Journal of Theoretical Biology 380 (2015) 98–102

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Revisiting the time until fixation of a neutral mutant in a finite population – A coalescent theory approach

Gili Greenbaum *

Department of Solar Energy and Environmental Physics and Mitrani Department of Desert Ecology, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, 84990, Israel

HIGHLIGHTS

- Fixation times of neutral mutations usually estimated by diffusion approximations.
- Here a coalescent theory approach is used to estimate these fixation times.
- The two approaches converge for large populations but differ for small populations.
- Coalescent approximations are more accurate for small population sizes.

ARTICLE INFO

Article history:
Received 14 December 2014
Received in revised form
26 April 2015
Accepted 11 May 2015
Available online 21 May 2015

Keywords:
Fokker-Planck
Coalescence
Conditional fixation
Genetic drift
Diffusion approximation
Population genetics

ABSTRACT

Evaluation of the time scale of the fixation of neutral mutations is crucial to the theoretical understanding of the role of neutral mutations in evolution. Diffusion approximations of the Wright–Fisher model are most often used to derive analytic formulations of genetic drift, as well as for the time scales of the fixation of neutral mutations. These approximations require a set of assumptions, most notably that genetic drift is a stochastic process in a continuous allele-frequency space, an assumption appropriate for large populations. Here equivalent approximations are derived using a coalescent theory approach which relies on a different set of assumptions than the diffusion approach, and adopts a discrete allele-frequency space. Solutions for the mean and variance of the time to fixation of a neutral mutation derived from the two approaches converge for large populations but slightly differ for small populations. A Markov chain analysis of the Wright–Fisher model for small populations is used to evaluate the solutions obtained, showing that both the mean and the variance are better approximated by the coalescent approach. The coalescence approximation represents a tighter upper-bound for the mean time to fixation than the diffusion approximation, while the diffusion approximation and coalescence approximation form an upper and lower bound, respectively, for the variance. The converging solutions and the small deviations of the two approaches strongly validate the use of diffusion approximations, but suggest that coalescent theory can provide more accurate approximations for small populations.

1. Introduction

The study of the fixation process of neutral mutations in the last decades has been instrumental to the advancement of theoretical population genetics, as well as to the understanding of the evolutionary process. The stochastic process that accompanies the genetic process, known as genetic drift, is responsible to drive newly arisen neutral mutations to one of two eventual outcomes, assuming the absence of other forces – fixation as the sole allele in the population or loss from the population. Both the probability with which these events occur (Kimura, 1962; McKane and Waxman, 2007; Otto and Whitlock, 1997; Whitlock, 2003) and the time scale of the fixation and loss processes (Burrows and Cockerham, 1974; Kimura and Ohta, 1969a, 1969b; Kimura, 1980; Waxman, 2012; Whitlock, 2003) has been extensively studied.

The most widely used model for description of the genetic process is the Wright–Fisher model (Fisher 1922; Wright 1931). In this model discrete generations and a discrete gene pool are assumed. The gene pool in each generation is generated by sampling with replacement using the allele frequency of the

* Tel.: +972 52 3689626.
E-mail address: gili.greenbaum@gmail.com

http://dx.doi.org/10.1016/j.jtbi.2015.05.019
0022-5193/

previous generation's gene pool, thus inducing a binomial distribution on the allele frequencies in the following generation. One of the main mathematical approaches to deal with genetic drift is the diffusion approximation of the Wright–Fisher model, first introduced by Fisher (1922) and Wright (1931), and later developed and extended by Kimura (1964). The diffusion approach is a *prospective* approach that approximates the process by a diffusion of the probability density along the frequency space, assumed to be continuous, over time. This approach has had a significant influence on the development of the theory of neutral genetic variation and modern evolutionary theory, as it enables the description of quantities for many phenomena, such as the time to fixation and loss of alleles (Kimura and Ohta, 1969a, 1969b). An important special case of these results describes the time of fixation of a neutral mutation conditioned on the occurrence of such a fixation, a process known as the 'conditional fixation process' (Kimura and Ohta, 1969a).

Another approach that was developed to study genetic drift and genetic processes is coalescent theory (Kingman, 1982a). This approach adopts a *retrospective* backwards-looking viewpoint on the genetic process. Instead of asking how a certain generation's allele frequency affects the allele frequencies in the following generation, the coalescent approach is to ask how long ago lineages of copies of a certain allele separated from their most recent common ancestor (MRCA). Put in other words, the coalescence time of copies of an allele is the time it would take for them to coalesce if the process was to be run backwards. This approach has considerable computational and practical advantages when addressing questions regarding processes that occurred in the past, since only surviving lineages need be taken into account (Rosenberg and Nordborg, 2002). The first papers of coalescent theory dealt with simple ideal populations, but like the diffusion approach, they were later supplemented with many studies extending the approach's principles to various violations of the ideal population (Hein et al., 2005; Nordborg, 2008).

Deriving analytic equations to describe the timescale of the conditional fixation process is crucial to the understanding of the polymorphism observed in many loci and in many organisms, since neutral mutations on their way to fixation induce polymorphism in a monomorphic locus, or enhance the polymorphism of an already polymorphic locus for the time scale of the fixation process (Kimura, 1984). Thus equations describing the time scale of the conditional fixation process, accompanied by the mutation rate and the probability of fixation of neutral mutations, can be applied to the analysis of the observed genetic variation in nature, and to explain the observed polymorphism in many organisms and loci. Deriving approximations to the Wright–fisher model in order to generate predictions for the timescales of the conditional fixation process of mutations is also instrumental in the study of the role of neutral mutations in evolution, as a source of genetic variation (Kimura, 1984; Ohta, 1992), and has been shown to effect population's evolvability (Draghi and Wagner, 2008; Wagner, 2008).

While the importance of Kimura's and Ohta's original result regarding the time to fixation of a neutral mutant has been acknowledged (Watterson, 1996), an analytic confirmation of this result using an alternative approach is of importance. This is so since the diffusion approach forces continuity on the frequency space, an assumption that is violated especially in small populations, which are of particular interest in population genetics and conservation genetics. In this study the standard coalescent (n-coalescent) model (Kingman, 1982b) of the Wright–Fisher model is used, a model that does not assume a continuous frequency space, to derive the mean time of the conditional fixation process of a mutation, as well as its variance. It is shown that the conditional fixation and the coalescence of the entire population are processes with identical time scales (although they do not occur simultaneously and only partially overlap; Campbell, 1999), but result in a different approximation of the conditional fixation process. The diffusion approximation and the coalescence approximation converge for large populations, but differ for finite populations. Explicit Markov chain analysis of the Wright–Fisher model focusing on small populations ($N \leq 100$) are used to compare the diffusion approximations with the coalescence approximations.

2. The diffusion and coalescence approximations

2.1. The diffusion approximation of the conditional fixation process

The diffusion approach is based on formulating the Wright–Fisher model using the Fokker-Plank diffusion equation (known also as the Kolmogorov forward equation), and approximates the probability density of the allele frequency over time. In order to describe a diffusion process, the frequency space is formulated as a continuous random variable rather than a discrete random variable, although in a finite population an allele can attain only a finite number of different allele frequencies (Waxman, 2011). Under this formulation, the mean and variance of $T(p,N)$, the time to fixation of a neutral allele with initial frequency p in a population of size N (the full formulation takes into account the variance effective population size N_e, but here the population is considered to be ideal, thus $N=N_e$), is given by (Kimura and Ohta, 1969a; Narain, 1970)

$$E[T(p,N)] = \frac{-4N(1-p)ln(1-p)}{p} \tag{1}$$

$$Var[T(p,N)] = 32N^2\left[\frac{(1-p)ln(1-p)}{p} + \sum_{k=1}^{\infty}\frac{1-p^k}{k^2}\right] - (E[T(p,N)])^2 \tag{2}$$

For a neutral mutation, which initially appears in the population at only one copy, the diffusion approximation for the mean time to (conditional) fixation is therefore

$$E\left[T_{fix}(N)\right] = -8N^2\left(1-\frac{1}{2N}\right)\ln\left(1-\frac{1}{2N}\right) \approx 4N-1 \tag{3}$$

Note that the second order approximation of $\ln\left(1-\frac{1}{2N}\right)$ is needed here, as the second order term is significant and cannot be neglected. The diffusion approximation for the variance is given by

$$\begin{aligned}\mathrm{Var}\left[T_{fix}(n)\right] &= 32N^2\left[(2N-1)\ln\left(1-\frac{1}{2N}\right) + \sum_{k=1}^{\infty}\frac{1-\left(\frac{1}{2N}\right)^k}{k^2}\right] \\ &\quad -(E\left[T\left(\frac{1}{2N},N\right)\right])^2 \approx \left(\frac{16}{3}\pi^2-48\right)N^2-\frac{1}{3}\end{aligned} \tag{4}$$

The two convergences in Eqs. (3) and (4) are rapid and can be applied to small population sizes as well. Note that when taking $p\to 0$ instead of $p=\frac{1}{2N}$ to demonstrate the conditional fixation time of a neutral mutation, the result is Eqs. (3) and (4) without the constant terms, and these are the results that most often appear in the literature (and are more convenient when dealing with large populations and continuous frequency spaces).

The diffusion approximation of the Wright–Fisher model is extensively used in population genetics to obtain approximations of biologically significant phenomena, such as the time to fixation of a neutral mutation. The diffusion approximation of a neutral allele converges to the Wright–Fisher model for large populations (Guess, 1973). The error of the diffusion approximation has been studied both numerically and analytically (Ethier and Norman, 1977; Ewens, 1963; Kimura, 1980; Zhao et al., 2013), and have mostly been found to be quite accurate. However, it has been pointed out that the diffusion theory may be vulnerable near the