez les mâles que chez les femelles. Plus spécifiquement, si on prend un des couples formés par les rejetons de $2N$ individus, chaque individu a chance égale d'être le géniteur du mâle de ce couple, et pareillement pour la femelle.

Pour déterminer p_{21}, on cherche la probabilité qu'un premier individu de la génération t ait le même géniteur qu'un deuxième individu de la même génération. Or, comme on vient de l'expliquer, les $2N$ individus de la génération précédente ont des chances équiprobables d'être le géniteur du premier individu. Ainsi,

$$p_{21} = \frac{1}{2N}. \tag{4.17}$$

En combinant les résultats (4.16) et (4.17), on trouve finalement

$$P_s(F_i) = x_i(0)\left[1 + 2Ns\frac{(1-2r)}{2r(1-r)}\left(d_i - \overline{d}(0)\right)\right] + o(s). \tag{4.18}$$

Toujours en se référant à (1.94), on observe que, sous une intensité de sélection assez faible, la

fixation de l'allèle A_i sera avantagée si

$$(1-2r)\left(d_i - \overline{d}(0)\right) > 0, \tag{4.19}$$

ce qui est possible dans deux situations : soit

$$\left(r < \frac{1}{2}\right) \text{ et } \left(d_i > \overline{d}(0)\right), \tag{4.20}$$

ou bien

$$\left(r > \frac{1}{2}\right) \text{ et } \left(d_i < \overline{d}(0)\right). \tag{4.21}$$

Ceci revient à dire que la sélection avantage les allèles dont la fonction mâle associée se rapproche de 1/2 par rapport à la fonction mâle moyenne dans la population inititale. De façon réciproque, ceci signifie que la sélection s'oppose à l'inéquité mâles/femelles.

Résultat 4.1

Dans le modèle de Wright-Fisher haploïde pour la détermination du sexe, la probabilité de fixation de l'allèle A_i sachant l'état initial de la population $\overrightarrow{x}(0)$ est

$$P_s(F_i) = x_i(0)\left[1 + 2Ns\frac{(1-2r)}{2r(1-r)}\left(d_i - \overline{d}(0)\right)\right] + o(s), \tag{4.22}$$

où la constante r est la probabilité de la fonction mâle sous neutralité. Pour une intensité de sélection suffisament faible, la fixation de l'allèle A_i est avantagée si

$$\left(r < \frac{1}{2}\right) \textit{ et } \left(d_i > \overline{d}(0)\right), \tag{4.23}$$

ou si

$$\left(r > \frac{1}{2}\right) \textit{ et } \left(d_i < \overline{d}(0)\right), \tag{4.24}$$

c'est-à-dire si la probabilité de la fonction mâle qui y est associée est comprise entre 1/2 et la probabilité moyenne de la fonction mâle dans la population initiale.

Chapitre 5

Modèle de Cannings

5.1 Description du modèle

On considère une population haploïde de taille $2N$ qui évolue selon un modèle de Cannings avec sélection. Ce modèle est l'extrapolation du cas à deux allèles décrit dans [2] et [5].

Il y a n allèles possibles à l'unique locus, soit A_1, A_2, ..., A_n. L'effectif de l'allèle A_j à la génération t est noté $y_j(t)$ et sa fréquence $x_j(t)$, i.e.

$$x_j(t) = \frac{y_j(t)}{2N}. \tag{5.1}$$

On ordonne arbitrairement les individus de la génération t qui sont d'allèle A_1, puis ceux qui sont d'allèle A_2, et ainsi de suite jusqu'à ce que tous les individus soient ordonnés. Cette opération nous permet de définir la variable aléatoire $Z_k(t)$, qui représente le nombre de rejetons du k-ieme individu de la génération t ainsi ordonnée. Comme les individus de même allèle A_i ont un comportement évolutif identique, les variables aléatoires

$$Z_{y_0(t)+y_1(t)+\ldots+y_{j-1}(t)+1}(t), \ldots, Z_{y_0(t)+y_1(t)+\ldots+y_j(t)}(t), \tag{5.2}$$

où $y_0(t) = 0$, sont échangeables pour $j \in \{1, \ldots, n\}$. C'est ce qui caractérise le modèle de Cannings. Pour la même raison, on considère que toute distribution marginale de $(Z_1(t), \ldots, Z_{2N}(t))$ ne dépend que du type d'allèle associé à chaque $Z_k(t)$. Ainsi, pour toute suite strictement croissante $1 \leqslant a_1 < a_2 < \ldots < a_m \leqslant 2N$, le vecteur aléatoire $(Z_{a_1}(t), Z_{a_2}(t), \ldots, Z_{a_m}(t))$ a la même distribution si on permute des composantes associées à un même allèle.

La sélection affectera la distribution des variables aléatoires $Z_i(t)$. Notre but est de déterminer la probabilité de fixation de chacun des allèles en fonction de la sélection.

5.1.1 Constance de la taille de la population

Nous venons de mentionner que la taille de la population est constante. Pour que ceci se réalise, il faut brider les variables aléatoires $Z_k(t)$ par certaines conditions. En fait, la taille de la population à la génération $(t+1)$ est la variable aléatoire $\sum_{k=1}^{2N} Z_k(t)$. Pour que $\sum_{k=1}^{2N} Z_k(t) = 2N$ à tout coup, il faut qu'elle soit de moyenne $2N$ et de variance nulle.

Premièrement, sous l'hypothèse que les variables

$$Z_{y_0(t)+\ldots+y_{j-1}(t)+1}(t), \ldots, Z_{y_0(t)+\ldots+y_j(t)}(t) \tag{5.3}$$

sont échangeables, elles sont identiquement distribuées et ont la même espérance. Définissons

$$\mu_j^s(t) := E_s\left[Z_{y_0(t)+y_1(t)+\ldots+y_j(t)}(t)\middle|\overrightarrow{x}(t)\right], \tag{5.4}$$

représentant le nombre moyen de rejetons engendrés par un individu d'allèle A_j à la génération t pour $j \in \{1,\ldots,n\}$. Ainsi, la condition

$$E_s\left[\sum_{k=1}^{2N} Z_k(t)\right] = 2N \tag{5.5}$$

est équivalente à

$$\sum_{j=1}^{n} y_j(t)\mu_j^s(t) = 2N. \tag{5.6}$$

En divisant par $2N$, on peut récrire sous la forme

$$\overline{\mu^s}(t) = 1. \tag{5.7}$$

Ceci revient à dire que les individus engendrent en moyenne un rejeton chacun.

Deuxièmement, toujours sous l'hypothèse que les variables aléatoires $Z_k(t)$ associées à un même allèle sont échangeables, plusieurs d'entre elles ont les mêmes variances et covariances. On définit trois groupes de covariances. D'abord, pour $i \neq j$,

$$Cov_{ij}^s(t) := Cov_s(Z_{y_1(t)+\ldots+y_i(t)}(t), Z_{y_1(t)+\ldots+y_j(t)}(t)) \tag{5.8}$$

est la covariance entre le nombre de rejetons de deux individus d'allèles A_i et A_j à la génération t. Si $y_i(t) \geqslant 2$, on définit

$$Cov_{ii}^s(t) := Cov_s(Z_{y_1(t)+\ldots+y_i(t)}(t), Z_{y_1(t)+\ldots+y_i(t)-1}(t)), \tag{5.9}$$

qui est la covariance, sous sélection s, entre le nombre de rejetons de deux individus différents de la génération t, mais tous deux d'allèle A_i. Cette covariance n'est définie que s'il existe deux tels individus. Finalement,

$$Var_i^s(t) := Var_s(Z_{y_1(t)+\ldots+y_i(t)}(t)) = Cov_s(Z_{y_1(t)+\ldots+y_i(t)}(t), Z_{y_1(t)+\ldots+y_i(t)}(t)) \tag{5.10}$$

est la variance, sous sélection s, du nombre de rejetons d'un individu d'allèle A_i à la génération t.

Comme on a mentionné plus haut, on a

$$Var_s\left(\sum_{j=1}^{2N} Z_j(t)\right) = 0, \tag{5.11}$$

ce qui est équivalent à

$$\sum_{k=1}^{2N}\sum_{l=1}^{2N} Cov_s(Z_k(t), Z_l(t)) = 0. \tag{5.12}$$

En utilisant les définitions ci-dessus, on transforme cette dernière égalité en

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{n} y_i(t)y_j(t)Cov_{ij}^s(t) - y_i(t)Cov_{ii}^s(t) + y_i(t)Var_i^s(t)\right) = 0, \tag{5.13}$$

que l'on récrit

$$\sum_{i=1}^{n}\sum_{j=1}^{n} y_i(t)y_j(t)Cov_{ij}^s(t) = \sum_{j=1}^{n} y_j(t)\left(Cov_{jj}^s(t) - Var_j^s(t)\right). \tag{5.14}$$

Cette condition et la condition (5.7) sont nécessaires pour que la taille de la population demeure constante. On supposera désormais qu'elles sont satisfaites.

5.1.2 L'effet de la sélection

En absence de sélection, chaque individu aurait en moyenne un rejeton, peu importe l'allèle qu'il porte. Nous introduisons maintenant une fonction $h_j(t)$ modifiant légèrement le nombre moyen de rejetons d'un individu d'allèle A_j de la génération t lorsqu'on considère une sélection faible d'intensité $s > 0$. On ajoute maintenant au modèle l'hypothèse que la moyenne est une fonction suffisamment régulière en s pour qu'elle puisse s'écrire sous la forme

$$\mu_j^s(t) = 1 + s\left(1 - x_j(t)\right)\left(h_j(t) + O(s)\right), \tag{5.15}$$

où la fonction $O(s)$ tend vers 0 quand s tend vers 0. Supposons aussi que cette fonction $O(s)$ soit dérivable et que sa dérivée soit continue en $s = 0$. Le facteur $(1 - x_j(t))$ se veut garant de la continuité de $\mu_j^s(t)$ quand on approche de $x_j(t) = 1$. En effet, lorsque la population est entièrement composée d'individus d'allèle A_j, chacun de ces individus devra avoir en moyenne un rejeton.

La fonction $h_j(t)$ peut prendre diverses formes. Nous considérerons ici la forme linéaire par rapport aux fréquences alléliques, c'est-à-dire

$$h_j(t) := a_0(j) + a_1(j)x_1(t) + a_2(j)x_2(t) + \ldots + a_n(j)x_n(t), \tag{5.16}$$

que nous noterons parfois comme le produit scalaire $a_0(i) + \vec{a}(j) \cdot \vec{x}(t)$. Cette forme nous laisse beaucoup de latitude, mais ces fonctions sont tout de même soumises à une restriction imposée par (5.7) qui, sous l'hypothèse (5.15), ne sera satisfaite que si

$$\sum_{j=1}^{n} x_j(t)\,(1 - x_j(t))\,h_j(t) = 0. \tag{5.17}$$

5.2 Espérance du changement de la fréquence allélique

Nous sommes maintenant prêts à entreprendre l'étude de la probabilité de fixation de l'allèle A_i. On rappelle que

$$E_s\left[\Delta x_i(t) \middle| \vec{x}(t)\right] = E_s\left[x_i(t+1) \middle| \vec{x}(t)\right] - x_i(t). \tag{5.18}$$

Dans le modèle présent, $x_i(t+1)$ s'exprime en fonction des $Z_j(t)$ par

$$E_s\left[\Delta x_i(t) \middle| \vec{x}(t)\right] = E_s\left[\frac{1}{2N} \sum_{k=y_0(t)+\ldots+y_{i-1}(t)+1}^{y_0(t)+\ldots+y_i(t)} Z_k(t) \middle| \vec{x}(t)\right] - x_i(t), \tag{5.19}$$

ce qui nous donne

$$E_s\left[\Delta x_i(t) \middle| \vec{x}(t)\right] = \frac{1}{2N} y_i(t)\mu_i^s(t) - x_i(t), \tag{5.20}$$

c'est-à-dire

$$E_s\left[\Delta x_i(t) \middle| \vec{x}(t)\right] = x_i(t)(\mu_i^s(t) - 1). \tag{5.21}$$

En remplaçant $\mu_i^s(t)$ par (5.15), on trouve

$$E_s\left[\Delta x_i(t)\middle|\vec{x}(t)\right] = s x_i(t)\left(1 - x_i(t)\right)\left(h_i(t) + O(s)\right). \tag{5.22}$$

De (5.22) et (1.28), on peut extraire $\phi_i(t)$, dont l'expression est

$$\phi_i(t) = x_i(t)\left(1 - x_i(t)\right) h_i(t), \tag{5.23}$$

car on a supposé sour 5.15 que $O(s)$ était dérivable et que sa dérivée était continue en $s = 0$.

Calculons maintenant

$$E_0\left[\phi_i(t)\right] = E_0\left[x_i(t)\left(1 - x_i(t)\right)\left(a_0(i) + \vec{a}(i)\cdot\vec{x}(t)\right)\right]. \tag{5.24}$$

En développant, on obtient

$$E_0\left[\phi_i(t)\right] = E_0\left[a_0(i)x_i(t) + \sum_{j=1}^{n} a_j(i)x_j(t)x_i(t) - a_0(i)x_i^2(t) - \sum_{j=1}^{n} a_j(i)x_j(t)x_i^2(t)\right]. \tag{5.25}$$

Si on isole les termes en $j = i$ de chaque somme, on trouve

$$\begin{aligned} E_0\left[\phi_i(t)\right] = E_0\Bigg[& a_0(i)x_i(t) + \left(a_i(i) - a_0(i)\right)x_i^2(t) - a_i(i)x_i^3(t) \\ & + \sum_{\substack{j=1 \\ j\neq i}}^{n}\left[a_j(i)x_j(t)x_i(t) - a_j(i)x_j(t)x_i^2(t)\right]\Bigg]. \end{aligned} \tag{5.26}$$

En substituant (1.54), (1.55), (1.57), (1.56) et (1.58) dans la dernière équation, on obtient que

$$\begin{aligned} E_0\left[\phi_i(t)\right] = {} & a_0(i)x_i(0) + \left(a_i(i) - a_0(i)\right)x_i(0)\left(1 - p_{22}^t\left(1 - x_i(0)\right)\right) \\ & - a_i(i)x_i\left(1 - \frac{1 - x_i(0)}{2}\left(3p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right)\right) \\ & + \sum_{\substack{j=1 \\ j\neq i}}^{n} a_j(i)x_i(0)x_j(0)\left(p_{22}^t - \frac{1}{2}\left(p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right)\right). \end{aligned} \tag{5.27}$$

Quelques termes s'annulent. On peut mettre en évidence $x_i(0)$, puis on ajoute et soustrait le

terme manquant de la somme pour compléter celle-ci. Ainsi, on trouve

$$\begin{aligned} E_0[\phi_i(t)] = x_i(0)\Bigg[& -p_{22}^t a_i(i)(1-x_i(0)) + a_0(i)p_{22}^t(1-x_i(0)) \\ & + \frac{a_i(i)(1-x_i(0))}{2}\left(3p_{22}^t + p_{33}^t(2x_i(0)-1)\right) \\ & - a_i(i)x_i(0)\left(p_{22}^t - \frac{1}{2}(p_{22}^t + p_{33}^t(2x_i(0)-1))\right) \\ & + \sum_{j=1}^{n} a_j(i)x_j(0)\left(p_{22}^t - \frac{1}{2}\left(p_{22}^t + p_{33}^t(2x_i(0)-1)\right)\right)\Bigg]. \end{aligned} \tag{5.28}$$

On développe maintenant les trois premiers termes, puis on effectue quelques modifications mineures aux derniers pour obtenir

$$\begin{aligned} E_0[\phi_i(t)] = x_i(0)\Bigg[& -p_{22}^t a_i(i) + x_i(0)p_{22}^t a_i(i) + a_0(i)p_{22}^t - x_i(0)a_0(i)p_{22}^t \\ & + \frac{a_i(i)}{2}\left(3p_{22}^t + p_{33}^t(2x_i(0)-1)\right) - \frac{a_i(i)x_i(0)}{2}\left(3p_{22}^t + p_{33}^t(2x_i(0)-1)\right) \\ & - \frac{a_i(i)x_i(0)}{2}\left(p_{22}^t - p_{33}^t(2x_i(0)-1)\right) + \sum_{j=1}^{n}\frac{a_j(i)x_j(0)}{2}\left(p_{22}^t - p_{33}^t(2x_i(0)-1)\right)\Bigg]. \end{aligned} \tag{5.29}$$

Ainsi, on observe quelques termes à regrouper. Ce faisant, on a que

$$\begin{aligned} E_0[\phi_i(t)] = x_i(0)\Bigg[& \frac{1}{2}p_{22}^t a_i(i) + x_i(0)p_{22}^t a_i(i) + a_0(i)p_{22}^t - x_i(0)a_0(i)p_{22}^t \\ & + \frac{a_i(i)}{2}p_{33}^t(2x_i(0)-1) - 2a_i(i)x_i(0)p_{22}^t \\ & + \frac{1}{2}\left(p_{22}^t - p_{33}^t(2x_i(0)-1)\right)\sum_{j=1}^{n} a_j(i)x_j(0)\Bigg], \end{aligned} \tag{5.30}$$

En mettant en évidence un facteur $1/2$, puis p_{22}^t et p_{33}^t, on trouve

$$\begin{aligned} E_0[\phi_i(t)] = \frac{x_i(0)}{2}\Big[& p_{22}^t\left(a_i(i) - 2x_i(0)a_i(i) + 2a_0(i) - 2x_i(0)a_0(i) + \vec{a}(i)\vec{x}(0)\right) \\ & + p_{33}^t\left(a_i(i)(2x_i(0)-1) - (2x_i(0)-1)\vec{a}(i)\vec{x}(0)\right)\Big]. \end{aligned} \tag{5.31}$$

Pour p_{33}^t, on a un facteur $(2x_i(0)-1)$. Tentons de faire apparaître le même facteur pour p_{22}^t en écrivant

$$\begin{aligned} E_0[\phi_i(t)] = \frac{x_i(0)}{2}\Big[& p_{22}^t\left(a_i(i) - 2x_i(0)a_i(i) + a_0(i) - 2x_i(0)a_0(i) + h_i(0)\right) \\ & + p_{33}^t(2x_i(0)-1)\left(a_i(i) - \vec{a}(i)\vec{x}(0)\right)\Big]. \end{aligned} \tag{5.32}$$

La réussite est partielle, car il reste le terme $h_i(0)$. En le faisant apparaître pour p_{33}^t aussi, on trouve que

$$E_0[\phi_i(t)] = \frac{x_i(0)}{2}\Big[p_{22}^t\left((1-2x_i(0))(a_0(i)+a_i(i))+h_i(0)\right) + p_{33}^t(2x_i(0)-1)\left(a_0(i)+a_i(i)-h_i(0)\right)\Big]. \tag{5.33}$$

5.3 Calcul de la probabilité de fixation

En rappelant le résultat (1.27) et en le combinant à (5.33), on obtient que

$$P_s(F_i) = x_i(0) + s\sum_{t\geq 0}\frac{x_i(0)}{2}\Big[p_{22}^t\left((1-2x_i(0))(a_0(i)+a_i(i))+h_i(0)\right) + p_{33}^t(2x_i(0)-1)\left(a_0(i)+a_i(i)-h_i(0)\right)\Big] + o(s). \tag{5.34}$$

Les seuls termes dans la somme dépendant de t sont les probabilités de non coalescence. On a donc deux séries géométriques que nous pouvons calculer, obtenant ainsi

$$P_s(F_i) = x_i(0) + \frac{sx_i(0)}{2}\left[\frac{(1-2x_i(0))(a_0(i)+a_i(i))+h_i(0)}{1-p_{22}} - \frac{(1-2x_i(0))\left(a_0(i)+a_i(i)-h_i(0)\right)}{1-p_{33}}\right] + o(s). \tag{5.35}$$

On obtient donc une expression pour la probabilité de fixation éventuelle de l'allèle A_i en fonction des probabilités $(1-p_{22})$ et $(1-p_{33})$ qu'il nous faut calculer. Pour cela, nous calculerons d'abord p_{21} et p_{31}.

La probabilité p_{21} est celle sous neutralité que deux individus sélectionnés au hasard à une génération $(t+1)$ proviennent du même géniteur de la génération t. N'oublions pas que, sous neutralité, tous les individus de toutes les générations agissent selon les mêmes règles, la valeur de p_{21} ne dépend donc pas de t. Nous utiliserons ici $t = 0$. De plus, tous les individus étant interchangeables, la probabilité recherchée est $2N$ fois la probabilité que l'individu numéro 1 de la génération 0 soit le géniteur des deux gènes pigés à la génération 1. En conditionnant sur le nombre de rejetons de cet individu, on a

$$p_{21} = 2N\sum_{k=2}^{2N} P_0(Z_1(0)=k)\frac{\binom{k}{2}}{\binom{2N}{2}}. \tag{5.36}$$

En explicitant les nombres de combinaisons, on s'aperçoit qu'on peut rajouter les deux premiers

termes de la somme, car ils sont nuls. Donc,

$$p_{21} = \sum_{k=0}^{2N} P_0(Z_1(0) = k)\frac{k(k-1)}{2N-1}, \tag{5.37}$$

ce qui nous permet de reformuler le tout en fonction d'espérances, soit

$$p_{21} = \frac{E_0[Z_1^2(0) - Z_1(0)]}{2N-1}. \tag{5.38}$$

On sait que $E_0[Z_1(t)] = 1$ et donc on peut soustraire $E_0[Z_1(0)]$ puis ajouter $E_0[1]$ sans changer la valeur de l'expression. De cette façon, on a

$$p_{21} = \frac{E_0[Z_1^2(0) - 2Z_1(0) + 1]}{2N-1}. \tag{5.39}$$

On voit ainsi que

$$p_{21} = \frac{E_0\left[(Z_1(0)-1)^2\right]}{2N-1}. \tag{5.40}$$

On procède par le même raisonnement pour p_{31}. D'abord, en conditionnant, on obtient que

$$p_{31} = 2N\sum_{k=3}^{2N} P_0(Z_1(0) = k)\frac{\binom{k}{3}}{\binom{2N}{3}}. \tag{5.41}$$

Cette fois, les trois premiers termes de la somme sont nuls, ce qui nous permet d'affirmer que

$$p_{31} = \sum_{k=0}^{2N} P_0(Z_1(0) = k)\frac{k(k-1)(k-2)}{(2N-1)(2N-2)}, \tag{5.42}$$

qui peut s'écrire sous la forme

$$p_{31} = \frac{E_0[Z_1^3(0) - 3Z_1^2(0) + 2Z_1(0)]}{(2N-1)(2N-2)}. \tag{5.43}$$

Nous pouvons ajouter $E_0[Z_1(0)]$ puis soustraire $E_0[1]$ pour obtenir

$$p_{31} = \frac{E_0[Z_1^3(0) - 3Z_1^2(0) + 3Z_1(0) - 1]}{(2N-1)(2N-2)}, \tag{5.44}$$

ce que l'on reconnaît comme étant

$$p_{31} = \frac{E_0\left[(Z_1(0)-1)^3\right]}{(2N-1)(2N-2)}. \tag{5.45}$$

On cherchait initialement $(1-p_{22})$ et $(1-p_{33})$. La première valeur est facile à obtenir. Par

(1.50) et (5.40), on a

$$1 - p_{22} = \frac{E_0\left[(Z_1(0) - 1)^2\right]}{2N - 1}. \tag{5.46}$$

L'équation (1.49) nous aide pour la suite. Comme

$$1 - p_{33} = 3p_{21} - 2p_{31}, \tag{5.47}$$

il suffit de remplacer les valeurs trouvées en (5.40) et (5.45) pour trouver

$$1 - p_{33} = \frac{3(N-2)E_0\left[(Z_1(0) - 1)^2\right] - 2E_0\left[(Z_1(0) - 1)^3\right]}{(2N-1)(2N-2)}. \tag{5.48}$$

Ceci nous permet de conclure.

Résultat 5.1

Dans un modèle de Cannings avec sélection respectant les conditions (5.14) *et* (5.17), *la probabilité de fixation de l'allèle* A_i *est donnée par*

$$P_s(F_i) = x_i(0) + \frac{sx_i(0)}{2}\left[\frac{(1-2x_i(0))(a_0(i)+a_i(i)) + h_i(\vec{x}(0))}{1-p_{22}} - \frac{(1-2x_i(0))\left(a_0(i)+a_i(i) - h_i(\vec{x}(0))\right)}{1-p_{33}}\right] + o(s), \tag{5.49}$$

où

$$1 - p_{22} = \frac{E_0\left[(Z_1(0) - 1)^2\right]}{2N - 1} \tag{5.50}$$

et

$$1 - p_{33} = \frac{3(N-2)E_0\left[(Z_1(0) - 1)^2\right] - 2E_0\left[(Z_1(0) - 1)^3\right]}{(2N-1)(2N-2)}. \tag{5.51}$$

Chapitre 6

Modèle de Wright-Fisher diploïde

6.1 Description du modèle

On considère une population diploïde de N individus qui évolue selon un modèle de Wright-Fisher à coefficients de sélection constants. À la différence des chapitres 2 et 4, on considère cette fois le modèle diploïde. Pour référence, voir [2].

Les n allèles présents dans la population, $A_1, A_2, \ldots, A_n$, peuvent former les n^2 génotypes ordonnés A_iA_j pour $i, j \in \{1, \ldots, n\}$. Nous devrons étudier les fréquences de ces génotypes pour connaître celles des différents allèles. Nous utiliserons comme d'habitude la variable aléatoire $x_j(t)$ pour représenter la fréquence de l'allèle A_j à la génération t, tandis que la variable aléatoire $y_{ij}(t)$ sera la fréquence du génotype ordonné A_iA_j à cette même génération.

D'un point de vue génétique, on devrait regrouper les individus A_iA_j et A_jA_i, mais les notations sont plus simples en les considérant séparément, ce qui ne change de toute façon aucunement le modèle.

On démarre une génération avec N individus reproducteurs qui créent un bassin infini de gamètes dont les fréquences alléliques sont identiques à celles chez les reproducteurs. Les gamètes mâles et femelles sont ensuite couplés au hasard pour former une infinité de jeunes individus. Ceux-ci survivent avec plus ou moins de succès, tout dépendant de leur génotype, pour devenir des individus matures. Ce succès variable crée une différence entre les fréquences alléliques et génotypiques chez les jeunes individus et chez les individus matures. Ces nouvelles fréquences seront notées $x_j^m(t)$ et $y_{ij}^m(t)$, où le m indique qu'il s'agit des fréquences parmi les individus matures. Il y a finalement échantillonnage de N individus parmi les individus matures, individus qui engendreront la prochaine génération.

Les fréquences alléliques sont les mêmes chez les géniteurs que chez les jeunes qu'ils ont produits. Par contre, les fréquences génotypiques ne sont pas les mêmes chez ces deux groupes

d'individus, le deuxième étant issu d'un brassage du matériel génétique du premier. Il est donc important de préciser que $y_{ij}(t)$ représente la fréquence du génotype A_iA_j chez les jeunes individus, et non pas chez leurs géniteurs.

On définit les coefficients de sélection u_{ij} pour mesurer la capacité de survie d'un individu de génotype A_iA_j. Elles nous donnent les valeurs sélectives correspondantes $f_{ij} := 1 + su_{ij}$, où $s > 0$ est l'intensité de la sélection. Les fréquences génotypiques mentionnées au paragraphe précédent sont ainsi liées par la relation

$$y_{ij}^m(t) = y_{ij}(t)\frac{f_{ij}}{\overline{f}(t)}, \tag{6.1}$$

où

$$\overline{f}(t) = \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}(t)f_{ij} \tag{6.2}$$

est la valeur sélective moyenne à la génération t. Les fréquences alléliques chez les jeunes individus et les individus matures sont liées par cette même relation. Il faut expliciter ce lien avant de débuter le calcul des probabilités de fixation des allèles.

Comme les gamètes sont infiniment nombreux et couplés au hasard, les génotypes A_iA_j et A_jA_i ont les mêmes fréquences chez les jeunes d'une même génération. En fait, on a

$$y_{ij}(t) = x_i(t)x_j(t) = y_{ji}(t). \tag{6.3}$$

De plus, l'ordre des allèles dans le génotype d'un individu n'influence pas son phénotype, c'est-à-dire l'expression de ses gènes. Ainsi, des individus A_iA_j et A_jA_i sont équivalents en tous points. Ils ont donc le même coefficient de sélection, c'est-à-dire que

$$u_{ij} = u_{ji}, \tag{6.4}$$

ce qui entraîne que

$$y_{ij}^m(t) = y_{ji}^m(t). \tag{6.5}$$

On s'intéresse aux fréquences des allèles, mais nous devons travailler sur les fréquences des génotypes pour appliquer la sélection. Nous pourrons par la suite retrouver les fréquences alléliques grâce à l'équation

$$x_i(t) = \frac{1}{2}\sum_{\substack{j=1\\ j\neq i}}^{n} y_{ij}(t) + \frac{1}{2}\sum_{\substack{j=1\\ j\neq i}}^{n} y_{ji}(t) + y_{ii}(t), \tag{6.6}$$

qui montre par (6.3) que

$$x_i(t) = \sum_{j=1}^{n} y_{ij}(t). \tag{6.7}$$

Par le même raisonnement et l'équation (6.5), on peut déduire que cette relation est aussi valide chez les individus matures.

Utilisons ceci pour trouver les fréquences génotypiques puis alléliques matures d'une génération à partir des fréquences chez les jeunes individus. On a déjà mentionné que

$$y_{ij}^{m}(t) = y_{ij}(t)\frac{f_{ij}}{\overline{f}(t)}, \tag{6.8}$$

ce qui est équivalent à

$$y_{ij}^{m}(t) = \frac{y_{ij}(t)(1+su_{ij})}{\overline{f}(t)}. \tag{6.9}$$

En insérant au dénominateur le coefficient de sélection génotypique moyen à la génération t, c'est-à-dire

$$\overline{u}(t) = \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}(t)u_{ij}, \tag{6.10}$$

on obtient que

$$y_{ij}^{m}(t) = \frac{y_{ij}(t)(1+su_{ij})}{1+s\overline{u}(t)}. \tag{6.11}$$

En insérant ceci dans (6.7), on trouve que

$$x_i^{m}(t) = \sum_{j=1}^{n} \frac{y_{ij}(t)(1+su_{ij})}{1+s\overline{u}(t)}. \tag{6.12}$$

On applique ensuite (6.3) pour avoir

$$x_i^{m}(t) = \frac{x_i(t)\left(1+s\sum_{j=1}^{n} x_j(t)u_{ij}\right)}{1+s\overline{u}(t)}, \tag{6.13}$$

qu'on écrira comme

$$x_i^{m}(t) = \frac{x_i(t)\left(1+s\overline{u_i}(t)\right)}{1+s\overline{u}(t)}, \tag{6.14}$$

où

$$\overline{u_i}(t) = \sum_{j=1}^{n} x_j(t) u_{ij} = \frac{1}{x_i(t)} \sum_{j=1}^{n} y_{ij}(t) u_{ij} \tag{6.15}$$

est le coefficient de sélection moyen de l'allèle A_i à la génération t. La dernière équation montre clairement que la moyenne des coefficients de sélection alléliques moyens est le coefficient de sélection génotypique moyen, i.e.

$$\sum_{i=1}^{n} x_i(t) \overline{u_i}(t) = \overline{u}(t). \tag{6.16}$$

6.2 Espérance du changement de la fréquence allélique

On a toujours

$$E_s\left[\Delta x_i(t) \middle| \overrightarrow{x}(t)\right] = E_s\left[x_i(t+1) \middle| \overrightarrow{x}(t)\right] - x_i(t). \tag{6.17}$$

On rappelle que $x_i(t+1)$ est la fréquence de l'allèle A_i chez les géniteurs de la génération $(t+1)$. Ces géniteurs sont choisis au hasard parmi les individus matures de la génération t et donc

$$E_s\left[\Delta x_i(t) \middle| \overrightarrow{x}(t)\right] = x_i^m(t) - x_i(t). \tag{6.18}$$

Insérons le résultat (6.14) pour obtenir

$$E_s\left[\Delta x_i(t) \middle| \overrightarrow{x}(t)\right] = \frac{x_i(t)\left(1 + s\overline{u_i}(t)\right)}{1 + s\overline{u}(t)} - x_i(t), \tag{6.19}$$

puis simplifions l'équation résultante, ce qui donne

$$E_s\left[\Delta x_i(t) \middle| \overrightarrow{x}(t)\right] = \frac{s x_i(t)\left(\overline{u_i}(t) - \overline{u}(t)\right)}{1 + s\overline{u}(t)}. \tag{6.20}$$

Il est facile par la suite de dériver par rapport à s puis d'évaluer en zéro pour obtenir

$$\phi_i(t) = x_i(t)\left(\overline{u_i}(t) - \overline{u}(t)\right). \tag{6.21}$$

6.3 Calcul de la probabilité de fixation

En reprenant le dernier résultat, on sait que

$$E_0\left[\phi_i(t)\right] = E_0\left[x_i(t)\left(\overline{u_i}(t) - \overline{u}(t)\right)\right]. \tag{6.22}$$

Écrivons le tout en fréquences alléliques à partir de (6.15) et (6.10) pour avoir

$$E_0\left[\phi_i(t)\right] = E_0\left[x_i(t)\left(\sum_{j=1}^{n} x_j(t)u_{ij} - \sum_{j=1}^{n}\sum_{k=1}^{n} x_j(t)x_k(t)u_{jk}\right)\right]. \tag{6.23}$$

Pour calculer l'espérance, il faut tenter de mettre en évidence les espérances calculées à la section 1.7. Pour ce faire, extrayons les termes en $j = i$ et $k = i$ des sommes, ce qui donne

$$\begin{aligned} E_0\left[\phi_i(t)\right] = E_0\Bigg[& x_i^2(t)u_{ii} + \sum_{\substack{j=1\\ j\neq i}}^{n} x_i(t)x_j(t)u_{ij} - x_i^3(t)u_{ii} - 2\sum_{\substack{j=1\\ j\neq i}}^{n} x_i^2(t)x_j(t)u_{ij} \\ & - \sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i}}^{n} x_i(t)x_j(t)x_k(t)u_{jk}\Bigg]. \end{aligned} \tag{6.24}$$

Complétons le travail sur les termes en $k = j$. On trouve alors

$$\begin{aligned} E_0\left[\phi_i(t)\right] = E_0\Bigg[& x_i^2(t)u_{ii} - x_i^3(t)u_{ii} + \sum_{\substack{j=1\\ j\neq i}}^{n}\Big(x_i(t)x_j(t)u_{ij} - 2x_i^2(t)x_j(t)u_{ij} \\ & - x_i(t)x_j^2(t)u_{jj}\Big) - \sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i,j}}^{n} x_i(t)x_j(t)x_k(t)u_{jk}\Bigg]. \end{aligned} \tag{6.25}$$

Nous remplaçons maintenant les espérances par leurs valeurs trouvées en (1.55), (1.57), (1.56), (1.58) et (1.59). Notons que les espérances des deux derniers termes de la somme médiane sont symétriques. On obtient que

$$\begin{aligned} E_0\left[\phi_i(t)\right] = {} & u_{ii}x_i(0)\left(1 + p_{22}^t\left(x_i(0) - 1\right)\right) \\ & - u_{ii}x_i(0)\left(1 + \frac{x_i(0) - 1}{2}\left(3p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right)\right) \\ & + \sum_{\substack{j=1\\ j\neq i}}^{n}\Bigg(u_{ij}p_{22}^t x_i(0)x_j(0) - 2u_{ij}\frac{x_i(0)x_j(0)}{2}\left(p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right) \\ & - u_{jj}\frac{x_i(0)x_j(0)}{2}\left(p_{22}^t + p_{33}^t\left(2x_j(0) - 1\right)\right)\Bigg) \\ & - \sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i,j}}^{n} u_{jk}p_{33}^t x_i(0)x_j(0)x_k(0). \end{aligned} \tag{6.26}$$

On peut maintenant annuler quelques termes. Profitons-en pour ajouter à la double somme les termes où $k = i$ ou $k = j$, ce qui nous laisse avec

$$\begin{aligned} E_0[\phi_i(t)] &= u_{ii}x_i(0)(x_i(0)-1)\left(-\frac{p_{22}^t}{2}-\frac{p_{33}^t}{2}(2x_i(0)-1)\right) \\ &+\sum_{\substack{j=1\\j\neq i}}^{n} x_i(0)x_j(0)\left(-u_{ij}p_{33}^t(2x_i(0)-1)-\frac{1}{2}u_{jj}\left(p_{22}^t+p_{33}^t(2x_j(0)-1)\right)\right) \\ &+\sum_{\substack{j=1\\j\neq i}}^{n}\left(u_{ij}p_{33}^t x_i^2(0)x_j(0)+u_{jj}p_{33}^t x_i(0)x_j^2(0)\right) \\ &-\sum_{\substack{j=1\\j\neq i}}^{n} p_{33}^t x_i(0)x_j(0)\sum_{k=1}^{n}u_{jk}x_k(0). \end{aligned} \tag{6.27}$$

On reconnaît $\overline{u_j}(0)$ à la toute fin de l'expression. On peut ensuite fusionner les trois sommes, obtenant dans l'exercice

$$\begin{aligned} E_0[\phi_i(t)] &= u_{ii}x_i(0)(x_i(0)-1)\left(-\frac{p_{22}^t}{2}-\frac{p_{33}^t}{2}(2x_i(0)-1)\right) \\ &+\sum_{\substack{j=1\\j\neq i}}^{n} x_i(0)x_j(0)\left(-u_{ij}p_{33}^t(x_i(0)-1)-\frac{1}{2}u_{jj}\left(p_{22}^t-p_{33}^t\right)-\overline{u_j}(0)p_{33}^t\right). \end{aligned} \tag{6.28}$$

En développant le premier terme et en ajoutant le terme en $j = i$ à la somme, on trouve que

$$\begin{aligned} E_0[\phi_i(t)] &= u_{ii}x_i(0)\left(\frac{p_{22}^t}{2}+\frac{p_{33}^t}{2}(2x_i(0)-1)\right)-u_{ii}x_i^2(0)\left(\frac{p_{22}^t}{2}+\frac{p_{33}^t}{2}(2x_i(0)-1)\right) \\ &+x_i^2(0)\left(u_{ii}p_{33}^t(x_i(0)-1)+\frac{1}{2}u_{ii}\left(p_{22}^t-p_{33}^t\right)+\overline{u_i}(0)p_{33}^t\right) \\ &-\sum_{j=1}^{n}x_i(0)x_j(0)\left(u_{ij}p_{33}^t(x_i(0)-1)+\frac{1}{2}u_{jj}\left(p_{22}^t-p_{33}^t\right)+\overline{u_j}(0)p_{33}^t\right). \end{aligned} \tag{6.29}$$

On a alors de nouvelles simplifications à effectuer dans les deux premières lignes. On peut du même coup extraire plusieurs coefficients de sélection moyens de la somme grâce à (6.15) et (6.16), ce qui nous donne

$$\begin{aligned} E_0[\phi_i(t)] &= u_{ii}x_i(0)\left(\frac{p_{22}^t}{2}+\frac{p_{33}^t}{2}(2x_i(0)-1)\right)-u_{ii}x_i^2(0)p_{33}^t+x_i^2(0)\overline{u_i}(0)p_{33}^t \\ &-x_i(0)\left(\overline{u_i}(0)p_{33}^t(x_i(0)-1)+\frac{1}{2}\left(p_{22}^t-p_{33}^t\right)\sum_{j=1}^{n}x_j(0)u_{jj}+\overline{u}(0)p_{33}^t\right). \end{aligned} \tag{6.30}$$

Il ne reste que deux termes en p_{22}^t. Regroupons-les pour obtenir

$$E_0[\phi_i(t)] = x_i(0)\left[p_{22}^t\left(\frac{u_{ii}}{2} - \frac{\sum_{j=1}^n x_j(0)u_{jj}}{2}\right) + p_{33}^t\left(u_{ii}\frac{1}{2}(2x_i(0)-1) - u_{ii}x_i(0) + x_i(0)\overline{u_i}(0) - \overline{u_i}(0)(x_i(0)-1) + \frac{1}{2}\sum_{j=1}^n x_j(0)u_{jj} - \overline{u}(0)\right)\right]. \tag{6.31}$$

Quelques simplifications sur le facteur de p_{33}^t nous mènent à

$$E_0[\phi_i(t)] = x_i(0)\left[p_{22}^t\left(\frac{u_{ii}}{2} - \frac{\sum_{j=1}^n x_j(0)u_{jj}}{2}\right) + p_{33}^t\left(-\frac{u_{ii}}{2} + \overline{u_i}(0) + \frac{1}{2}\sum_{j=1}^n x_j(0)u_{jj} - \overline{u}(0)\right)\right]. \tag{6.32}$$

On peut récrire le tout de façon à mettre en évidence la ressemblance entre les deux termes. En effet, on a

$$E_0[\phi_i(t)] = \frac{x_i(0)}{2}\left[p_{22}^t\left(u_{ii} - \sum_{j=1}^n x_j(0)u_{jj}\right) - p_{33}^t\left(u_{ii} - \sum_{j=1}^n x_j(0)u_{jj} - 2(\overline{u_i}(0) - \overline{u}(0))\right)\right]. \tag{6.33}$$

On peut maintenant utiliser cette espérance pour obtenir la probabilité de fixation éventuelle de l'allèle A_i. L'équation (1.27) stipule que

$$P_s(F_i) = x_i(0) + s\sum_{t\geqslant 0} E_0[\phi_i(t)] + o(s). \tag{6.34}$$

Substituons-y (6.33), ce qui nous donne

$$P_s(F_i) = x_i(0) + s\sum_{t\geqslant 0}\frac{x_i(0)}{2}\left[p_{22}^t\left(u_{ii} - \sum_{j=1}^n x_j(0)u_{jj}\right) - p_{33}^t\left(u_{ii} - \sum_{j=1}^n x_j(0)u_{jj} - 2(\overline{u_i}(0) - \overline{u}(0))\right)\right] + o(s). \tag{6.35}$$

Il n'y a dans la somme que les probabilités de non coalescence qui dépendent de la variable t. On peut calculer les séries géométriques résultantes et obtenir

$$P_s(F_i) = x_i(0) + \frac{sx_i(0)}{2}\left[\frac{u_{ii} - \sum_{j=1}^n x_j(0)u_{jj}}{1-p_{22}} - \frac{u_{ii} - \sum_{j=1}^n x_j(0)u_{jj} - 2(\overline{u_i}(0) - \overline{u}(0))}{1-p_{33}}\right] + o(s). \tag{6.36}$$

Une dernière manipulation algébrique nous donne l'expression

$$P_s(F_i) = x_i(0) + \frac{sx_i(0)}{2}\left[\left(\frac{1}{1-p_{22}} - \frac{1}{1-p_{33}}\right)\left(u_{ii} - \sum_{j=1}^{n} x_j(0)u_{jj}\right) + \frac{2\left(\overline{u_i}(0) - \overline{u}(0)\right)}{1-p_{33}}\right] + o(s). \tag{6.37}$$

Il ne reste qu'à calculer, sous neutralité, les probabilités p_{22} et p_{33}. Dans ce modèle de Wright-Fisher, elles sont plutôt simples à déterminer.

En effet, p_{21} est la probabilité que deux gènes distincts soient la copie du gène présent au même locus d'un même géniteur, qui est égale à $1/2N$. Pareillement, on sait aussi que p_{31} vaut $1/(2N)^2$. Par (1.49), on a que

$$\begin{aligned} 1 - p_{33} &= 3\frac{1}{2N} - 2\frac{1}{(2N)^2} \\ &= \frac{6N-2}{4N^2} \\ &= \frac{3N-1}{2N^2}. \end{aligned} \tag{6.38}$$

Donc, on a

$$\frac{2}{1-p_{33}} = \frac{4N^2}{3N-1}. \tag{6.39}$$

Par (1.50), on a aussi

$$\frac{1}{1-p_{22}} = \frac{1}{p_{21}} = 2N, \tag{6.40}$$

d'où

$$\begin{aligned} \frac{1}{1-p_{22}} - \frac{1}{1-p_{33}} &= 2N - \frac{2N^2}{3N-1} \\ &= 2N\left(\frac{3N-1-N}{3N-1}\right) \\ &= \frac{(2N)(2N-1)}{3N-1}. \end{aligned} \tag{6.41}$$

En remarquant que les coefficients trouvés à la dernière section ont un facteur commun et en le mettant en évidence dans (6.37), on obtient le résultat final, i.e.

$$P_s(F_i) = x_i(0) + \frac{Nsx_i(0)}{3N-1}\left[(2N-1)\left(u_{ii} - \sum_{j=1}^{n} x_j(0)u_{jj}\right) + 2N\left(\overline{u_i}(0) - \overline{u}(0)\right)\right] + o(s). \tag{6.42}$$

Ceci peut sembler rébarbatif à première vue, mais on peut en tirer une analyse simple sur l'effet de premier ordre de la sélection sur un allèle particulier. D'après la définition donnée en

(1.94), le dernier résultat implique que la fixation de l'allèle A_i est avantagée par la sélection si

$$(2N-1)\left(u_{ii} - \sum_{j=1}^{n} x_j(0)u_{jj}\right) + 2N\left(\overline{u_i}(0) - \overline{u}(0)\right) > 0, \tag{6.43}$$

pour $s > 0$ suffisamment petit. L'équation (6.16) nous rappelle que $\overline{u}(t) = \sum_{j=1}^{n} x_j(t)\overline{u_j}(t)$. On a donc deux termes sur les quatre comportant des coefficients de sélection qui concernent directement l'allèle A_i tandis que les deux autres traitent de la population entière. En faisant les regroupements correspondants, la dernière inégalité est équivalente à

$$(2N-1)u_{ii} + 2N\overline{u_i}(0) > \sum_{j=1}^{n}(2N-1)x_j(0)u_{jj} + 2N\sum_{j=1}^{n} x_j(0)\overline{u_j}(0). \tag{6.44}$$

Divisons maintenant par $2N$ de chaque côté et mettons en évidence un facteur $x_j(0)$ pour obtenir

$$\left(1 - \frac{1}{2N}\right)u_{ii} + \overline{u_i}(0) > \sum_{j=1}^{n} x_j(0)\left[\left(1 - \frac{1}{2N}\right)u_{jj} + \overline{u_j}(0)\right]. \tag{6.45}$$

Introduisons la notation

$$v_i(N, \vec{x}(0)) := \left(1 - \frac{1}{2N}\right)u_{ii} + \overline{u_i}(0), \tag{6.46}$$

qui nous permet de récrire (6.45) comme

$$v_i(N, \vec{x}(0)) > \overline{v}(N, \vec{x}(0)). \tag{6.47}$$

Ainsi, $v_i(N, \vec{x}(0))$ agit comme coefficient de sélection effectif de l'allèle A_i.

Pour une grande population, on pourrait négliger le terme $1/2N$ dans (6.46), ce qui rendrait l'expression du coefficient de sélection effectif de A_i plus élégante.

Résultat 6.1

Dans ce modèle de Wright-Fisher diploïde à coefficients de sélection constants, la probabilité de fixation éventuelle de l'allèle A_i avec une intensité de sélection $s > 0$ étant donné les fréquences alléliques initiales dans la population est

$$P_s(F_i) = x_i(0) + \frac{Nsx_i(0)}{3N-1}\left[(2N-1)\left(u_{ii} - \sum_{j=1}^{n} x_j(0)u_{jj}\right) + 2N\left(\overline{u_i}(0) - \overline{u}(0)\right)\right] + o(s). \tag{6.48}$$

La fixation de l'allèle A_i est avantagée par une faible intensité de sélection si son coefficient de sélection effectif

$$v_i(N, \overrightarrow{x}(0)) = \left(1 - \frac{1}{2N}\right)u_{ii} + \overline{u_i}(0) \tag{6.49}$$

est supérieur à sa moyenne dans la population.

Chapitre 7

Modèle de Wright-Fisher diploïde avec interactions

7.1 Description du modèle

Nous allons modifier légèrement le modèle de Wright-Fisher diploïde décrit au chapitre précédent. Cette fois-ci, les coefficients de sélection varieront selon la composition de la population en raison des interactions entre les individus. Ce modèle est inspiré de celui de Lessard [6, p. 21], mais transposé à une population diploïde à plusieurs allèles.

La population comptera N individus diploïdes où n allèles différents, A_1, A_2, ..., A_n, peuvent apparaître au locus considéré. La taille de la population est constante dans le temps. La fréquence de l'allèle A_i à la génération t sera notée par la variable aléatoire $x_i(t)$.

Comme la population est diploïde, il y aura n^2 génotypes ordonnés possibles. Nous utiliserons la variable aléatoire $y_{ij}(t)$ pour représenter la fréquence du génotype A_iA_j à la génération t. On parlera souvent des génotypes A_iA_j et A_jA_i comme s'ils étaient différents. C'est dans le but de simplifier les notations, surtout celles des sommes.

Les individus de la population peuvent être de deux phénotypes différents : P_1 et P_2. Un individu A_iA_j sera de phénotype P_1 avec probabilité $g_{ij} := (h_i + h_j)/2$ et de phénotype P_2 avec la probabilité complémentaire, ces probabilités étant les moyennes de celles d'individus A_iA_i et A_jA_j.

La sélection prend place lors des interactions entre les individus. Si un individu P_a rencontre un individu P_b, l'effet de leur interaction sur le premier individu est représenté numériquement par la valeur m_{ab}. C'est un bonus si cette valeur est positive, mais un malus si elle est négative. On regroupe ces valeurs dans la matrice d'interaction M de dimension deux par deux. Ainsi, la

valeur sélective du génotype A_iA_j est

$$f_{ij}(t) := 1 + s(g_{ij}, 1 - g_{ij})M \cdot (\overline{g}(t), 1 - \overline{g}(t)), \tag{7.1}$$

où $(g_{ij}, 1 - g_{ij})$ représente les probabilités phénotypiques d'un individu A_iA_j, $(\overline{g}(t), 1 - \overline{g}(t))$ les probabilités phénotypiques d'un individu rencontré au hasard à la génération t (voir (7.3)) et M donne les répercutions de ces interactions. Finalement, le tout est modulé en fonction de l'intensité de sélection $s > 0$.

Une génération débute avec une infinité de gamètes qui sont couplés au hasard pour donner une infinité de jeunes individus. Ces jeunes sont affectés par la sélection, ce qui changera les fréquences génotypiques chez les individus matures qui seront données par

$$y_{ij}^{m}(t) = y_{ij}(t)\frac{f_{ij}(t)}{\overline{f}(t)}. \tag{7.2}$$

Parmi ces individus matures seront échantillonnés N géniteurs qui produiront une infinité de gamètes, gamètes à l'origine de la génération suivante.

Notons tout d'abord que certaines relation établies sur le modèle de Wright-Fisher diploïde à coefficients de sélection constants sont toujours valides dans le cas présent. Ainsi, nous pouvons réutiliser les équations (6.3), (6.5) et (6.7).

Il y a un lien entre les variables g_{ij} et h_i. Par définition, on a

$$\overline{g}(t) = \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}(t)g_{ij}. \tag{7.3}$$

Remplaçons la valeur de g_{ij} pour obtenir

$$\overline{g}(t) = \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}(t)\left(\frac{h_i + h_j}{2}\right), \tag{7.4}$$

ce qui s'écrit aussi comme

$$\overline{g}(t) = \frac{1}{2}\left(\sum_{i=1}^{n} h_i \sum_{j=1}^{n} y_{ij}(t) + \sum_{j=1}^{n} h_j \sum_{i=1}^{n} y_{ij}(t)\right). \tag{7.5}$$

À l'aide de (6.3), on remarque que les deux sommes sont identiques, et donc

$$\overline{g}(t) = \sum_{i=1}^{n} h_i \sum_{j=1}^{n} y_{ij}(t), \tag{7.6}$$

ce qui donne

$$\overline{g}(t) = \sum_{i=1}^{n} h_i x_i(t), \tag{7.7}$$

où $x_i(t) = \sum_{j=1}^{n} y_{ij}(t)$. On a donc montré que les deux moyennes sont égales, i.e.

$$\overline{g}(t) = \overline{h}(t). \tag{7.8}$$

Sachant les fréquences génotypiques chez les jeunes, on a mentionné dans les spécificités du modèle la valeur de $y_{ij}^m(t)$, mais qu'en est-il de $x_i^m(t)$? Sa valeur est induite par (6.7), qui dit que

$$x_i^m(t) = \sum_{j=1}^{n} y_{ij}^m(t), \tag{7.9}$$

et donc que

$$x_i^m(t) = \sum_{j=1}^{n} y_{ij}(t) \frac{f_{ij}(t)}{\overline{f}(t)}, \tag{7.10}$$

ou bien encore

$$x_i^m(t) = \sum_{j=1}^{n} y_{ij}(t) \left(\frac{1 + s(g_{ij}, 1-g_{ij}) M \cdot (\overline{g}(t), 1-\overline{g}(t))}{1 + s(\overline{g}(t), 1-\overline{g}(t)) M \cdot (\overline{g}(t), 1-\overline{g}(t))} \right). \tag{7.11}$$

On a déjà mentionné que $y_{ij}(t) = x_i(t)x_j(t)$. Du coup, profitons-en pour récrire le dénominateur à l'aide de (7.8). On a

$$x_i^m(t) = x_i(t) \left(\frac{1 + s\sum_{j=1}^{n} x_j(t)(g_{ij}, 1-g_{ij}) M \cdot (\overline{g}(t), \mathbf{1} - \overline{g}(t))}{1 + s(\overline{g}(t), 1-\overline{g}(t)) M \cdot (\overline{g}(t), 1-\overline{g}(t))} \right), \tag{7.12}$$

ce qu'on écrira

$$x_i^m(t) = x_i(t) \left(\frac{1 + s(\overline{g_i}(t), 1-\overline{g_i}(t)) M \cdot (\overline{g}(t), \mathbf{1} - \overline{g}(t))}{1 + s(\overline{g}(t), 1-\overline{g}(t)) M \cdot (\overline{g}(t), 1-\overline{g}(t))} \right), \tag{7.13}$$

où

$$\overline{g_i}(t) := \sum_{j=1}^{n} x_j(t) g_{ij} \tag{7.14}$$

prend aussi comme valeur

$$\overline{g_i}(t) = \sum_{j=1}^{n} x_j(t) \left(\frac{h_i + h_j}{2} \right), \tag{7.15}$$

ou bien encore

$$\overline{g_i}(t) = \frac{1}{2}\left(h_i + \overline{h}(t)\right). \tag{7.16}$$

7.2 Espérance du changement de la fréquence allélique

Nous somme maintenant prêts à calculer

$$E_s\left[\Delta x_i(t)|\vec{x}(t)\right] = x_i^m(t) - x_i(t). \tag{7.17}$$

En effet, l'espérance de $x_i(t+1)$ est aussi l'espérance de la fréquence de l'allèle A_i chez les N géniteurs de la génération t. Ceux-ci sont pigés au hasard parmi les individus matures, cette dernière vaut donc $x_i^m(t)$. Reprenant (7.13), on peut dire que

$$E_s\left[\Delta x_i(t)|\vec{x}(t)\right] = x_i(t)\left(\frac{1 + s(\overline{g_i}(t), 1-\overline{g_i}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t))}{1 + s(\overline{g}(t), 1-\overline{g}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t))}\right) - x_i(t), \tag{7.18}$$

ce qui nous donne la dérivée

$$\begin{aligned}
&E_s'\left[\Delta x_i(t)|\vec{x}(t)\right] = \\
&\quad x_i(t)\left[1 + s(\overline{g}(t), 1-\overline{g}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t))\right]^{-2} \\
&\quad \left[(\overline{g_i}(t), 1-\overline{g_i}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t))\,(1 + s(\overline{g}(t), 1-\overline{g}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t)))\right. \\
&\quad \left. - (\overline{g}(t), 1-\overline{g}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t))\,(1 + s(\overline{g_i}(t), 1-\overline{g_i}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t)))\right],
\end{aligned} \tag{7.19}$$

qu'on peut évaluer à $s = 0$ pour obtenir

$$\phi_i(t) = x_i(t)\left[(\overline{g_i}(t), 1-\overline{g_i}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t)) - (\overline{g}(t), 1-\overline{g}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t))\right]. \tag{7.20}$$

Ceci est égal à

$$\phi_i(t) = x_i(t)(\overline{g_i}(t) - \overline{g}(t), \overline{g}(t) - \overline{g_i}(t))M\cdot(\overline{g}(t), 1-\overline{g}(t)). \tag{7.21}$$

En se rappelant (7.8) et (7.16), on peut récrire ceci comme

$$\phi_i(t) = x_i(t)\left(\frac{h_i + \overline{h}(t)}{2} - \overline{h}(t)\right)(1,-1)M\cdot\left(\overline{h}(t), 1-\overline{h}(t)\right), \tag{7.22}$$

qui se simplifie en

$$\phi_i(t) = \frac{x_i(t)}{2}\left(\overline{h_i}(t) - \overline{h}(t)\right)(1,-1)M \cdot \left(\overline{h}(t), 1-\overline{h}(t)\right). \tag{7.23}$$

7.3 Calcul de la probabilité de fixation

Pour calculer l'espérance sous neutralité de la dernière expression, il faut mettre en évidence les fréquences alléliques. En effectuant le produit matriciel, on obtient

$$\phi_i(t) = \frac{x_i(t)}{2}\left(\overline{h_i}(t) - \overline{h}(t)\right)\left((m_{11}-m_{21})\overline{h}(t) + (m_{12}-m_{22})\left(1-\overline{h}(t)\right)\right). \tag{7.24}$$

On cherche à regrouper les termes en puissances de $\overline{h}(t)$. On a premièrement que

$$\phi_i(t) = \frac{x_i(t)}{2}\left(\overline{h_i}(t) - \overline{h}(t)\right)\left((m_{11}-m_{21}-m_{12}+m_{22})\overline{h}(t) + (m_{12}-m_{22})\right), \tag{7.25}$$

ce qui nous mène à

$$\begin{aligned}\phi_i(t) = \frac{x_i(t)}{2}\Big[&h_i(m_{12}-m_{22}) + \overline{h}(t)\left(h_i(m_{11}-m_{21}-m_{12}+m_{22}) - (m_{12}-m_{22})\right)\\ &-\overline{h}^2(t)(m_{11}-m_{21}-m_{12}+m_{22})\Big].\end{aligned} \tag{7.26}$$

Si on définit

$$a := \tfrac{1}{2}(m_{12}-m_{22}), \tag{7.27}$$

$$b := \tfrac{1}{2}(m_{11}-m_{21}-m_{12}+m_{22}), \tag{7.28}$$

on peut récrire

$$\phi_i(t) = x_i(t)\left[ah_i + (bh_i - a)\,\overline{h}(t) - b\overline{h}^2(t)\right]. \tag{7.29}$$

Vient alors le temps d'expliciter les moyennes, ce qui donne

$$\phi_i(t) = ax_i(t)h_i + (bh_i - a)\sum_{j=1}^{n} x_i(t)x_j(t)h_j - b\sum_{j=1}^{n}\sum_{k=1}^{n} x_i(t)x_j(t)x_k(t)h_jh_k. \tag{7.30}$$

Il faut extirper des sommes les termes où les indices sont identiques. Ce faisant, on trouve

$$\phi_i(t) = ax_i(t)h_i + (bh_i - a)\,x_i^2(t)h_i + (bh_i - a)\sum_{\substack{j=1\\ j\neq i}}^{n} x_i(t)x_j(t)h_j - bx_i^3(t)h_i^2$$
$$- b\sum_{\substack{j=1\\ j\neq i}}^{n} x_i(t)x_j^2(t)h_j^2 - 2b\sum_{\substack{j=1\\ j\neq i}}^{n} x_i^2(t)x_j(t)h_ih_j - b\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i,j}}^{n} x_i(t)x_j(t)x_k(t)h_jh_k. \qquad (7.31)$$

On utilise maintenant les espérances (1.54) à (1.59) pour affirmer que

$$E_0\left[\phi_i(t)\right] = ax_i(0)h_i + (bh_i - a)\,x_i(0)\left(1 + p_{22}^t\left(x_i(0) - 1\right)\right)h_i$$
$$- bx_i(0)\left(1 + \left(\frac{x_i(0) - 1}{2}\right)\left(3p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right)\right)h_i^2$$
$$+ (bh_i - a)\sum_{\substack{j=1\\ j\neq i}}^{n} p_{22}^t x_i(0)x_j(0)h_j - b\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{x_i(0)x_j(0)}{2}\left(p_{22}^t + p_{33}^t\left(2x_j(0) - 1\right)\right)h_j^2$$
$$- 2b\sum_{\substack{j=1\\ j\neq i}}^{n}\left(\frac{x_i(0)x_j(0)}{2}\right)\left(p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right)h_ih_j$$
$$- b\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i,j}}^{n} p_{33}^t x_i(0)x_j(0)x_k(0)h_jh_k. \qquad (7.32)$$

On peut mettre $x_i(0)$ en évidence. Il faut aussi compléter les sommes pour avoir la chance de voir apparaître les moyennes. En annulant deux paires de termes opposés par leurs signes, nous transformons la dernière expression en

$$E_0\left[\phi_i(t)\right] = x_i(0)\Bigg[bh_i^2p_{22}^t\left(x_i(0) - 1\right) - ah_ip_{22}^t\left(x_i(0) - 1\right)$$
$$- bh_i^2\left(\frac{x_i(0) - 1}{2}\right)\left(3p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right) - \left(bh_i - a\right)h_ip_{22}^tx_i(0)$$
$$+ (bh_i - a)\sum_{j=1}^{n} h_jp_{22}^tx_j(0) + bh_i^2\frac{x_i(0)}{2}\left(p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right)$$
$$- b\sum_{j=1}^{n} h_j^2\frac{x_j(0)}{2}\left(p_{22}^t + p_{33}^t\left(2x_j(0) - 1\right)\right) + 2bh_i^2\frac{x_i(0)}{2}\left(p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right)$$
$$- 2bh_i\sum_{j=1}^{n} h_j\frac{x_j(0)}{2}\left(p_{22}^t + p_{33}^t\left(2x_i(0) - 1\right)\right) + b\sum_{j=1}^{n} h_j^2p_{33}^tx_j^2(0)$$
$$+ 2bh_i\sum_{j=1}^{n} h_jp_{33}^tx_i(0)x_j(0) - 2bh_i^2p_{33}^tx_i^2(0) - b\sum_{j=1}^{n}\sum_{k=1}^{n} h_jh_kp_{33}^tx_j(0)x_k(0)\Bigg]. \qquad (7.33)$$

On peut maintenant regrouper quelques termes. De plus, il y a simplification entre la première somme de la 4^e ligne et la dernière de la 5^e, pareillement pour les premières sommes des deux

dernières lignes. On trouve

$$E_0[\phi_i(t)] = x_i(0)\Bigg[-\frac{1}{2}bh_i^2p_{22}^t(x_i(0)-1) + ah_ip_{22}^t - bh_i^2\frac{(x_i(0)-1)}{2}\left(p_{33}^t(2x_i(0)-1)\right)$$
$$- bh_i^2p_{22}^tx_i(0) + (bh_i-a)p_{22}^t\sum_{j=1}^n h_jx_j(0) + \frac{3}{2}bh_i^2x_i(0)\left(p_{22}^t + p_{33}^t(2x_i(0)-1)\right)$$
$$- b\frac{p_{22}^t-p_{33}^t}{2}\sum_{j=1}^n h_j^2x_j(0) - bh_i\left(p_{22}^t-p_{33}^t\right)\sum_{j=1}^n h_jx_j(0) - 2bh_i^2p_{33}^tx_i^2(0)$$
$$- bp_{33}^t\sum_{j=1}^n h_jx_j(0)\sum_{k=1}^n h_kx_k(0)\Bigg]. \tag{7.34}$$

On a profité de la dernière réécriture pour mettre en évidence deux moyennes, que nous remplacerons ici. En continuant les simplifications, on obtient

$$E_0[\phi_i(t)] = x_i(0)\Bigg[\frac{1}{2}bh_i^2p_{22}^t + ah_ip_{22}^t + bh_i^2\left(\frac{2x_i(0)+1}{2}\right)\left(p_{33}^t(2x_i(0)-1)\right) - ap_{22}^t\overline{h}(0)$$
$$- b\left(\frac{p_{22}^t-p_{33}^t}{2}\right)\overline{h^2}(0) + bh_ip_{33}^t\overline{h}(0) - 2bh_i^2p_{33}^tx_i^2(0) - bp_{33}^t\overline{h}^2(0)\Bigg], \tag{7.35}$$

où $\overline{h}(0) = \sum_{j=1}^n h_jx_j(0)$ et $\overline{h^2}(0) = \sum_{j=1}^n h_j^2x_j(0)$. Il reste une dernière simplification à faire. Pour réaliser ceci, développons le troisième terme. Ainsi, on a

$$E_0[\phi_i(t)] = x_i(0)\Bigg[\frac{1}{2}bh_i^2p_{22}^t + ah_ip_{22}^t + bh_i^2p_{33}^t\left(2x_i^2(0) - x_i(0) + x_i(0) - \frac{1}{2}\right) - ap_{22}^t\overline{h}(0)$$
$$- b\frac{p_{22}^t-p_{33}^t}{2}\overline{h^2}(0) + bh_ip_{33}^t\overline{h}(0) - 2bh_i^2p_{33}^tx_i^2(0) - bp_{33}^t\overline{h}^2(0)\Bigg]. \tag{7.36}$$

Après avoir annulé les deux termes ainsi exposés, il ne reste qu'à regrouper les termes en facteurs de p_{22}^t et p_{33}^t. Ce faisant, on trouve

$$E_0[\phi_i(t)] = x_i(0)\Bigg[p_{22}^t\left(\frac{1}{2}bh_i^2 + ah_i - a\overline{h}(0) - b\frac{1}{2}\overline{h^2}(0)\right)$$
$$+ p_{33}^t\left(-\frac{1}{2}bh_i^2 + b\frac{1}{2}\overline{h^2}(0) + bh_i\overline{h}(0) - b\overline{h}^2(0)\right)\Bigg]. \tag{7.37}$$

On peut finalement manipuler un peu l'expression pour lui donner du sens. La dernière égalité est équivalente à

$$E_0[\phi_i(t)] = x_i(0)\Bigg[p_{22}^t\left(a\left(h_i-\overline{h}(0)\right) + \frac{b}{2}\left(h_i^2-\overline{h^2}(0)\right)\right)$$
$$+ p_{33}^tb\left(\overline{h}(0)\left(h_i-\overline{h}(0)\right) - \frac{1}{2}\left(h_i^2-\overline{h^2}(0)\right)\right)\Bigg]. \tag{7.38}$$

Le dernier résultat intégré dans (1.27) nous dit que la probabilité de fixation éventuelle de l'allèle

A_i est égale à

$$P_s(F_i) = x_i(0) + s\sum_{t\geq 0} x_i(0)\Big[p_{22}^t\Big(a\left(h_i - \overline{h}(0)\right) + \frac{b}{2}\left(h_i^2 - \overline{h^2}(0)\right)\Big) + p_{33}^t b\Big(\overline{h}(0)\left(h_i - \overline{h}(0)\right) - \frac{1}{2}\left(h_i^2 - \overline{h^2}(0)\right)\Big)\Big] + o(s). \tag{7.39}$$

Rappelons que a et b ne sont que des constantes. Il n'y a dans la série que p_{22}^t et p_{33}^t qui dépendent de l'indice t. On a donc deux séries géométriques dont on peut calculer les valeurs pour obtenir

$$P_s(F_i) = x_i(0) + sx_i(0)\left[\frac{a\left(h_i - \overline{h}(0)\right) + \frac{b}{2}\left(h_i^2 - \overline{h^2}(0)\right)}{1 - p_{22}} + \frac{b\Big(\overline{h}(0)\left(h_i - \overline{h}(0)\right) - \frac{1}{2}\left(h_i^2 - \overline{h^2}(0)\right)\Big)}{1 - p_{33}}\right] + o(s). \tag{7.40}$$

Les probabilités de non coalescence sont les mêmes que dans le modèle de Wright-Fisher à coefficients de sélection constants, on peut donc réutiliser (6.40) et (6.38) pour obtenir la probabilité de fixation

$$P_s(F_i) = x_i(0) + sNx_i(0)\Big[2\Big(a\left(h_i - \overline{h}(0)\right) + \frac{b}{2}\left(h_i^2 - \overline{h^2}(0)\right)\Big) + \frac{2N}{3N-1}b\Big(\overline{h}(0)\left(h_i - \overline{h}(0)\right) - \frac{1}{2}\left(h_i^2 - \overline{h^2}(0)\right)\Big)\Big] + o(s). \tag{7.41}$$

Étudions maintenant quelles conditions sont favorables à l'allèle A_i. En observant (7.47), on voit que $P_s(F_i) > x_i(0)$ pour s assez petit si et seulement si

$$2\left(a + \frac{Nb\overline{h}(0)}{3N-1}\right)\left(h_i - \overline{h}(0)\right) + b\left(1 - \frac{N}{3N-1}\right)\left(h_i^2 - \overline{h^2}(0)\right) > 0. \tag{7.42}$$

En multipliant par $(3N-1)/N$, ceci est équivalent à

$$2\left(a\left(3 - \frac{1}{N}\right) + b\overline{h}(0)\right)\left(h_i - \overline{h}(0)\right) + b\left(3 - \frac{1}{N} - 1\right)\left(h_i^2 - \overline{h^2}(0)\right) > 0, \tag{7.43}$$

ou, une fois simplifié, à

$$\left(3a + b\overline{h}(0) - \frac{a}{N}\right)h_i + \left(b - \frac{b}{2N}\right)h_i^2 > \left(3a + b\overline{h}(0) - \frac{a}{N}\right)\overline{h}(0) + \left(b - \frac{b}{2N}\right)\overline{h^2}(0). \tag{7.44}$$

Ainsi,

$$v_i\left(N, \vec{x}(0)\right) := \left(3a + b\overline{h}(0) - \frac{a}{N}\right)h_i + \left(b - \frac{b}{2N}\right)h_i^2, \tag{7.45}$$

agit en tant que coefficient de sélection effectif, dans ce sens que la fixation de l'allèle A_i est avantagée si

$$v_i\left(N, \vec{x}(0)\right) > \overline{v}\left(N, \vec{x}(0)\right) = \sum_{j=1}^{n} v_j\left(N, \vec{x}(0)\right) x_j(0). \tag{7.46}$$

Pour une grande population, on pourrait approximer cette condition en tronquant les termes en $1/2N$ dans (7.45).

Résultat 7.1

Dans le modèle de Wright-Fisher diploïde avec interactions, la probabilité de fixation éventuelle de l'allèle A_i étant donné les fréquences alléliques initiales dans la population est

$$\begin{aligned} P_s(F_i) = x_i(0) + sNx_i(0)\Bigg[2\left(a + \frac{Nb\overline{h}(0)}{3N-1}\right)\left(h_i - \overline{h}(0)\right) \\ + b\left(1 - \frac{N}{3N-1}\right)\left(h_i^2 - \overline{h^2}(0)\right)\Bigg] + o(s), \end{aligned} \tag{7.47}$$

où

$$a = \tfrac{1}{2}\left(m_{12} - m_{22}\right), \tag{7.48}$$

$$b = \tfrac{1}{2}\left(m_{11} - m_{21} - m_{12} + m_{22}\right). \tag{7.49}$$

La fixation de l'allèle A_i est avantagée par la sélection si son coefficient de sélection effectif

$$v_i\left(N, \vec{x}(0)\right) = \left(3a + b\overline{h}(0) - \frac{a}{N}\right)h_i + \left(b - \frac{b}{2N}\right)h_i^2 \tag{7.50}$$

est supérieur à sa moyenne dans la population.

Conclusion

La méthode générale présentée dans ce mémoire est une généralisation aux modèles polymorphes de celle exposée par Lessard et Ladret dans [5] pour le modèle de Cannings à deux allèles. Cette adaptation s'est réalisée sans heurt, dans ce sens que les hypothèses de continuité de l'annexe A ne sont pas plus sévères que celles mises en place par Lessard et Ladret dans leur article. La méthode en question nous permet de trouver l'effet de premier ordre de la sélection sur la probabilité de fixation d'un allèle particulier dans un modèle génétique de population à n allèles. Ce résultat est d'autant plus précieux dans les cas polymorphes que ceux-ci, contrairement aux modèles à deux allèles, peuvent difficilement être approximés par des processus de diffusion.

Le travail effectué au premier chapitre nous permet par la suite de trouver aisément les probabilités de fixation pour les modèles haploïdes de Wright-Fisher et de Moran. Ces résultats respectent la même logique que leurs équivalents pour les cas bimorphiques, c'est-à-dire qu'un allèle est avantagé par la sélection si son coefficient de viabilité est plus élevé que la moyenne pour le modèle de Wright-Fisher, ou que son coefficient de mortalité est plus faible que la moyenne pour le modèle de Moran. La même cohérence apparaît pour le modèle de Wright-Fisher haploïde pour la sélection du sexe. Un allèle est avantagé si la probabilité de la fonction mâle qui y est associée est comprise entre 1/2 et la probabilité moyenne de la fonction mâle dans la population, ce qui est équivalent au résultat énoncé dans [6] pour le modèle à deux allèles. Il n'y a donc aucune surprise à avoir pour ces trois modèles. Du travail effectué sur les modèles haploïdes, le résultat à retenir est principalement la probabilité de fixation pour le modèle de Cannings à plusieurs allèles. Celui-ci est intéressant puisque le modèle de Cannings englobe potentiellement tous les autres modèles haploïdes.

L'étude du modèle de Wright-Fisher diploïde par le procédé exposé mène à un résultat très élégant. Pour une grande population évoluant selon ce modèle et pour une sélection assez faible, un allèle A_i est avantagé par la sélection si la somme du coefficient de sélection d'un individu homozygote A_iA_i et du coefficient de sélection moyen d'un individu hétérozygote A_iA_j (pour $j \in \{1,\ldots,n\}$) est supérieure à la moyenne de cette somme dans la population. Ce résultat est nouveau et suffisamment simple pour qu'on en tire une compréhension approfondie du modèle.

C'est un résultat intéressant. Pour ce qui est de l'autre modèle diploïde, celui de Wright-Fisher avec interactions, le résultat trouvé se rapporte sous certaines conditions au modèle haploïde à deux allèles.

Les modèles considérés ici sont variés et notre méthode peut s'appliquer à virtuellement n'importe quel modèle. Certains de ceux que nous avons étudiés, mais qui ne sont pas présentés dans ce mémoire, débouchaient sur des espérances de produits de fréquences alléliques de degré quatre. Ces espérances peuvent être calculées de la même manière que celles de degré inférieur en extrayant plus d'information du principe d'inclusion-exclusion. Cette restriction sur le degré des produits de fréquences servait de cadre de travail, mais ne doit aucunement être vu comme une limite inhérante à la procédure. Dans un autre ordre d'idée, il serait aussi intéressant de pousser l'analyse à l'effet de deuxième ordre de la sélection. Celle-ci serait nécessaire pour compléter rigoureusement l'étude du modèle de Wright-Fisher haploïde pour la détermination du sexe. Nous croyons que l'étude d'autres modèles génétiques de populations polymorphes par la méthode présentée dans ce mémoire pourrait mener rapidement à de nouveaux résultats.

Annexe A

Preuve de la convergence uniforme

Nous devons montrer que, dans le cadre général décrit au chapitre 1, la série

$$\sum_{t\geqslant 0} \frac{\partial}{\partial s} E_s\left[\Delta x_i(t)\right] \tag{A.1}$$

converge uniformément pour un voisinage de $s = 0$. Considérons la somme partielle jusqu'à $T-1 \geqslant 0$. On a que

$$\sum_{t=0}^{T-1} \frac{\partial}{\partial s} E_s\left[\Delta x_i(t)\right] = \frac{\partial}{\partial s} E_s\left[\sum_{t=0}^{T-1} \Delta x_i(t)\right]. \tag{A.2}$$

L'équation (1.12) nous rappelle que $\sum_{t=0}^{T-1} \Delta x_i(t)$ est $(x_i(T) - x_i(0))$, ce qui nous permet de dire que

$$\sum_{t=0}^{T-1} \frac{\partial}{\partial s} E_s\left[\Delta x_i(t)\right] = \frac{\partial}{\partial s} E_s\left[x_i(T)\right]. \tag{A.3}$$

Il suffit donc de montrer la convergence uniforme du membre de droite de l'équation ci-dessus. Or, si on écrit

$$E_s\left[x_i(T)\right] = \sum_{k=0}^{2N} P_s\left(x_i(T) = \frac{k}{2N}\right)\frac{k}{2N}, \tag{A.4}$$

on réalise qu'il suffit de montrer la convergence uniforme de la dérivée de la matrice de transition en T pas, $\mathbf{P}_s^{(T)}$, de la chaîne de Markov associée aux états de la population de génération en génération, c'est-à-dire de la chaîne $\vec{x}(t)$ décrite à la section 1.3. Nous donnons dans cet annexe la preuve de la convergence de $\partial \mathbf{P}_s^{(T)}/\partial s$ sous l'hypothèse que la matrice de transition en un pas, $\mathbf{P}_s$, et sa dérivée par rapport à s sont continues à $s = 0$. Cette preuve a été faite par Lessard et Ladret pour le cas d'une population à deux allèles dans [5, p.740].

On a déjà remarqué que la chaîne comporte

$$\binom{2N+n-1}{n-1} \tag{A.5}$$

états possibles, dont n états absorbants. On définit

$$m := \binom{2N+n-1}{n-1} - n \tag{A.6}$$

comme le nombre d'états transitoires. En plaçant les états transitoires en premier et les états absorbants en dernier, on obtient une matrice de transition de la forme

$$\mathbf{P}_s = \left[\begin{array}{c:c} Q_s & R_s \\ \hdashline O_{n\times m} & I \end{array}\right], \tag{A.7}$$

où Q_s est la matrice $m \times m$ des probabilités de transition entre les états transitoires, R_s la matrice $m \times n$ des probabilités de transition des états transitoires vers les états absorbants, I la matrice identité $n \times n$ associée aux états absorbants, et $O_{n\times m}$ une matrice nulle de dimension $n \times m$. On peut déduire la structure de la matrice de transition en T pas en calculant la T-ième puissance de $\mathbf{P}_s$. Par induction, on obtient

$$\mathbf{P}_s^{(T)} = \mathbf{P}_s^T = \left[\begin{array}{c:c} Q_s^T & \sum_{i=0}^{T-1} Q_s^i R_s \\ \hdashline O_{n\times m} & I \end{array}\right]. \tag{A.8}$$

La règle de dérivation d'un produit matriciel donne

$$\frac{\partial \mathbf{P}_s^{(T)}}{\partial s} = \frac{\partial \mathbf{P}_s^T}{\partial s} = \sum_{t=0}^{T-1} \mathbf{P}_s^t \frac{\partial \mathbf{P}_s}{\partial s} \mathbf{P}_s^{T-t-1}. \tag{A.9}$$

Si on combine ces deux résultats, on a

$$\frac{\partial \mathbf{P}_s^{(T)}}{\partial s} = \sum_{t=0}^{T-1} \left[\begin{array}{c:c} Q_s^t & \sum_{i=0}^{t-1} Q_s^i R_s \\ \hdashline O_{n\times m} & I \end{array}\right] \left[\begin{array}{c:c} \frac{\partial Q_s}{\partial s} & \frac{\partial R_s}{\partial s} \\ \hdashline O_{n\times m} & O_{n\times n} \end{array}\right] \left[\begin{array}{c:c} Q_s^{T-t-1} & \sum_{i=0}^{T-t-1} Q_s^i R_s \\ \hdashline O_{n\times m} & I \end{array}\right]. \tag{A.10}$$

Suite à la multiplication par bloc des deux premières matrices, on trouve que

$$\frac{\partial \mathbf{P}_s^{(T)}}{\partial s} = \sum_{t=0}^{T-1} \left[\begin{array}{c:c} Q_s^t \frac{\partial Q_s}{\partial s} & Q_s^t \frac{\partial R_s}{\partial s} \\ \hdashline O_{n\times m} & O_{n\times n} \end{array}\right] \left[\begin{array}{c:c} Q_s^{T-t-1} & \sum_{i=0}^{T-t-1} Q_s^i R_s \\ \hdashline O_{n\times m} & I \end{array}\right], \tag{A.11}$$

puis, en répétant la même opération, que

$$\frac{\partial \mathbf{P}_s^{(T)}}{\partial s} = \sum_{t=0}^{T-1} \left[\begin{array}{c:c} Q_s^t \frac{\partial Q_s}{\partial s} Q_s^{T-t-1} & Q_s^t \frac{\partial Q_s}{\partial s} \sum_{i=0}^{T-t-1} Q_s^i R_s + Q_s^t \frac{\partial R_s}{\partial s} \\ \hdashline O_{n\times m} & O_{n\times n} \end{array} \right]. \tag{A.12}$$

En définissant les matrices

$$A_s^{(T)} := \sum_{t=0}^{T-1} Q_s^t, \tag{A.13}$$

$$B_s^{(T)} := \sum_{t=0}^{T-1} Q_s^t \frac{\partial Q_s}{\partial s} Q_s^{T-t-1}, \tag{A.14}$$

$$C_s^{(T)} := \sum_{t=0}^{T-1} Q_s^t \frac{\partial Q_s}{\partial s} \sum_{i=0}^{T-t-1} Q_s^i, \tag{A.15}$$

on peut récrire (A.12) comme

$$\frac{\partial \mathbf{P}_s^{(T)}}{\partial s} = \left[\begin{array}{c:c} B_s^{(T)} & C_s^{(T)} R_s + A_s^{(T)} \frac{\partial R_s}{\partial s} \\ \hdashline O_{n\times m} & O_{n\times n} \end{array} \right]. \tag{A.16}$$

On va montrer que cette matrice converge uniformément vers

$$\frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} := \left[\begin{array}{c:c} O_{m\times m} & (I - Q_s)^{-1} \frac{\partial Q_s}{\partial s} (I - Q_s)^{-1} R_s + (I - Q_s)^{-1} \frac{\partial R_s}{\partial s} \\ \hdashline O_{n\times m} & O_{n\times n} \end{array} \right]. \tag{A.17}$$

Avant de procéder, on va devoir montrer que l'inverse de la matrice $(I - Q_s)$ existe.

Si on considère λ, une valeur propre de Q_s, et $\vec{v} = (v_1, \ldots, v_m) \neq \vec{0}$, un vecteur propre associé, on a par définition

$$Q_s \vec{v} = \lambda \vec{v}. \tag{A.18}$$

Par induction, on obtient que

$$Q_s^T \vec{v} = \lambda^T \vec{v} \tag{A.19}$$

pour $T \geqslant 1$. Comme $\vec{v}$ est non nul, il existe un indice i pour lequel $v_i \neq 0$. Pour tout indice, dont celui-ci, on a

$$\left[Q_s^T \vec{v} \right]_i = \lambda^T v_i. \tag{A.20}$$

Or, le théorème ergodique pour les chaînes de Markov à temps discret garantit que la matrice Q_s^T, dont les entrées sont des probabilités de transition en T pas vers des états transitoires, tend

vers $O_{m\times m}$ quand T tend vers l'infini (pour référence, voir [1, p.48]). Donc, on a

$$\lim_{T\to\infty}\left[Q_s^T\vec{v}\right]_i = 0, \tag{A.21}$$

d'où

$$\lim_{T\to\infty}\lambda^T v_i = 0. \tag{A.22}$$

Puisque $v_i \neq 0$, ceci entraîne que $|\lambda| < 1$, et donc que 1 n'est pas une valeur propre de la matrice Q_s, i.e.

$$\det(I - Q_s) \neq 0, \tag{A.23}$$

d'où la matrice $(I - Q_s)$ est inversible.

Pour montrer la convergence uniforme vers (A.17), nous devons définir une norme sur ces matrices, qui sont carrées. On utilisera la norme infinie qui, pour une matrice carrée X de dimension l dont les entrées sont X_{ij} pour $i, j \in \{1, \ldots, l\}$, est définie par

$$\|X\| = \max_{1\leqslant i\leqslant l}\sum_{j=1}^{l}|X_{ij}|. \tag{A.24}$$

Cette norme est sous-additive et sous-multiplicative. C'est-à-dire que, pour toutes matrices carrées de même dimension X et Y, on a

$$\|X + Y\| \leqslant \|X\| + \|Y\| \tag{A.25}$$

et

$$\|XY\| \leqslant \|X\|\|Y\|. \tag{A.26}$$

On considère

$$\begin{aligned}&\left\|\frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s}\right\| \\ &= \left\| \begin{array}{c|c} B_s^{(T)} & \left[C_s^{(T)} - (I-Q_s)^{-1}\frac{\partial Q_s}{\partial s}(I-Q_s)^{-1}\right]R_s + \left[A_s^{(T)} - (I-Q_s)^{-1}\right]\frac{\partial R_s}{\partial s} \\ \hline O_{n\times m} & O_{n\times n} \end{array} \right\|.\end{aligned} \tag{A.27}$$

Par sous-additivité et sous-multiplicativité, on obtient que

$$\left\|\frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s}\right\| \leqslant \left\|B_s^{(T)}\right\| + \left\|C_s^{(T)} - (I-Q_s)^{-1}\frac{\partial Q_s}{\partial s}(I-Q_s)^{-1}\right\| \left\|\begin{array}{c|c} O_{m\times m} & R_s \\ \hline O_{n\times m} & O_{n\times n} \end{array}\right\| + \left\|A_s^{(T)} - (I-Q_s)^{-1}\right\| \left\|\begin{array}{c|c} O_{m\times m} & \frac{\partial R_s}{\partial s} \\ \hline O_{n\times m} & O_{n\times n} \end{array}\right\|. \tag{A.28}$$

Sous l'hypothèse que $\mathbf{P}_s$ et $\partial \mathbf{P}_s/\partial s$ sont continues à $s = 0$, il en est de même des sous-matrices R_s et $\partial R_s/\partial s$. Il existe donc un voisinage V_1 de $s = 0$ dans lequel

$$\left\|\begin{array}{c|c} O_{m\times m} & R_s \\ \hline O_{n\times m} & O_{n\times n} \end{array}\right\| \leqslant \alpha \tag{A.29}$$

et

$$\left\|\begin{array}{c|c} O_{m\times m} & \frac{\partial R_s}{\partial s} \\ \hline O_{n\times m} & O_{n\times n} \end{array}\right\| \leqslant \alpha, \tag{A.30}$$

pour une constante $0 < \alpha < \infty$, et donc où

$$\left\|\frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s}\right\| \leqslant \left\|B_s^{(T)}\right\| + \left\|C_s^{(T)} - (I-Q_s)^{-1}\frac{\partial Q_s}{\partial s}(I-Q_s)^{-1}\right\|\alpha + \left\|A_s^{(T)} - (I-Q_s)^{-1}\right\|\alpha. \tag{A.31}$$

Il est possible de récrire $C_s^{(T)}$ donné en (A.15) sous la forme

$$C_s^{(T)} = \sum_{t=0}^{T-1} Q_s^t \frac{\partial Q_s}{\partial s}(1 - Q_s^{T-t})(1-Q_s)^{-1}. \tag{A.32}$$

On remarque alors la présence de $A_s^{(T)}$ et $B_s^{(T)}$, donnés en (A.13) et (A.14) respectivement, ce qu'on peut exploiter pour écrire

$$C_s^{(T)} = A_s^{(T)}\frac{\partial Q_s}{\partial s}(1-Q_s)^{-1} - B_s^{(T)}Q_s(I-Q_s)^{-1}. \tag{A.33}$$

Ainsi, on a

$$C_s^{(T)} - (I-Q_s)^{-1}\frac{\partial Q_s}{\partial s}(I-Q_s)^{-1} = \left(A_s^{(T)} - (I-Q_s)^{-1}\right)\frac{\partial Q_s}{\partial s}(1-Q_s)^{-1} - B_s^{(T)}Q_s(I-Q_s)^{-1}, \tag{A.34}$$

ce qui permet d'obtenir par sous-additivité et sous-multiplicativité de la norme infinie que

$$\left\| C_s^{(T)} - (I - Q_s)^{-1} \frac{\partial Q_s}{\partial s} (I - Q_s)^{-1} \right\| \leqslant \left\| A_s^{(T)} - (I - Q_s)^{-1} \right\| \left\| \frac{\partial Q_s}{\partial s} \right\| \left\| (1 - Q_s)^{-1} \right\| + \left\| B_s^{(T)} \right\| \left\| Q_s \right\| \left\| (I - Q_s)^{-1} \right\|. \tag{A.35}$$

En utilisant cette borne dans (A.31), on trouve

$$\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant \left\| B_s^{(T)} \right\| \left(1 + \left\| Q_s \right\| \left\| (I - Q_s)^{-1} \right\| \alpha \right) + \alpha \left\| A_s^{(T)} - (I - Q_s)^{-1} \right\| \left(\left\| \frac{\partial Q_s}{\partial s} \right\| \left\| (1 - Q_s)^{-1} \right\| + 1 \right). \tag{A.36}$$

Sous l'hypothèse que $\mathbf{P}_s$ et $\partial \mathbf{P}_s / \partial s$ sont continues à $s = 0$, il en est de même pour Q_s, $\partial Q_s / \partial s$ et aussi $(I - Q_s)^{-1}$. Ainsi, il existe un voisinage $V_2 \subseteq V_1$ de $s = 0$ où les normes $\|\partial Q_s / \partial s\|$ et $\left\|(I - Q_s)^{-1}\right\|$ sont bornées par une constante $0 < \beta < \infty$. Notons que pour tout $s \geqslant 0$, la norme de Q_s est inférieure ou égale à 1, car c'est une sous-matrice d'une matrice de transition. Pour tout s dans V_2, on a donc

$$\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant \left\| B_s^{(T)} \right\| (1 + \beta\alpha) + \alpha \left\| A_s^{(T)} - (I - Q_s)^{-1} \right\| \left(\beta^2 + 1 \right). \tag{A.37}$$

Puisque

$$A_s^{(T)} = \sum_{t=0}^{T-1} Q_s^t = (I - Q_s)^{-1} (I - Q_s^T), \tag{A.38}$$

on trouve que

$$A_s^{(T)} - (I - Q_s)^{-1} = (I - Q_s)^{-1} (-Q_s^T), \tag{A.39}$$

et donc

$$\left\| A_s^{(T)} - (I - Q_s)^{-1} \right\| \leqslant \left\| (I - Q_s)^{-1} \right\| \left\| Q_s^T \right\| \leqslant \beta \left\| Q_s^T \right\| \tag{A.40}$$

dans le voisinage V_2.

D'autre part, on a

$$B_s^{(T)} = \sum_{t=0}^{\lfloor T/2 \rfloor - 1} Q_s^t \frac{\partial Q_s}{\partial s} Q_s^{T-t-1} + \sum_{t=\lfloor T/2 \rfloor}^{T-1} Q_s^t \frac{\partial Q_s}{\partial s} Q_s^{T-t-1}, \tag{A.41}$$

pour $T \geqslant 2$, où $\lfloor T/2 \rfloor$ désigne la valeur entière de $T/2$. Dans la première des deux sommes du

membre de droite de l'équation ci-dessus, on a

$$T - t - 1 \geqslant T - (\lfloor T/2 \rfloor - 1) - 1 = T - \lfloor T/2 \rfloor \geqslant \lfloor T/2 \rfloor, \tag{A.42}$$

tandis que, dans la deuxième, $t \geqslant \lfloor T/2 \rfloor$. Ces deux inégalités nous permettent d'écrire

$$B_s^{(T)} = \left(\sum_{t=0}^{\lfloor T/2 \rfloor - 1} Q_s^t \frac{\partial Q_s}{\partial s} Q_s^{\lfloor T/2 \rfloor} Q_s^{T-t-1-\lfloor T/2 \rfloor} + \sum_{t=\lfloor T/2 \rfloor}^{T-1} Q_s^{\lfloor T/2 \rfloor} Q_s^{t-\lfloor T/2 \rfloor} \frac{\partial Q_s}{\partial s} Q_s^{T-t-1} \right), \tag{A.43}$$

où tous les exposants sont non négatifs. Par sous-additivité et sous-multiplicativité, on trouve

$$\left\| B_s^{(T)} \right\| \leqslant \left\| \frac{\partial Q_s}{\partial s} \right\| \left(\sum_{t=0}^{\lfloor T/2 \rfloor - 1} \left\| Q_s^t \right\| \left\| Q_s^{\lfloor T/2 \rfloor} \right\| \left\| Q_s \right\|^{T-t-1-\lfloor T/2 \rfloor} + \sum_{t=\lfloor T/2 \rfloor}^{T-1} \left\| Q_s^{\lfloor T/2 \rfloor} \right\| \left\| Q_s \right\|^{t-\lfloor T/2 \rfloor} \left\| Q_s^{T-t-1} \right\| \right). \tag{A.44}$$

Comme $\|Q_s\| \leqslant 1$, on obtient

$$\left\| B_s^{(T)} \right\| \leqslant \left\| \frac{\partial Q_s}{\partial s} \right\| \left\| Q_s^{\lfloor T/2 \rfloor} \right\| \left(\sum_{t=0}^{\lfloor T/2 \rfloor - 1} \left\| Q_s^t \right\| + \sum_{t=\lfloor T/2 \rfloor}^{T-1} \left\| Q_s^{T-t-1} \right\| \right), \tag{A.45}$$

d'où, en complétant les sommes partielles en séries jusqu'à l'infini,

$$\left\| B_s^{(T)} \right\| \leqslant 2 \left\| \frac{\partial Q_s}{\partial s} \right\| \left\| Q_s^{\lfloor T/2 \rfloor} \right\| \sum_{t \geqslant 0} \left\| Q_s^t \right\|, \tag{A.46}$$

et donc

$$\left\| B_s^{(T)} \right\| \leqslant 2\gamma \left\| Q_s^{\lfloor T/2 \rfloor} \right\| \sum_{t \geqslant 0} \left\| Q_s^t \right\|, \tag{A.47}$$

dans le voisinage V_2.

Si on intègre les résultats (A.40) et (A.47) à l'inégalité (A.37), on a que

$$\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant 2\beta (1 + \alpha\beta) \left\| Q_s^{\lfloor T/2 \rfloor} \right\| \sum_{t \geqslant 0} \left\| Q_s^t \right\| + \alpha\beta \left(\beta^2 + 1 \right) \left\| Q_s^T \right\|, \tag{A.48}$$

pour $T \geqslant 2$, dans le voisinage V_2. Il ne reste qu'à borner $\|Q_s^t\|$. Le théorème ergodique nous informe que Q_0^t tend vers $O_{n \times m}$ quand t tend vers l'infini. Du coup, $\|Q_0^t\|$ tend vers 0 et il existe un entier T_0 tel que

$$\left\| Q_0^{T_0} \right\| < 1. \tag{A.49}$$

Comme la matrice de transition $\mathbf{P}_s$ est continue à $s = 0$, alors Q_s, $Q_s^{T_0}$ et donc $\left\| Q_s^{T_0} \right\|$ le sont

aussi. Il est donc possible de trouver une constante $0 < \gamma < 1$ et un voisinage $V_3 \subseteq V_2$ de $s = 0$ dans lequel

$$\left\| Q_s^{T_0} \right\| < \gamma. \tag{A.50}$$

Ceci nous permettra de conclure.

Pour tout entier $k \geq 0$, il existe un unique entier positif $r(k)$ tel que $k = \lfloor k/T_0 \rfloor T_0 + r(k)$. On peut écrire $\lfloor T/2 \rfloor$, t et T sous cette forme dans (A.48), ce qui nous donne

$$\begin{aligned}\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant 2\beta(1+\beta\alpha) \left\| Q_s^{\lfloor \lfloor T/2 \rfloor / T_0 \rfloor T_0 + r(\lfloor T/2 \rfloor)} \right\| \sum_{t \geqslant 0} \left\| Q_s^{\lfloor t/T_0 \rfloor T_0 + r(t)} \right\| \\ + \alpha\beta\left(\beta^2+1\right) \left\| Q_s^{\lfloor T/T_0 \rfloor T_0 + r(T)} \right\|.\end{aligned} \tag{A.51}$$

En appliquant (A.25) et (A.26), on a

$$\begin{aligned}\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant 2\beta(1+\beta\alpha) \left\| Q_s^{T_0} \right\|^{\lfloor \lfloor T/2 \rfloor / T_0 \rfloor} \left\| Q_s \right\|^{r(\lfloor T/2 \rfloor)} \sum_{t \geqslant 0} \left\| Q_s^{T_0} \right\|^{\lfloor t/T_0 \rfloor} \left\| Q_s \right\|^{r(t)} \\ + \alpha\beta\left(\beta^2+1\right) \left\| Q_s^{T_0} \right\|^{\lfloor T/T_0 \rfloor} \left\| Q_s \right\|^{r(T)}.\end{aligned} \tag{A.52}$$

En utilisant le fait que $\|Q_s\| \leqslant 1$, on obtient

$$\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant 2\beta(1+\beta\alpha) \left\| Q_s^{T_0} \right\|^{\lfloor \lfloor T/2 \rfloor / T_0 \rfloor} \sum_{t \geqslant 0} \left\| Q_s^{T_0} \right\|^{\lfloor t/T_0 \rfloor} + \alpha\beta\left(\beta^2+1\right) \left\| Q_s^{T_0} \right\|^{\lfloor T/T_0 \rfloor}, \tag{A.53}$$

ce qui peut se récrire

$$\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant 2\beta(1+\beta\alpha) \left\| Q_s^{T_0} \right\|^{\lfloor \lfloor T/2 \rfloor / T_0 \rfloor} T_0 \sum_{t \geqslant 0} \left\| Q_s^{T_0} \right\|^{t} + \alpha\beta\left(\beta^2+1\right) \left\| Q_s^{T_0} \right\|^{\lfloor T/T_0 \rfloor}. \tag{A.54}$$

Dans le voisinage V_3, on a donc

$$\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant 2\beta(1+\beta\alpha) \gamma^{\lfloor \lfloor T/2 \rfloor / T_0 \rfloor} T_0 \sum_{t \geqslant 0} \gamma^t + \alpha\beta\left(\beta^2+1\right) \gamma^{\lfloor T/T_0 \rfloor}, \tag{A.55}$$

c'est-à-dire,

$$\left\| \frac{\partial \mathbf{P}_s^{(T)}}{\partial s} - \frac{\partial \mathbf{P}_s^{(\infty)}}{\partial s} \right\| \leqslant \frac{2\beta(1+\beta\alpha)T_0}{1-\gamma} \gamma^{\lfloor \lfloor T/2 \rfloor / T_0 \rfloor} + \alpha\beta\left(\beta^2+1\right) \gamma^{\lfloor T/T_0 \rfloor}. \tag{A.56}$$

Ceci est valide pour $T \geqslant 2$. De plus, la borne à droite, valide pour tout s dans V_3, tend vers 0 lorsque T tend vers l'infini, ce qui termine la preuve.

Annexe B

Ancêtres d'un échantillon de gènes sous neutralité

Dans une population évoluant en absence de sélection naturelle, si on pige au hasard un échantillon de b gènes à la génération t et que cet échantillon a a ancêtres à la génération 0, on affirme que tout groupe de a gènes de la génération 0 a chance égale d'être le groupe d'ancêtres de ces gènes. Cette assertion intuitivement raisonnable est utilisée tout au long de la section 1.7 sans plus de justification. Nous montrons ici qu'elle est valide.

Pour ce faire, on utilisera le modèle de Cannings sous neutralité. On ordonne les gènes de la génération 0 en plaçant d'abord ceux d'allèle A_1, puis ceux d'allèle A_2, etc, ce qui permet de définir $Z_1(0), \ldots, Z_{2N}(0)$ comme leur nombre de rejetons respectifs et $Z_1(t-1), \ldots, Z_{2N}(t-1)$ comme leur nombre de descendants à la génération $t \geqslant 1$. Remarquez que la notation est légèrement différente de celle utilisée au chapitre 5 où on réordonnait les gènes à chaque génération et où $Z_1(t-1)$ aurait représenté le nombre de rejetons du premier gène de la génération $(t-1)$. Le reste est identique : on a $2N$ gènes, n allèles et les variables aléatoires $x_j(t)$ sont leurs fréquences à la génération t. La somme des variables aléatoires $Z_1(t-1), \ldots, Z_{2N}(t-1)$ est de $2N$ à chaque génération pour que la taille de la population demeure constante. La caractérisation du modèle de Cannings neutre est que les variables aléatoires $Z_1(0), \ldots, Z_{2N}(0)$ sont échangeables. Ceci implique que les nombres de rejetons de chaque individu de la génération 1 sont des variables aléatoires échangeables et ainsi de suite pour les générations suivantes. On montrera plus loin que ceci implique l'échangeabilité pour tout $t \geqslant 1$ des variables aléatoires $Z_1(t-1), \ldots, Z_{2N}(t-1)$.

Dans ce modèle neutre, les entrées de la matrice de transition $\mathbf{P}_0\left(\vec{v}, \vec{v'}\right)$ décrites à la section

1.3 prennent comme valeur

$$\begin{aligned}\mathbf{P}_0\left(\vec{v},\vec{v'}\right) &= P_0\left(\vec{x}(1)=\vec{v'}\middle|\vec{x}(0)=\vec{v}\right)\\ &= P_0\left(\forall j\in\{1,\ldots,n\},\sum_{k=2N(v_0(0)+\ldots+v_{j-1}(0))+1}^{2N(v_0(0)+\ldots+v_j(0))} Z_k(0)=v'_j\right),\end{aligned} \tag{B.1}$$

où $v_0(0)=0$.

Les modèles neutres de Wright-Fisher et de Moran s'inscrivent dans ce modèle. En effet, il suffit de définir la distribution conjointe des variables aléatoires $Z_1(0),\ldots,Z_{2N}(0)$ de façon appropriée pour se retrouver avec l'un ou l'autre des modèles.

Dans le modèle neutre de Wright-Fisher, les possibilités de configurations pour les nombres de rejetons des gènes de la génération initiale sont tous les vecteurs de $2N$ composantes entières non négatives dont la somme est de $2N$. La distribution conjointe des variables aléatoires $Z_1(0),\ldots,Z_{2N}(0)$ donnée par

$$P_0\left(\vec{Z}(0)=\vec{v}\right)=\begin{cases}\binom{2N}{v_1,\ldots,v_{2N}}\left(\frac{1}{2N}\right)^{2N} & \text{si } \sum_{j=1}^{2N}v_j=2N \text{ et } \forall j, v_j\geqslant 0,\\ 0 & \text{sinon,}\end{cases} \tag{B.2}$$

correspond au modèle neutre de Wright-Fisher. La symétrie du coefficient multinomial

$$\binom{2N}{v_1,\ldots,v_{2N}}=\frac{(2N)!}{v_1!\cdots v_{2N}!} \tag{B.3}$$

montre alors l'échangeabilité des variables aléatoires $Z_1(0),\ldots,Z_{2N}(0)$.

Dans le cas du modèle de Moran, les configurations possibles des nombres de rejetons de la première génération sont très restreintes. Si l'événement de naissance-mort concerne un seul gène, les nombres de rejetons auront la configuration

$$(\overbrace{1,\ldots,1}^{2N}). \tag{B.4}$$

Sinon, ils prendront comme configuration une permutation du vecteur

$$(0,\overbrace{1,\ldots,1}^{2N-2},2). \tag{B.5}$$

où le "2" est associé au gène qui s'est reproduit et le "0" au gène qui est mort. En définissant

la distribution conjointe des variables aléatoires $Z_1(0), \ldots, Z_{2N}(0)$ par

$$P_0\left(\vec{Z}(0) = \vec{v}\right) = \begin{cases} \frac{1}{2N} & \text{si } \vec{v} = (\overbrace{1, \ldots, 1}^{2N}), \\ \left(\frac{1}{2N}\right)^2 & \text{si } \vec{v} \text{ est une permutation de } (0, \overbrace{1, \ldots, 1}^{2N-2}, 2), \\ 0 & \text{sinon,} \end{cases} \tag{B.6}$$

on obtient le modèle neutre de Moran. Remarquons encore une fois que cette définition est symétrique et donc que les variables aléatoires sont échangeables.

Revenons maintenant au modèle neutre de Cannings (qui inclut les deux autres) et montrons l'échangeabilité, pour tout $t \geqslant 0$, des variables aléatoires $Z_1(t), \ldots, Z_{2N}(t)$. Pour une bijection γ de $\{1, 2, \ldots, 2N\}$ sur lui-même, on définit la permutation vectorielle associée

$$\begin{aligned} \Gamma : \mathbf{Z}^{2N} &\longrightarrow \mathbf{Z}^{2N} \\ (u_1, u_2, \ldots, u_{2N}) &\longmapsto \left(u_{\gamma(1)}, u_{\gamma(2)}, \ldots, u_{\gamma(2N)}\right). \end{aligned} \tag{B.7}$$

Notons qu'à chaque bijection est associée une unique permutation vectorielle et qu'à chaque permutation vectorielle est associée une unique bijection. Supposons que les variables aléatoires $Z_1(t-1), \ldots, Z_{2N}(t-1)$ soient échangeables. Ceci signifie que pour toute bijection γ et pour tout vecteur $\vec{u}$ admissible, on a

$$P_0\left(\vec{Z}(t-1) = \vec{u}\right) = P_0\left(\vec{Z}(t-1) = \Gamma\left(\vec{u}\right)\right). \tag{B.8}$$

Pour un vecteur $\vec{v}$ admissible, on a que

$$P_0\left(\vec{Z}(t) = \vec{v}\right) = \sum_{\vec{u} \in U} P_0\left(\vec{Z}(t-1) = \vec{u}\right) P_0\left(\vec{Z}(t) = \vec{v} \middle| \vec{Z}(t-1) = \vec{u}\right), \tag{B.9}$$

où U est l'ensemble des vecteurs $\vec{u}$ de $2N$ composantes non négatives dont la somme est de $2N$. Puisque les nombres de rejetons des gènes de la génération t sont échangeables, on a

$$P_0\left(\vec{Z}(t) = \vec{v}\right) = \sum_{\vec{u} \in U} P_0\left(\vec{Z}(t-1) = \vec{u}\right) P_0\left(\vec{Z}(t) = \Gamma\left(\vec{v}\right) \middle| \vec{Z}(t-1) = \Gamma\left(\vec{u}\right)\right). \tag{B.10}$$

Ajoutons à cela l'hypothèse d'induction B.8 pour obtenir

$$P_0\left(\vec{Z}(t) = \vec{v}\right) = \sum_{\vec{u} \in U} P_0\left(\vec{Z}(t-1) = \Gamma\left(\vec{u}\right)\right) P_0\left(\vec{Z}(t) = \Gamma\left(\vec{v}\right) \middle| \vec{Z}(t-1) = \Gamma\left(\vec{u}\right)\right). \tag{B.11}$$

Comme la bijection γ de $\{1, \ldots, 2N\}$ sur lui-même induit une bijection Γ de U sur lui-même, pour tout vecteur $\vec{u'} \in U$, il existe un unique vecteur $\vec{u} \in U$ tel que $\Gamma\left(\vec{u}\right) = \vec{u'}$, et donc la

somme ci-dessus n'est qu'une réorganisation de la somme

$$P_0\left(\vec{Z}(t)=\vec{v}\right)=\sum_{\vec{u}\in U}P_0\left(\vec{Z}(t-1)=\vec{u}\right)P_0\left(\vec{Z}(t)=\Gamma\left(\vec{v}\right)\middle|\vec{Z}(t-1)=\vec{u}\right), \tag{B.12}$$

qu'on reconnaît comme équivalente à

$$P_0\left(\vec{Z}(t)=\vec{v}\right)=P_0\left(\vec{Z}(t)=\Gamma\left(\vec{v}\right)\right). \tag{B.13}$$

Par induction, on a donc que les variables aléatoires $Z_1(t),\ldots,Z_{2N}(t)$ sont échangeables pour chaque génération $t\geqslant 0$ dans le modèle neutre de Cannings.

Cette propriété permettra de prouver la proposition initiale. Supposons qu'on choisisse au hasard un échantillon de b gènes à la génération t et que cet échantillon ait a ancêtres a la génération 0. Considérons un groupe $G\subseteq\{1,2,\ldots,2N\}$ de a gènes de la génération 0. On s'intéresse à la probabilité que ce groupe G soit le groupe d'ancêtres de l'échantillon choisi, sachant que celui-ci à bien a ancêtres à la génération 0. On notera cette probabilité conditionnelle ρ. On voudrait montrer que tout groupe de cette taille a chance égale d'être le dit groupe d'ancêtres. Comme les groupes de ce type sont au nombre de $\binom{2N}{a}$, il faut montrer que

$$\rho=\frac{1}{\binom{2N}{a}}. \tag{B.14}$$

Or, en conditionnant sur le nombre de descendants à la génération t de chaque gène présent a la génération 0, on a que

$$\rho=\frac{\displaystyle\sum_{\substack{\vec{v}\\ \forall j,v_j\geqslant 0\\ v_1+\ldots+v_{2N}=2N}}P_0\left(\vec{Z}(t-1)=\vec{v}\right)\sum_{\substack{\vec{w}\\ \forall j\in G,1\leqslant w_j\leqslant v_j\\ \forall j\in G^C,w_j=0\\ w_1+\ldots+w_{2N}=b}}\frac{\prod_{k=1}^{2N}\binom{v_k}{w_k}}{\binom{2N}{b}}}{\displaystyle\sum_{\substack{H\subseteq\{1,\ldots,2N\}\\ |H|=a}}\sum_{\substack{\vec{v}\\ \forall j,v_j\geqslant 0\\ v_1+\ldots+v_{2N}=2N}}P_0\left(\vec{Z}(t-1)=\vec{v}\right)\sum_{\substack{\vec{w}\\ \forall j\in H,1\leqslant w_j\leqslant v_j\\ \forall j\in H^C,w_j=0\\ w_1+\ldots+w_{2N}=b}}\frac{\prod_{k=1}^{2N}\binom{v_k}{w_k}}{\binom{2N}{b}}}. \tag{B.15}$$

Bien sûr, les vecteurs indices $\vec{v}$ et $\vec{w}$ ont $2N$ composantes non négatives. D'autre part, précisons que G^C et H^C sont les compléments de ces ensembles dans $\{1,\ldots,2N\}$. Au numérateur se trouve la probabilité que le groupe G soit le groupe d'ancêtres de l'échantillon pigé en t, tandis qu'au dénominateur est représentée la probabilité que l'échantillon ait a ancêtres à la génération 0. Il y a $\binom{2N}{a}$ possibilités pour H qu'on peut numéroter $H_1,\ldots,H_{\binom{2N}{a}}$, et ces ensembles peuvent être obtenus par des bijections $\gamma_1,\ldots,\gamma_{\binom{2N}{a}}$ de $\{1,\ldots,2N\}$ sur lui-même (ces bijections ne sont pas uniques) appliquées à G élément par élément. Sous cette transformation, la dernière égalité est

équivalente à

$$\rho = \frac{\sum\limits_{\substack{\vec{v} \\ \forall j, v_j \geqslant 0 \\ v_1+\ldots+v_{2N}=2N}} P_0\left(\vec{Z}(t-1)=\vec{v}\right) \sum\limits_{\substack{\vec{w} \\ \forall j \in G, 1 \leqslant w_j \leqslant v_j \\ \forall j \in G^C, w_j=0 \\ w_1+\ldots+w_{2N}=b}} \prod_{k=1}^{2N} \binom{v_k}{w_k}}{\sum\limits_{i=1}^{\binom{2N}{a}} \sum\limits_{\substack{\vec{v} \\ \forall j, v_j \geqslant 0 \\ v_1+\ldots+v_{2N}=2N}} P_0\left(\vec{Z}(t-1)=\vec{v}\right) \sum\limits_{\substack{\vec{w} \\ \forall j \in \gamma_i(G), 1 \leqslant w_j \leqslant v_j \\ \forall j \in \gamma_i(G^C), w_j=0 \\ w_1+\ldots+w_{2N}=b}} \prod_{k=1}^{2N} \binom{v_k}{w_k}}. \tag{B.16}$$

Une bijection appliquée sur G à l'indice de la somme a pour effet de permuter les conditions sur les éléments de $\vec{w}$. De façon équivalente, on pourrait plutôt permuter les éléments de $\vec{v}$ et laisser ces conditions intactes, ce qui nous donnerait

$$\rho = \frac{\sum\limits_{\substack{\vec{v} \\ \forall j, v_j \geqslant 0 \\ v_1+\ldots+v_{2N}=2N}} P_0\left(\vec{Z}(t-1)=\vec{v}\right) \sum\limits_{\substack{\vec{w} \\ \forall j \in G, 1 \leqslant w_j \leqslant v_j \\ \forall j \in G^C, w_j=0 \\ w_1+\ldots+w_{2N}=b}} \prod_{k=1}^{2N} \binom{v_k}{w_k}}{\sum\limits_{i=1}^{\binom{2N}{a}} \sum\limits_{\substack{\vec{v} \\ \forall j, v_j \geqslant 0 \\ v_1+\ldots+v_{2N}=2N}} P_0\left(\vec{Z}(t-1)=\vec{v}\right) \sum\limits_{\substack{\vec{w} \\ \forall j \in G, 1 \leqslant w_j \leqslant v_{\gamma_i(j)} \\ \forall j \in G^C, w_j=0 \\ w_1+\ldots+w_{2N}=b}} \prod_{k=1}^{2N} \binom{v_{\gamma_i(k)}}{w_k}}. \tag{B.17}$$

En profitant du fait que les variables aléatoires $Z_1(t-1), \ldots, Z_{2N}(t-1)$ sont échangeables, ceci donne rapidement

$$\rho = \frac{\sum\limits_{\substack{\vec{v} \\ \forall j, v_j \geqslant 0 \\ v_1+\ldots+v_{2N}=2N}} P_0\left(\vec{Z}(t-1)=\vec{v}\right) \sum\limits_{\substack{\vec{w} \\ \forall j \in G, 1 \leqslant w_j \leqslant v_j \\ \forall j \in G^C, w_j=0 \\ w_1+\ldots+w_{2N}=b}} \prod_{k=1}^{2N} \binom{v_k}{w_k}}{\sum\limits_{i=1}^{\binom{2N}{a}} \sum\limits_{\substack{\vec{v} \\ \forall j, v_j \geqslant 0 \\ v_1+\ldots+v_{2N}=2N}} P_0\left(\vec{Z}(t-1)=\Gamma_i\left(\vec{v}\right)\right) \sum\limits_{\substack{\vec{w} \\ \forall j \in G, 1 \leqslant w_j \leqslant v_{\gamma_i(j)} \\ \forall j \in G^C, w_j=0 \\ w_1+\ldots+w_{2N}=b}} \prod_{k=1}^{2N} \binom{v_{\gamma_i(k)}}{w_k}}, \tag{B.18}$$

où Γ_i est la permutation vectorielle associée à γ_i, i.e.

$$\begin{aligned} \Gamma_i &: \mathbf{Z}^{2N} \longrightarrow \mathbf{Z}^{2N} \\ &(v_1, v_2, \ldots, v_{2N}) \longmapsto \left(v_{\gamma_i(1)}, v_{\gamma_i(2)}, \ldots, v_{\gamma_i(2N)}\right). \end{aligned} \tag{B.19}$$

Comme pour tout i, la permutation Γ_i met en bijection l'ensemble des vecteurs $\vec{v}$ sur lequel la

somme

$$\sum_{\substack{\vec{v} \\ \forall j, v_j \geq 0 \\ v_1 + \ldots + v_{2N} = 2N}} P_0\left(\vec{Z}(t-1) = \Gamma_i\left(\vec{v}\right)\right) \sum_{\substack{\vec{w} \\ \forall j \in G, 1 \leq w_j \leq v_{\gamma_i(j)} \\ \forall j \in G^C, w_j = 0 \\ w_1 + \ldots + w_{2N} = b}} \prod_{k=1}^{2N} \binom{v_{\gamma_i(k)}}{w_k} \tag{B.20}$$

est effectuée, celle-ci n'est en fait que la somme du numérateur de (B.18) effectuée dans un ordre différent. Par commutativité de la somme, on peut donc conclure que

$$\rho = \frac{1}{\binom{2N}{a}}, \tag{B.21}$$

ce qui confirme que tout groupe de a gènes de la génération 0 a chance égale d'être le groupe d'ancêtres d'un échantillon de b gènes choisi au hasard à la génération t sachant que celui-ci a bien a ancêtres.

Bibliographie

[1] Durret, R., *Essentials of Stochastic Processes*, Springer-Verlag, New York, 1999

[2] Ewens, W. J., *Mathematical Population Genetics : I. Theoretical Introduction*, Springer-Verlag, New York, 2004

[3] Grimaldi, R. P., *Discrete and Combinatorial Mathematics : An Applied Introduction*, Pearson Addison Wesley, Boston, 2004

[4] Karlin, S. and Lessard, S., *Theoretical Studies on Sex Ratio Evolution*, Princeton University Press, Princeton, 1986

[5] Lessard, S., et Ladret, V., The probability of fixation of a single mutant in an exchangeable selection model, *Journal of Mathematical Biology* **54**, pp. 721-744, 2007

[6] Lessard, S., Long-term stability from fixation probabilities in finite populations : New perspectives for ESS theory, *Theoretical Population Biology* **68**, pp. 19-27, 2005

[7] Lessard, S., *Évolution du rapport numérique des sexes et modèles dynamiques connexes*, dans : *Mathematical and Statistical Developments of Evolutionary Theory, NATO ASI Series C : Mathematical and Physical Sciences*, vol. **299** (S. Lessard, Éd.), Kluwer Academic Publishers, Dordrecht, Pays-Bas, pp. 269-325, 1987

[8] Resnick, S. I., *A Probability Path*, Birkhäuser, Boston, 1999