



Perturbation expansions of multilocus fixation probabilities for frequency-dependent selection with applications to the Hill–Robertson effect and to the joint evolution of helping and punishment

Laurent Lehmann^{a,*}, François Rousset^b

^a Department of Biological Sciences, Stanford University, USA

^b University of Montpellier II, CNRS, Institut des Sciences de l'Evolution, France

ARTICLE INFO

Article history:

Received 16 December 2008

Available online 2 May 2009

Keywords:

Multilocus models

Fixation probabilities

Coalescence times

Frequency-dependent selection

Hill–Robertson effect

Coevolution of helping and punishment

ABSTRACT

Natural populations are of finite size and organisms carry multilocus genotypes. There are, nevertheless, few results on multilocus models when both random genetic drift and natural selection affect the evolutionary dynamics. In this paper we describe a formalism to calculate systematic perturbation expansions of moments of allelic states around neutrality in populations of constant size. This allows us to evaluate multilocus fixation probabilities (long-term limits of the moments) under arbitrary strength of selection and gene action. We show that such fixation probabilities can be expressed in terms of selection coefficients weighted by mean first passages times of ancestral gene lineages within a single ancestor. These passage times extend the coalescence times that weight selection coefficients in one-locus perturbation formulas for fixation probabilities. We then apply these results to investigate the Hill–Robertson effect and the coevolution of helping and punishment. Finally, we discuss limitations and strengths of the perturbation approach. In particular, it provides accurate approximations for fixation probabilities for weak selection regimes only ($Ns \leq 1$), but it provides generally good prediction for the direction of selection under frequency-dependent selection.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

In the absence of mutation, any finite population will ultimately become genetically homogeneous as a result of random genetic drift affecting the evolutionary dynamics. The fate of any allele is then either its fixation or loss from the population. For this reason, the fixation probability of a mutant allele is a standard measure used to evaluate its evolutionary success (Crow and Kimura, 1970; Hartl and Clark, 2007; Ewens, 2004; Gillespie, 2004). The fixation probability captures the effects of both genetic drift and natural selection on gene frequency change, from the appearance to the eventual fixation or loss of an allele from the population. But the interaction between genetic drift and selection is complicated and, besides some special cases, population biologists have renounced evaluating fixation probabilities exactly. We are thus left with the necessity of approximating under even the simplest biological scenarios.

Under one-locus genetics, the fixation probability of a mutant allele is generally approximated by three different approaches.

First, the population is assumed to be so large that the interactions between mutants is neglected. This leads to the branching process approximation of fixation probabilities (e.g., Haldane, 1927; Bartlett, 1955; Ewens, 2004). Second, the population is assumed to be large and selection weak, which leads to the diffusion approximation of fixation probabilities (e.g., Crow and Kimura, 1970; Gillespie, 2004; Ewens, 2004). Finally, regardless of population size, selection is assumed to be weak and exact perturbation results are obtained for fixation probabilities near neutrality. This leads to the evaluation of first-order perturbation of fixation probabilities in the strength of selection, which allows one to compute analytically measures of convergence stability for continuous traits under a large spectrum of demographic scenarios in spatially structured populations (e.g., Rousset, 2004; Lessard and Ladret, 2007a; Lehmann, 2008).

For multilocus genetics, there are few results on fixation probabilities when both selection and drift simultaneously affect the evolutionary dynamics, and most results rely on the branching process approximation (e.g., Barton, 1995; Otto and Barton, 1997). Yet how these fixation probabilities depend on population size appears important for Hill–Robertson effects, the evolution of recombination, or the frequency-dependent selection occurring in social evolution (e.g., Hill and Robertson, 1966; Otto and Barton, 2001; Wild and Taylor, 2004; Grafen, 2007).

* Corresponding author.

E-mail addresses: lehmann@stanford.edu (L. Lehmann), rousset@isem.univ-montp2.fr (F. Rousset).

In this paper we extend the first-order one-locus perturbation method to an n th-order multilocus perturbation approach. We describe a formalism to calculate systematic perturbation expansions of moments of allelic states (expectations of sets of genes sampled within or between individuals) under arbitrary gene action, which covers epistatic interactions within genomes as well as synergistic interactions between them (thus including social behaviors). This allows us to evaluate n th-order Taylor polynomials in the strength of selection of multilocus fixation probabilities (long-term limits of the moments). We then extend algorithms for deterministic models in order to obtain analytical approximations of fixation probabilities under a Wright–Fisher model of reproduction. We illustrate the method and the approximations by evaluating fixation probabilities of alleles and gametes under two two-locus two-allele settings, one involving the Hill–Robertson effect and the other the coevolution of helping and punishment.

2. Model

2.1. Notation for L -locus fixation probabilities

The analysis presented in this paper is restricted to a haploid population of constant size N with no further division into demographic classes (e.g. no age structure, separate sexes, or geographic structure). We assume that each individual bears L loci and that a mutant and a resident allele segregate at each locus (i.e., L -locus two-allele model, Slatkin, 1971; Christiansen, 1999), which gives 2^L different gamete types in the population.

Our notation (Table 1) draws on the general notation for multilocus models in infinite populations of Kirkpatrick et al. (2002) and Roze and Rousset (2008). We denote by $p_{l(i)}$ the frequency (0 or 1) of the mutant allele at locus l in individual i of the population. Products of such random variables are noted $p_{V(i)} = \prod_{l \in V} p_{l(i)}$, where V is a set of distinct loci sampled from the same individual in the population (either the L distinct loci or a subset of it). The average of $p_{V(i)}$ over all individuals in the population is denoted

$$p_V = \frac{1}{N} \sum_i p_{V(i)}, \quad (1)$$

which is itself a random variable. In the abovementioned works, indicator variables are defined as $\zeta_{l(i)} = p_{l(i)} - \wp_{l(i)}$ relative to some reference values $\wp_{l(i)}$ usually (but not necessarily) taken as the average allele frequency at locus l in the population. This choice of reference values reduces the number of variables that need to be considered in infinite population models, but as no such simplification appears in the finite population model we do not use such variables here.

Sets of loci can also be sampled from distinct individuals; for instance from two individuals, in which case the average of the product of mutant alleles $p_{V(i)}p_{W(j)}$ sampled from the same or distinct loci in two distinct individuals is denoted

$$p_{V/W} = \frac{1}{N(N-1)} \sum_i \sum_{i \neq j} p_{V(i)}p_{W(j)}, \quad (2)$$

where the “/” symbol separates sets of loci sampled from distinct individuals. We will refer to sets of distinct loci sampled from the same individual and sets of similar or distinct loci sampled from different individuals as sets of position. A position refers to a particular locus in a particular context (Kirkpatrick et al., 2002, p. 1729); for instance, the place in the genome where the maternally inherited copy of a given gene resides, or the place in the genome of a neighbor of a given individual where a particular gene affecting a social behavior resides.

More generally sets of positions can be sampled from an arbitrary number of distinct individuals in the population. We thus introduce the random variable p_S , where $S \equiv S_1/S_2/\dots/S_{|S|}$, and where each S_i is a set of loci sampled from the same individual

Table 1
List of symbols.

Symbol	Definition
L	Number of loci.
N	Population size.
r	Recombination rate.
w_i	Fitness of individual i .
f_i	Fertility of individual i .
$p_{l(i)}$	Frequency (0 or 1) of the mutant allele at locus l in individual i .
$p_{V(i)}$	Frequency (0 or 1) that all loci in the set V in individual i carry the mutant allele.
$p_{V/W}/\dots$	Average of $p_{V(i)}p_{W(j)}\dots$ variables over all individuals in the population sampled without replacement.
$p_{\widehat{V}/\widehat{W}/\dots}$	Average of $p_{V(i)}p_{W(j)}\dots$ variables over all individuals in the population sampled with replacement.
Ep_S	Expectation of p_S .
\mathbf{p}	Vector of the p_S variables (vector of random variables).
$\mathbf{E}\mathbf{p}$	Vector of the expectations of the p_V variables (vector of Ep_S variables).
\wp	Vector of expectations of reference variables.
\mathbf{v}	Vector of deviations ($\mathbf{v} = \mathbf{E}\mathbf{p} - \wp$).
\mathbf{d}	Vector of the sum of deviations ($\mathbf{d} \equiv \sum_{t=0}^{\infty} \mathbf{v}(t)$).
\mathbf{a}_U	Row of \mathbf{A} giving the coefficients describing the dynamics of p_U .
\mathbf{u}	Left eigenvector of \mathbf{A}° associated to a unit eigenvalue.
\mathbf{e}	Right eigenvector of \mathbf{A} associated to a unit eigenvalue.
\mathbf{A}	Transition matrix of the Ep_S variables.
\mathbf{A}_\wp	Matrix obtained by row replacement of the rows of \mathbf{A} .
\mathbf{A}_c	Matrix with each row consisting of the left eigenvector of \mathbf{A}° associated to its unit eigenvalue.
\mathbf{Z}	Fundamental matrix associated to \mathbf{A}° .

and $|S|$ is the total number of different individuals from which sets of loci have been sampled. The variable p_S is obtained in a direct generalization of Eq. (2), by averaging (without replacement) over all $|S|$ distinct individuals. Conditional on some initial state of the population, the expectation of $p_S(t)$ at time t is given by

$$E[p_S(t)] = \sum_{\mathbf{p}(t)} p_S(t) \Pr(\mathbf{p}(t)), \quad (3)$$

where $\mathbf{p}(t)$ is a vector collecting the frequencies of the 2^L gamete types in the population at time t ; $\Pr(\mathbf{p}(t))$ is the distribution of $\mathbf{p}(t)$ at time t conditional on the initial state of the population being $\mathbf{p}(0)$; and the sum is over all possible values of $\mathbf{p}(t)$. We will use the expectation operator notation without brackets as a shorthand for expectation conditional on the initial state $\mathbf{p}(0)$ of the population, e.g., $Ep_S(t) \equiv E[p_S(t)]$.

For a single locus denoted A , Ep_A represents the probability that an individual sampled at random from the population carries the mutant allele at that locus, and $Ep_{A/A}$ is the probability that two distinct individuals sampled at random carry the mutant allele. For two loci, denoted A and B , Ep_{AB} is the probability that a single individual carries the mutant alleles at both loci while $Ep_{AB/A/B}$ is the probability that, among three randomly sampled individuals, one carries the mutant alleles at both loci, another individual carries the mutant allele at locus A and the third individual carries the mutant allele at locus B . Note that one can also interpret the subscripts in the Ep_S variables as being “alleles” or “gametes” (e.g., Ep_A and Ep_{AB} , are respectively, the frequency of allele A and gamete AB if the mutant alleles are denoted by the same letter as the locus at which they segregate). The Ep_S variables can thus be interpreted as moments of allelic states in exactly the same way as is usually done in population genetics (Wright, 1931; Crow and Kimura, 1970; Gillespie, 2004), and we will refer to them as such.

We will generically denote by $p_U(t)$ the frequency of a focal gametic combination U at time t in the population, where U is the set of mutant alleles at L distinct loci that can be sampled from the same individual. In the absence of mutation, the focal gametic combination goes either to fixation ($p_U(\infty) = 1$) or is lost from the population ($p_U(\infty) = 0$). The fixation probability π_U of this gamete is thus its expected frequency in the long-term:

$\pi_U = Ep_U(\infty)$. Our main aim in this paper is to obtain the n th-order Taylor polynomials of π_U with respect to an overall measure δ of the strength of selection, which may conveniently be defined as the largest (in absolute value) of a set of selection coefficients. The n th-order Taylor polynomial of the fixation probability π_U of the focal gamete (or of any other quantity z) will be written as

$$\pi_U^{[n]} = \sum_{k=0}^n \frac{\delta^k}{k!} \pi_U^{(k)}, \quad (4)$$

where $\pi_U^{(k)}$ is the k th derivative of π_U with respect to δ evaluated at $\delta = 0$ (n th-order perturbation). Generically the superscripts $[n]$ and (n) thus denote, respectively, the n th-order Taylor polynomial and derivative of any quantity. But a special role in our formalization will be played by zero-order Taylor polynomials, which refer to quantities evaluated in the neutral process (no selection, $\delta = 0$). For ease of presentation, we will use the superscript \circ to identify any quantity evaluated under neutrality (e.g., $\pi_U^{[0]} = \pi_U^{(0)} = \pi_U^\circ$). Further, in all explicit examples presented below, we will use only first- and second-order derivatives and Taylor polynomials. In order to simplify the presentation of these examples, we will use dot and double dot accents (\dot{z} , \ddot{z}) to denote first and second derivatives of any quantity z with respect to δ , and dot and double dot exponents to denote (z' , z''), first-, and second-order Taylor polynomials (e.g., $z' = z^\circ + \delta \dot{z}$, $z'' = z^\circ + \delta \dot{z} + \delta^2 \ddot{z}/2$).

We will evaluate $\pi_U^{[n]}$ by evaluating the perturbations $\pi_U^{(n)}$, which will be carried out by decomposing π_U into a sum of expected gamete frequency changes over one generation (Rousset, 2003; Lessard and Ladret, 2007b), and then compute the perturbations of expected gamete frequency change. To this end, we denote the change in expected gamete frequency from generation t to $t+1$ by $\Delta Ep_U(t) \equiv Ep_U(t+1) - Ep_U(t)$. With this, the fixation probability can be written as

$$\begin{aligned} \pi_U &= Ep_U(\infty) \\ &= p_U(0) + \sum_{t=0}^{\infty} \Delta Ep_U(t). \end{aligned} \quad (5)$$

The change in expected gamete frequency can itself be expanded in terms of conditional changes as

$$\Delta Ep_U(t) = \sum_{\mathbf{p}(t)} \Pr(\mathbf{p}(t)) E[\Delta p_U | \mathbf{p}(t)], \quad (6)$$

where $E[\Delta p_U | \mathbf{p}(t)]$ is a shorthand for the expected conditional change in allele frequency $[E[p_U(t+1) | \mathbf{p}(t)] - p_U(t)]$. Our goal is to express $E[\Delta p_U | \mathbf{p}(t)]$ as a linear function of allelic states, and then integrate such expressions over all sample paths. For ease of illustration, we first present the calculations under neutrality and first-order effects of selection, then illustrate these calculations with simple examples, and finally generalize to higher-order perturbations.

2.2. Neutrality

Under neutrality, Eq. (6) becomes

$$\Delta Ep_U^\circ(t) = \sum_{\mathbf{p}(t)} \Pr^\circ(\mathbf{p}(t)) E^\circ[\Delta p_U | \mathbf{p}(t)], \quad (7)$$

where both the expected change in gamete frequency and the distribution of gamete frequencies are evaluated in the neutral process ($\delta = 0$). As in infinite populations, $E^\circ[\Delta p_U | \mathbf{p}(t)]$ can be expressed as $\sum_{X \in \mathcal{X}} a_X^\circ p_X(t)$ for some set of positions \mathcal{X} , where the a_X° 's are transmission coefficients (Kirkpatrick et al., 2002) describing the effect of recombination, drift, or both (no selection) on moments of gamete frequency change, and the set X may

involve sets of gene copies sampled from the same and from different individuals.

In order to illustrate these notions, we now introduce a two-locus model with a focal gamete consisting of the mutant alleles at loci A and B. Because recombination entails that alleles in a gamete may have distinct parents of origin, we also have to consider the dynamics of $Ep_{A/B}^\circ$ in order to account for that of Ep_{AB}° (Kimura, 1963; Ewens, 2004); where $Ep_{A/B}^\circ$ is the probability that, among two different individuals, one carries the mutant allele at locus A and the other carries the mutant allele at locus B. In order to calculate these variables explicitly, we assume a Wright–Fisher random union of gamete model (Karlin, 1968; Ewens, 2004, p. 130). Namely, N haploid parents produce a large number of gametes that fuse randomly to form diploid zygotes (self-fertilization thus occurs with probability $1/N$), which is immediately followed by meiosis with recombination rate r to produce haploid juveniles. Finally, N adults are randomly sampled among the juveniles to form the next generation.

Under this life-cycle, the probability that two genes, sampled from two distinct loci from the same individual, descend from the same individual in the previous generation is $(1-r) + r/N$, where the second term accounts for self-fertilization. In addition, we need the probability that two genes sampled from distinct loci from two distinct individuals, descend from the same individual in the previous generation, which is given by $1/N$, whereby

$$\begin{aligned} E^\circ[p_{AB}(t+1) | \mathbf{p}(t)] &= \left(1 - r \left(\frac{N-1}{N}\right)\right) p_{AB}(t) \\ &\quad + r \left(\frac{N-1}{N}\right) p_{A/B}(t) \\ E^\circ[p_{A/B}(t+1) | \mathbf{p}(t)] &= \frac{1}{N} p_{AB}(t) + \left(\frac{N-1}{N}\right) p_{A/B}(t). \end{aligned} \quad (8)$$

By subtracting $p_{AB}(t)$ on both sides of the first equation and $p_{A/B}(t)$ from the second, we obtain

$$\begin{aligned} E^\circ[\Delta p_{AB} | \mathbf{p}(t)] &= -r \left(\frac{N-1}{N}\right) (p_{AB}(t) - p_{A/B}(t)) \\ E^\circ[\Delta p_{A/B} | \mathbf{p}(t)] &= \frac{1}{N} (p_{AB}(t) - p_{A/B}(t)), \end{aligned} \quad (9)$$

which shows that the transmission coefficients determining $E^\circ[\Delta p_{AB} | \mathbf{p}(t)]$ are $-r(N-1)/N$ and $r(N-1)/N$ and those determining $E^\circ[\Delta p_{A/B} | \mathbf{p}(t)]$ are $1/N$ and $-1/N$.

More generally, we have

$$\begin{aligned} \Delta Ep_U^\circ(t) &= \sum_{\mathbf{p}(t)} \Pr^\circ(\mathbf{p}(t)) \sum_{X \in \mathcal{X}} a_X^\circ p_X(t) \\ &= \sum_{X \in \mathcal{X}} a_X^\circ \sum_{\mathbf{p}(t)} \Pr^\circ(\mathbf{p}(t)) p_X(t), \end{aligned} \quad (10)$$

where the last sum is the expectation of $p_X(t)$ in the neutral process, denoted $Ep_X^\circ(t)$, so that

$$\Delta Ep_U^\circ(t) = \sum_{X \in \mathcal{X}} a_X^\circ Ep_X^\circ(t), \quad (11)$$

and when U stands for a single gene copy, all a_X° are null, whenever there are no mutations.

For the random union of gametes model described by Eq. (9), application of Eq. (11) gives

$$\begin{aligned} \Delta Ep_{AB}^\circ(t) &= -r \left(\frac{N-1}{N}\right) (Ep_{AB}^\circ(t) - Ep_{A/B}^\circ(t)) \\ \Delta Ep_{A/B}^\circ(t) &= \frac{1}{N} (Ep_{AB}^\circ(t) - Ep_{A/B}^\circ(t)). \end{aligned} \quad (12)$$

By subtracting the second equation from the first one, the measure of expected linkage-disequilibrium (LD) $D_{AB}^{\circ}(t) \equiv Ep_{AB}^{\circ}(t) - Ep_{A/B}^{\circ}(t)$ can be found to satisfy the relation $\Delta D_{AB}^{\circ}(t) = -(r + (1-r)/N)D_{AB}^{\circ}(t)$, which yields $D_{AB}^{\circ}(t) = [(1-r)(N-1)/N]^t D_{AB}^{\circ}(0)$, where $D_{AB}^{\circ}(0) = p_{AB}(0) - p_{A/B}(0)$. By substituting this result into Eq. (12), itself inserted into Eq. (5), we find that the fixation probability of the focal gamete is

$$\begin{aligned} \pi_{AB}^{\circ} &= p_{AB}(0) - \sum_{t=0}^{\infty} r \left(\frac{N-1}{N} \right)^t D_{AB}^{\circ}(t) \\ &= p_{AB}(0) - \frac{r(N-1)D_{AB}^{\circ}(0)}{1+r(N-1)}. \end{aligned} \quad (13)$$

This result has been derived previously with different approaches [e.g., Karlin and McGregor, 1968, Table 1; Ewens, 2004, eq. 3.138; we mention that the measure of LD defined in this earlier work was $p_{AB} - p_A p_B = (p_{AB} - p_{A/B})(N-1)/N$ since $p_A p_B = p_{AB}/N + p_{A/B}(N-1)/N$].

2.3. First-order effects

Taking the derivative on both sides of Eq. (6) with respect to δ produces the first-order perturbation

$$\Delta \dot{E}p_U(t) = \sum_{\mathbf{p}(t)} [\dot{\Pr}(\mathbf{p}(t))E^{\circ}[\Delta p_U|\mathbf{p}(t)] + \Pr^{\circ}(\mathbf{p}(t))\dot{E}[\Delta p_U|\mathbf{p}(t)]]. \quad (14)$$

Using the expression $E^{\circ}[\Delta p_U|\mathbf{p}(t)] = \sum_{X \in \mathcal{X}} a_X^{\circ} p_X(t)$ allows us to simplify the first term in Eq. (14) as

$$\begin{aligned} \sum_{\mathbf{p}(t)} \dot{\Pr}(\mathbf{p}(t))E^{\circ}[\Delta p_U|\mathbf{p}(t)] &= \sum_{X \in \mathcal{X}} a_X^{\circ} \sum_{\mathbf{p}(t)} \dot{\Pr}(\mathbf{p}(t))p_X(t) \\ &= \sum_{X \in \mathcal{X}} a_X^{\circ} \dot{E}p_X(t). \end{aligned} \quad (15)$$

Likewise to the neutral case, $\dot{E}[\Delta p_U|\mathbf{p}(t)]$ can be written as $\sum_{Y \in \mathcal{Y}} \dot{a}_Y p_Y(t)$ for some set \mathcal{Y} , where the \dot{a}_Y 's are first-order selection coefficients that do not depend on any random variable. For instance, for a one-locus selection model in a random mating population with a mutant allele at locus A having a selective advantage δs_A over the resident, we have to first-order

$$\dot{E}[\Delta p_A(t)|\mathbf{p}(t)] = s_A (p_A(t) - p_A^2(t)). \quad (16)$$

The right-hand side of this equation is the classical first-order approximation of the haploid one-locus selection model [Wright, 1969, p. 30; Crow and Kimura, 1970; Gillespie, 2004, i.e., first-order Taylor series of $\delta s_A p_A(t)(1-p_A(t))/(1+\delta s_A p_A(t))$ evaluated at $\delta = 0$]. The first-order selection coefficients determining $\dot{E}[\Delta p_A(t)|\mathbf{p}(t)]$ are thus given by s_A and $-s_A$.

More generally, we have

$$\begin{aligned} \sum_{\mathbf{p}(t)} \Pr^{\circ}(\mathbf{p}(t))\dot{E}[\Delta p_U|\mathbf{p}(t)] &= \sum_{Y \in \mathcal{Y}} \dot{a}_Y \sum_{\mathbf{p}(t)} \Pr^{\circ}(\mathbf{p}(t))p_Y(t) \\ &= \sum_{Y \in \mathcal{Y}} \dot{a}_Y Ep_Y^{\circ}(t), \end{aligned} \quad (17)$$

and collecting all terms gives

$$\Delta \dot{E}p_U(t) = \sum_{Y \in \mathcal{Y}} \dot{a}_Y Ep_Y^{\circ}(t) + \sum_{X \in \mathcal{X}} a_X^{\circ} \dot{E}p_X(t). \quad (18)$$

This equation extends the formula for the first-order perturbation of moments of allelic states in a one-locus model (Rousset, 2003; Lessard and Ladret, 2007b) to the multilocus case. For a one-locus model the second term in Eq. (18) is null in the absence of mutation, since all a_X° are null in that case.

For the one-locus selection model described by Eq. (16), application of Eq. (18) gives

$$\Delta \dot{E}p_A(t) = s_A \left(Ep_A^{\circ}(t) - Ep_{A/A}^{\circ}(t) \right), \quad (19)$$

where $Ep_{A/A}^{\circ}(t)$ is the probability that two individuals sampled with replacement carry the mutant allele ($Ep_{A/A}^{\circ}(t) = E^{\circ}[p_A(t)^2]$).

This probability can be expanded as

$$Ep_{A/A}^{\circ}(t) = \frac{1}{N} Ep_A^{\circ}(t) + \left(\frac{N-1}{N} \right) Ep_{A/A}^{\circ}(t), \quad (20)$$

where $Ep_{A/A}^{\circ}$ is the probability that two individuals sampled without replacement carry the mutant, which gives

$$\Delta \dot{E}p_A(t) = s_A \left(\frac{N-1}{N} \right) (Ep_A^{\circ}(t) - Ep_{A/A}^{\circ}(t)). \quad (21)$$

For the Wright–Fisher scheme of reproduction, the variable $D_A^{\circ}(t) \equiv Ep_A^{\circ}(t) - Ep_{A/A}^{\circ}(t)$ satisfies $D_A^{\circ}(t) = [(N-1)/N]^t D_A^{\circ}(0)$ because $Ep_{A/A}^{\circ}(t+1) = (1/N)Ep_A^{\circ}(t) + [(N-1)/N]Ep_{A/A}^{\circ}(t)$, where $1/N$ is the probability that two individuals descend from the same parent. By substituting this result into Eq. (21), itself inserted into Eq. (5), we find that the first-order Taylor polynomial of the fixation probability is given by

$$\begin{aligned} \pi_A^{\circ} &= p_A(0) + \delta \left[\sum_{t=0}^{\infty} s_A \left(\frac{N-1}{N} \right)^t D_A^{\circ}(t) \right] \\ &= p_A(0) + \delta s_A \left(\frac{N-1}{N} \right) N D_A^{\circ}(0). \end{aligned} \quad (22)$$

This result has been derived previously, and when there is initially a single mutant allele in the population we have $p_A(0) = D_A^{\circ}(0) = 1/N$ (Hill, 1972, eq. 7; Rousset and Billiard, 2000; Lessard and Ladret, 2007b). As pointed out by Lessard and Ladret, we can compute the derivative of the ultimate expected frequency (fixation probability) as the sum of expected derivatives in each generation if the series of expected derivatives converges uniformly in a neighborhood of $\delta = 0$. They showed this for the one-locus Cannings exchangeable model, and this follows more generally from Eq. (A.10) (see explanations in the Appendix right after this equation).

2.4. nth-order effects

2.4.1. Recursions for moments of allelic states

The n th-order perturbation of the conditional expected change in gamete frequency can be written as

$$E^{(n)}[\Delta p_U|\mathbf{p}(t)] = \sum_{X_n \in \mathcal{X}_n} a_{X_n}^{(n)} p_{X_n}(t) \quad (23)$$

for some set \mathcal{X}_n , where the $a_{X_n}^{(n)}$ are n th-order selection coefficients that do not depend on any random variable. For any $n \geq 0$, these coefficients satisfy

$$\sum_{X_n \in \mathcal{X}_n} a_{X_n}^{(n)} = 0, \quad (24)$$

because, when every individual in the population bears the same genotype (e.g., all $p_{X_n} = 1$), no systematic change in gamete frequency can occur over a selection, recombination or migration phase. This logic applies to the expected change in frequency of any genotype for any L -locus model, and for any submodel involving only a subset of loci of positions. Thus, in a three-locus model, $\sum_{W_n \in \mathcal{W}_n} a_{W_n}^{(n)} = 0$ for every subset \mathcal{W}_n of all sets of positions involving only one locus, or at most two loci. This implies that $\sum_{W \in \mathcal{W}_n} a_{W_n}^{(n)} = 0$ for every subset \mathcal{W}_n of sets of positions each involving exactly two given loci. Thus with 3 loci, 7 non-overlapping sets of coefficients each add up to zero, one set for each locus, one set for each pair of locus, and the set of coefficients involving all three loci.

Differentiating the left- and right-hand side of Eq. (6) n times with respect to δ , using Leibniz's rule for the differentiation of products, inserting Eq. (23), and rearranging allows us to write

$$Ep_U^{(n)}(t+1) = Ep_U^{(n)}(t) + \sum_{j=0}^n \binom{n}{j} \sum_{X \in \mathcal{X}_j} a_X^{(j)} Ep_X^{(n-j)}(t). \quad (25)$$

On substitution of this equation into the expansion $Ep_U^{[n]}(t+1)$ (see Eq. (4)), we obtain after rearrangements a recursion on Taylor polynomials:

$$Ep_U^{[n]}(t+1) = Ep_U^{[n]}(t) + \sum_{j=0}^n \frac{\delta^j}{j!} \sum_{X \in \mathcal{X}_j} a_X^{(j)} Ep_X^{[n-j]}(t). \quad (26)$$

Since a similar recursion can be written for the Taylor polynomials $Ep_X^{[n-j]}(t)$ of any set of allelic states X appearing on the right-hand side of Eq. (26), the recursion for $Ep_U^{[n]}$ can be closed (the same logic also applies to Eq. (25)). More generally, one can evaluate in this way the n th Taylor polynomial $Ep_S^{[n]}$ of any moment of allelic state S in the population, be it the complete genetic description of the population.

2.4.2. Matrix representation

The dynamics of the Taylor polynomials $Ep_U^{[n]}(t)$ of the focal gamete (or more generally of any moment of allelic state) as described by Eq. (26) are linear in the expectations, which suggests that they can be evaluated using matrix algebra. To this end, we gather all the moments affecting the dynamics of $Ep_U^{[n]}(t)$ into a vector, except the lower-order polynomials of the highest-order moment of any set S of positions involved in the recursion (if we have $Ep_S^{[n]}(t)$, we do not consider $Ep_S^{[n-1]}(t)$ nor $Ep_S^{[n-2]}(t)$ in the vector). We then augment this vector with any additional moment required to close the system of recursions. The vector gathering the total set \mathcal{S} of positions closing the system is denoted $\mathbf{Ep}(t) \equiv (Ep_S(t))_{S \in \mathcal{S}}$, and it satisfies the recursion

$$\mathbf{Ep}(t+1) = \mathbf{A}\mathbf{Ep}(t), \quad (27)$$

for some matrix \mathbf{A} collecting the transmission and selection coefficients appearing in Eq. (26). Importantly, Eq. (27) is not a direct matrix formulation of Eq. (26), as it involves only the highest-order polynomial of any set of positions S involved in Eq. (26). This allows us to shrink the size of \mathbf{A} , but with the charge of carrying terms in excess powers in δ , which can subsequently be disposed of by taking a Taylor series (see below).

In order to illustrate these concepts we use the two explicit examples introduced above. For the two-locus neutral random union of gamete model (Eq. (12)), we need the vector $\mathbf{Ep}(t) \equiv (Ep_{AB}^{[n]}(t), Ep_{A/B}^{[n]}(t))$ to follow the dynamics of $Ep_{AB}^{[n]}(t)$. The dynamics of this vector is described by the matrix with neutral coefficients given by

$$\mathbf{A}^\circ = \begin{pmatrix} 1-r \left(\frac{N-1}{N} \right) & r \left(\frac{N-1}{N} \right) \\ \frac{1}{N} & \frac{N-1}{N} \end{pmatrix}. \quad (28)$$

For the one-locus selection model (Eq. (21)), we need the vector $\mathbf{Ep}(t) \equiv (Ep_A^{[n]}(t), Ep_{A/A}^{[n]}(t))$ to track the dynamics of $Ep_A^{[n]}(t)$. The dynamics of this vector can be described by the transition matrix

$$\mathbf{A} = \mathbf{A}^\circ + \delta \dot{\mathbf{A}}, \quad (29)$$

where

$$\mathbf{A}^\circ = \begin{pmatrix} 1 & 0 \\ \frac{1}{N} & \frac{N-1}{N} \end{pmatrix}, \quad (30)$$

gathers all neutral coefficients, while

$$\dot{\mathbf{A}} = \begin{pmatrix} s_A \left(\frac{N-1}{N} \right) & -s_A \left(\frac{N-1}{N} \right) \\ 0 & 0 \end{pmatrix}, \quad (31)$$

gathers all first-order selection coefficients. More generally, we can express \mathbf{A} as a linear combination of matrices $\mathbf{A}^{(n)}$ gathering all the n th-order $a_X^{(n)}$ selection coefficients.

These examples illustrate that \mathbf{A}° is a row stochastic matrix (each element ij is either null or positive, and the rows sum up to one, which entails at least one eigenvalue 1). But more generally, \mathbf{A}° may involve several independent neutral systems, so that \mathbf{A}° is a diagonal block matrix, where each block is the transition matrix of a Markov chain (Grinstead and Snell, 1997; Iosifescu, 2007), one for each set of sets of positions involving exactly a given number of distinct loci. For instance, for a two-locus model we may need to evaluate $\mathbf{Ep}^\circ(t) \equiv (Ep_A^\circ(t), Ep_{A/A}^\circ(t), Ep_{AB}^\circ(t), Ep_{A/B}^\circ(t))$, whose dynamics under Wright–Fisher random union of gametes is described by the two-block diagonal matrix

$$\mathbf{A}^\circ = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{N} & \frac{N-1}{N} & 0 & 0 \\ 0 & 0 & 1-r \left(\frac{N-1}{N} \right) & r \left(\frac{N-1}{N} \right) \\ 0 & 0 & \frac{1}{N} & \frac{N-1}{N} \end{pmatrix} \quad (32)$$

(see also Eq. (A.2) of the Appendix). The element ij of \mathbf{A}° gives the probability that a set of genes sampled in the set of positions i descend from the set of positions j in the previous generation. As can be seen from Eq. (32) (or Eq. (A.2)), all elements ij of \mathbf{A}° are null when i and j involves sets of positions with a different number of loci (one loci is involved in the set of positions $\{A, A/A\}$ while two distinct loci are involved in the set of positions $\{AB, A/B\}$). All properties of \mathbf{A}° can thus be deduced from the properties of each block because each neutral system in a block is a Markov chain independent from those in the other blocks, and we assume, without obvious biological restrictions, that each block is a mixing Markov chain (Iosifescu, 2007, p. 126; a Markov chain is mixing if it eventually converges to a stable stationary distribution). These examples also illustrate that in the presence of selection, the \mathbf{A} matrix no longer has the block structure of \mathbf{A}° , and is no longer a stochastic matrix as some of its elements can be negative (e.g., Eq. (29)), though all its row sums are still equal to 1, so it still has eigenvalue 1.

2.5. Fixation probabilities and mean first passage times

2.5.1. Taylor polynomials of fixation probabilities

The Taylor polynomial $\pi_U^{[n]}$ of the fixation probability is given by $Ep_U^{[n]}(\infty)$; namely, the long-term limit of the moments in the absence of mutation, which can either be directly obtained from the vector $\mathbf{Ep}^{[n]}(\infty)$ (Eq. (27)) or by summing up $\Delta Ep_U^{[n]}(t)$ from the time of the appearance of the mutant gamete until its eventual fixation in, or loss from, the population (Eq. (5), see also Eqs. (13) and (22)). While the first approach seems at first glance more direct, it involves the evaluation of the eigenvectors of \mathbf{A} associated with its unit eigenvalues, and the deletion of the excess powers in δ mentioned above by taking a Taylor series. As will be shown below, the second approach involves only the inversion of a matrix considered at neutrality (derived from \mathbf{A}°). This approach has been used previously in the one-locus model because it allows one to express fixation probabilities in terms of selection coefficients and coalescence times (Rousset, 2003; Lessard and Ladret, 2007b).

