Theoretical Population Biology 101 (2015) 40–46

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

A simple, semi-deterministic approximation to the distribution of selective sweeps in large populations

Guillaume Martin [a], Amaury Lambert [b,c,*]

[a] Institut des Sciences de l'Evolution, UMR 5554 – CNRS – Université Montpellier 2, Place Eugène Bataillon C.C. 065, 34095 Montpellier cedex 05, France
[b] Laboratoire de Probabilités et Modèles Aléatoires CNRS UMR 7599, UPMC Université Paris 06, Paris, France
[c] Center for Interdisciplinary Research in Biology CNRS UMR 7241, Collège de France, Paris, France

ARTICLE INFO

Article history:
Received 30 October 2014
Available online 24 February 2015

Keywords:
Selective sweep
Genetic drift
Selection
Fixation time
Wright–Fisher diffusion
Feller diffusion

ABSTRACT

In large populations, the distribution of the trajectory of allele frequencies under selection and genetic drift approaches a semi-deterministic behavior: a deterministic trajectory started and ended at stochastic boundary values. This provides simple yet accurate approximations for the distribution of allelic frequencies over time (conditional on fixation), and of extinction and fixation times, for both hard and soft sweeps, and under arbitrary inbreeding and dominance.

1. Introduction

The stochastic dynamics of beneficial allele frequencies under selection and genetic drift have been the focus of theoretical population genetics for decades (Kimura and Ohta, 1969; Maruyama, 1974). This process is typically analyzed through the Wright–Fisher diffusion (Kimura and Ohta, 1969), which approximates the exact process, even for relatively small populations. This tool has allowed the derivation of quantities of central evolutionary importance, such as the probability of ultimate fixation of an allele with given fitness effect. However, these results only describe long-term behaviors: rigorously, what is known in simple form is the probability that an allele fixes or is lost after an *infinite* time has elapsed. Even in the approximate diffusion framework, it proves much more challenging to derive shorter-term (somewhat more basic) quantities such as the distribution of allele frequencies at a given time, or the distribution of the time to fixation or loss of an allele (sojourn time distributions). To cite J.S. Gale's book (p. 81 1990) '*calculating the distributions themselves […] is a very formidable problem when natural selection operates*'.

Deriving such shorter term quantities has obvious applications too: for example, the stochastic time dynamics of any trait encoded by one or several alleles under directional selection can only be fully captured by knowing the distribution of allele frequency trajectories. The analysis of genetic time series would also benefit from an analytical prediction regarding the expected dynamics of genotypes undergoing selection and drift (discussed in Song and Steinrücken, 2012). The sojourn time distribution (which shows a one-to-one relationship to the frequency trajectory) is also a central descriptor of the evolutionary process; to cite J.S. Gale again (p. 81 1990): "*given an evolutionary process, what could be more natural than to ask: 'How long will this process take?'*". In more directly applied terms, since Maynard Smith and Haigh's first model (1974), most if not all models of hitch-hiking between a neutral and a selected allele (or between two selected alleles) depend on a description of the time to fixation of the allele driving the hitch-hiking effect. To date, they are typically handled by using a deterministic approximation, or the mean time to fixation, thus ignoring the full distribution of the sojourn time.

To date, the main tool to obtain the distribution of allelic frequencies at a given time is by numerical solutions of the diffusion equation, by perturbation analysis for small $N_e s$ (Kimura, 1957), or more recently, via spectral analysis of the diffusion operator (Song and Steinrücken, 2012). The latter provides approximate formulae for the frequency distribution, at any time, including the effects of selection, drift and mutation, with extensions to account for dominance (Steinrücken et al., 2013). These formulae are useful for statistical analysis where the frequency at time t is the end-result, the focus of the statistical analysis. However, they remain semi-explicit and still relatively complex to implement (an algorithm

* Corresponding author at: Laboratoire de Probabilités et Modèles Aléatoires CNRS UMR 7599, UPMC Université Paris 06, Paris, France.
E-mail address: amaury.lambert@upmc.fr (A. Lambert).

http://dx.doi.org/10.1016/j.tpb.2015.01.004
0040-5809/

must be iterated, involving the inversion of large-dimensional matrices). Therefore, they can for example not be easily plugged into other evolutionary models (a phenotypic model for example, or a hitch-hiking model), and their implementation for statistical purposes can also prove technical. Also, they do not readily provide a distribution of sojourn times, because the approach does not 'follow' dynamically a set of independent frequency trajectories.

The goal of this article is to provide a complementary tool on these issues. We seek to derive expressions for the distributions of allele frequencies and sojourn times, in a simple closed-form, where simplicity is also a requirement to plug these expressions into more integrated evolutionary models. To do so, we start from the Wright–Fisher diffusion approximation of the exact stochastic process of allelic frequency under selection and genetic drift. Then we approximate this diffusion itself, via a well-known separation of timescales, when the population is large and the allele starts at low frequency (e.g. Barton, 1998; Durrett and Schweinsberg, 2004; Ewing et al., 2011; Kaplan et al., 1989; Stephan et al., 1992). This separation of timescales is explicitly connected to the simpler Feller diffusion process (Feller, 1951), in which many short-term results can be obtained explicitly. We use exact individual-based simulations of a Wright–Fisher model of genetic drift, to check our approximations. Mathematical computations are detailed in Appendix, most can also be re-obtained and checked using an online supplementary (see Appendix B) *Mathematica* ® (Wolfram Research, 2012) notebook file.

2. Material and methods

Consider a diploid population with census size N ($2N$ chromosomes) and "variance" effective size $N_{e(v)}$, where a beneficial allele segregates at frequency p_t ($q_t = 1 - p_t$) at time t. At some locus, a beneficial allele with effect $h\,s$ (resp. s) in heterozygous (resp. homozygous) state starts initially in $k \geq 1$ copies. Following Glemin (2012), we characterize inbreeding by Wright's fixation index F (deviation from Hardy–Weinberg proportion), and allow for (non-extreme) dominance ($0 < h < 1$) and arbitrary inbreeding ($0 \leq F \leq 1$). The latter merely reduces effective size: $N_e = N_{e(v)}/(1+F)$ (Glemin, 2012), which will be implicit in what follows. The dynamics of p_t follow a stochastic differential equation (SDE) corresponding to the Wright–Fisher diffusion approximation (Ewing et al., 2011) adapted to include inbreeding (from Glemin, 2012):

$$\begin{aligned} dp_t &= p_t q_t\,(s_a q_t + s_b p_t)\,dt + \sqrt{\frac{p_t q_t}{2N_e}}dB_t \\ s_a &= s\,(h + (1-h)\,F) \quad \text{and} \quad s_b = s\,(1-(1-F)\,h) \end{aligned} \tag{1}$$

where B_t is a standard Brownian motion, $p_0 = k/2N$ is the initial frequency with k initial copies of the allele, and $s_a > 0$ and $s_b > 0$ for a beneficial allele. This SDE is equivalent to the more classic formulation in terms of their corresponding master equations: it approximates the dynamics of p_t as long as N_e is large enough ($N_e \gg 1$) and s is not too large ($s_a, s_b \ll 1$). The haploid model is characterized by $s_a = s_b = s_*$. The diploid codominant model ($h = 1/2$) reduces to a haploid model with effective size N_e and $s_* = s(1+F)/2$. With complete inbreeding ($F = 1$), the diploid model also reduces to a haploid model with effective size $N_e = N_{e(v)}/2$ and $s_* = s$. Notice that the scaling of time in Eq. (1) is per generations (as is classic in deterministic population genetics models), while, elsewhere in the literature (especially in the context of diffusion approaches), it can also be presented with a scaling of N generations time units.

From Eq. (1), one can compute the probability of fixation from a few copies ($p_0 \ll 1$), or the mean time to fixation ($\tau = \min(t;\, p_t = 1)$, conditional on fixation ($p_\infty = 1$)), or to extinction ($\tau_E = \min(t;\, p_t = 0)$, conditional on extinction ($p_\infty = 0$)). This has been done for a variety of scenarios, including with dominance (Ewing et al., 2011; Ohta and Kimura, 1972), inbreeding (Glemin, 2012), or population structure (Roze and Rousset, 2003; Whitlock, 2003). However, even in the simplest haploid models ($s_a = s_b$), the non-linearity of Eq. (1) with respect to p_t makes analytic progress very cumbersome, beyond averages. To date, no known explicit form exists for the full *distribution* of allele frequencies at t or for sojourn times.

A well-known separation of timescales applies in Eq. (1), when the population is large and the beneficial allele starts at low frequency (e.g. Barton, 1998; Durrett and Schweinsberg, 2004; Ewing et al., 2011; Kaplan et al., 1989; Stephan et al., 1992). Conditional on fixation, p_t is in fact only 'stochastic' while $p_t \to 0$ (early phase A: $t \to 0$) or $q_t \to 0$ (late phase C: $t \to \tau$), the intermediate phase B being approximately deterministic. The resulting early phase A is characterized by a simpler (linear) diffusion (Feller, 1951), akin to a branching process with independently growing types, as first noted by Haldane (1927). It turns out that the same goes for the late phase C. Indeed, series expansions of Eq. (1) to leading order in p_t (phase A), $1/\sqrt{N_e}$ (phase B) or q_t (phase C) illustrate this simplification:

$$\begin{aligned} &\text{phase A}: p_t = o(1): dp_t \approx s_a\,p_t\,dt + \sqrt{\frac{p_t}{2N_e}}dB_t + O(p_t^2 dt) \\ &\text{phase B}: p_t = O(1): dp_t \approx p_t q_t\,(s_a q_t + s_b\,p_t)\,dt \\ &\qquad\qquad + O\left(\frac{1}{\sqrt{N_e}}dt\right) \\ &\text{phase C}: q_t = o(1): dp_t = -dq_t \approx s_b\,q_t\,dt \\ &\qquad\qquad + \sqrt{\frac{q_t}{2N_e}}dB_t + O(q_t^2 dt). \end{aligned} \tag{2}$$

Phase A in Eq. (2) is the SDE of a supercritical Feller diffusion (Feller, 1951) with 'drift' term $s_a > 0$ and 'diffusion' term $1/2N_e$: $p_t \sim \mathit{Feller}(s_a,\, 1/2N_e)$. Conversely, q_t in phase C satisfies the SDE of a subcritical Feller diffusion with drift term $-s_b < 0$ and the same diffusion term $1/2N_e$: $q_t \sim \mathit{Feller}(-s_b,\, 1/2N_e)$. Because Feller diffusions yield many more explicit results than Wright–Fisher diffusions, these two approximations will be helpful.

Fixation probability: Phase A in Eq. (2) is characterized by $p_t \sim \mathit{Feller}(s_a,\, 1/2N_e)$. This directly provides an approximation for the probability of establishment (avoiding loss while rare), which amounts to ultimate fixation here, as $s_a, s_b > 0$. This is valid whenever ultimate extinction *vs.* establishment is decided during the early phase A, namely whenever $k \ll N$ and $N_e s \gg 1$, so that an allele destined to get lost remains at low frequency all along. The probability of fixation from k copies is given by

$$\begin{aligned} P_k &\approx 1 - e^{-\alpha_a\,k/2N} \\ \alpha_a &= 4\,N_e\,s_a. \end{aligned} \tag{3}$$

It can be checked that Eq. (3) converges in the limit $\alpha_a \gg 1$, with the more complex but accurate expression in Eq. (5a) of Glemin (2012), derived from the exact Wright–Fisher diffusion (Eq. (1)).

Semi-deterministic approximation: We first state our central result: In large populations, a selective sweep from n established copies behaves as if it was deterministic from $t = 0$ to $t = \tau$, but started from some gamma distributed "equivalent initial frequency" $p_0 = \tilde{p}_0 \sim \Gamma\,(n,\, 1/(2NP_1))$, and ended at some exponentially distributed "equivalent end frequency" $1 - p_\tau = \tilde{q}_\tau \sim \mathrm{Exp}\,(2NP_1)$.

This result stems from a useful property of Feller diffusions, applied to the early sweep during phase A, namely to $p_t \sim \mathit{Feller}(s_a,\, 1/2N_e)$. Indeed, conditional on non-extinction, a Feller diffusion starting at some $p_0 > 0$, when properly rescaled by its expectation $E\,(p_t) = p_0 e^{s_a t}$, converges to some fixed distribution as