Theoretical Population Biology 111 (2016) 51–64

Contents lists available at ScienceDirect

Theoretical Population Biology

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

The non-equilibrium allele frequency spectrum in a Poisson random field framework

Ingemar Kaj [a,*], Carina F. Mugal [b]

[a] Department of Mathematics, Uppsala University, Uppsala, Sweden
[b] Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden

ARTICLE INFO

Article history:
Received 15 December 2015
Available online 1 July 2016

Keywords:
Population genetics
Non-equilibrium allele frequency spectrum
Poisson random field
Coalescent theory
Duality relation
Natural selection

ABSTRACT

In population genetic studies, the allele frequency spectrum (AFS) efficiently summarizes genome-wide polymorphism data and shapes a variety of allele frequency-based summary statistics. While existing theory typically features equilibrium conditions, emerging methodology requires an analytical understanding of the build-up of the allele frequencies over time. In this work, we use the framework of Poisson random fields to derive new representations of the non-equilibrium AFS for the case of a Wright–Fisher population model with selection. In our approach, the AFS is a scaling-limit of the expectation of a Poisson stochastic integral and the representation of the non-equilibrium AFS arises in terms of a fixation time probability distribution. The known duality between the Wright–Fisher diffusion process and a birth and death process generalizing Kingman's coalescent yields an additional representation. The results carry over to the setting of a random sample drawn from the population and provide the non-equilibrium behavior of sample statistics. Our findings are consistent with and extend a previous approach where the non-equilibrium AFS solves a partial differential forward equation with a non-traditional boundary condition. Moreover, we provide a bridge to previous coalescent-based work, and hence tie several frameworks together. Since frequency-based summary statistics are widely used in population genetics, for example, to identify candidate loci of adaptive evolution, to infer the demographic history of a population, or to improve our understanding of the underlying mechanics of speciation events, the presented results are potentially useful for a broad range of topics.

1. Introduction

The allele frequency spectrum (AFS) describes the distribution of allele frequencies over a large number of identical and independent loci. In practice, the AFS is estimated by allele frequencies recorded in a sample of individuals. Here, the recent progress in whole-genome re-sequencing has significantly improved the accessibility of the AFS, and several allele frequency-based summary statistics have become central measurements in population genetic studies. The estimation of the AFS is then often based on polymorphic nucleotide sites, where the frequency of the derived allele over a finite collection of sites in the sample is summarized. In this context, the otherwise equivalent term 'site frequency spectrum' (SFS) is frequently used. For the purpose of generality, we here use the term AFS.

The theory of the AFS was initiated in the 1930s with the classical work of Fisher and Wright in the framework of diffusion theory including effects of natural selection (Fisher, 1930; Wright, 1931, 1938). Subsequently, Kimura (1964) pioneered the systematic use of stochastic processes in population genetics, and developed the theory further. In particular, he considered the equilibrium distribution of allele frequencies under irreversible mutation in an ensemble of polymorphic loci (Kimura, 1970b). Central to these successful applications of diffusion theory in describing the equilibrium limit AFS for various mutation and selection scenarios is the Green function representation of diffusion process occupation time functionals (Karlin and Taylor, 1981). Then, in order to study the impact of natural selection on the number of fixations in diverging species, Sawyer and Hartl (1992) introduced the Poisson random field framework. The basic assumptions of this approach are that new mutant alleles arise according to a Poisson process, mutations are irreversible, and the frequencies of the descendants of each mutation are described by independent Markov processes (no linkage). The loss or fixation of a mutant allele is captured by the separate events of extinction or

* Corresponding author.
E-mail addresses: ikaj@math.uu.se (I. Kaj), carina.mugal@ebc.uu.se (C.F. Mugal).

http://dx.doi.org/10.1016/j.tpb.2016.06.003
0040-5809/

fixation of the Markov process. This collection of Markov processes forms a Poisson random field in the sense that the limiting distributions of the allele frequencies are independent Poisson random variables. In particular, the number of fixations is a Poisson random variable with expected value increasing linearly over time. Segregating mutations are, on the other hand, in equilibrium with respect to time, and hence the marginal distributions of the corresponding Poisson variables are stationary. In other words, the AFS is assumed to be in equilibrium with respect to time.

More recently, Evans et al. (2007) initiated the study of the non-equilibrium AFS in a single population including the effects of natural selection, in the sense of deriving a function $f(t, x)$ which represents the expected fraction of alleles of frequency x existing at some time t, given an initial fraction $f(0, x)$ of alleles at time $t = 0$. Some of the modeling parameters, such as population size and selection intensity, are also allowed to depend on time. The resulting non-equilibrium AFS $f(t, x)$ is provided as a solution to a partial differential equation (PDE), essentially the Kolmogorov forward equation for the corresponding diffusion, linked to a given rate of mutational influx via a specific boundary condition of $f(t, x)$ as $x \to 0$. An additional approximation method using moments is employed to study the resulting allele frequencies in a sample. Building on this approach, Zivkovic and Stephan (2011) provide analytical results on the non-equilibrium AFS for the neutral case, focusing on time-dependence arising due to changes in population size. In the same direction, Zivkovic et al. (2015) consider the case of natural selection and develop the moment approximation method for a scenario of piecewise-constant population size starting from an equilibrium.

In a parallel methodological track the AFS has been studied using the view of coalescent theory, where mutations are randomly placed on the branches of a genealogy of a sample of individuals (Kingman, 1982). First, Fu (1995) obtained a representation of the stationary AFS for a single population under the assumptions of neutrality and constant population size, by deriving the mean and variance of the number of mutations on each branch of a given length. Griffiths and Tavaré (1998) explored the duality relation between the neutral Wright–Fisher diffusion process and Kingman's pure death coalescent process further and addressed deterministic changes in population size. Moreover, Wakeley and Hey (1997) obtained a description of the joint AFS of two isolated populations descending from a common ancestor under neutrality. Chen (2012) elaborated on their work and extended it to multiple populations and also modeled scenarios such as selective sweeps, influx of migration and changes in population size.

Here, we build on the work of Sawyer and Hartl (1992) and develop the approach of Mugal et al. (2014) further to derive a representation of the non-equilibrium AFS as the limiting expected value of a suitable Poisson stochastic integral. The model is developed in steps starting with finite population size N, where individuals are represented by a collection of L independent sites, subject to mutational influx of derived alleles and Wright–Fisher reproduction in discrete generation time. Assuming that the mutation rates and selection coefficients per individual per generation are of order $1/N$ and rescaling evolutionary time t so that it is measured in units of N generations, we then pass to the continuous-time Wright–Fisher diffusion approximation, but follow Evans et al. (2007) in keeping N as a modeling parameter. In the next stage of approximation, the mutation rate per site is assumed to tend to zero in such a way that the total mutation rate across the collection of sites is constant, a procedure which we interpret and implement as a limit in distribution as $L \to \infty$. The result is a Poisson random field parametrized by N, which we study in some detail. Then, we find the limiting expected values as $N \to \infty$ and identify the time-dependent AFS which arises in the limit. Thereby, we provide a link between the Poisson random field approach by Sawyer and Hartl (1992) and the setting of Evans et al. (2007), in particular by identifying the PDE solution $f(t, x)$ in terms of a Wright–Fisher fixation time probability distribution. An additional representation is obtained by elaborating on the duality relation between the Wright–Fisher diffusion process and a class of birth and death processes, where birth rates are proportional to the strength of selection (Shiga and Uchiyama, 1986; Athreya and Swart, 2005).

2. Poisson random field model

2.1. Basic markov chain model

Consider a population containing N haploid individuals, where each individual is represented by a collection of L independent sites. Random mutation events act on sites, independently and uniformly over individuals, replacing an ancestral allele by a derived allele. Only mono-allelic sites are affected by mutation. Thus, the setting of the model only allows for two alleles, the derived and the ancestral, in each site. The composition of ancestral and derived alleles per site changes in discrete steps from one generation to the next according to the Wright–Fisher model with selection, which relies on the following assumptions (1) non-overlapping generations, (2) constant population size and (3) random mating. The population dynamics are then given by a collection of independent, identically distributed Markov chains in discrete time, $\{(X_n^i)_{n\geq 0},\ 1 \leq i \leq L\}$, one component for each site. The state variable is

$$X_n^i = \#\text{ of individuals in generation } n \text{ with the derived allele in site } i$$

and the state space of each chain is $\{0, 1, \ldots, N\}$. An example path of the Markov chain is visualized in Fig. 1. Site i is said to be mono-allelic at time n if it carries the ancestral allele throughout the entire population, so that $X_n^i = 0$. A trajectory $(X_n^i)_{n\geq 0}$ consists of subsequent mono-allelic periods in state 0 and active polymorphic periods with both ancestral and derived alleles present. Whenever a derived allele reaches fixation in generation n, that is $X_n^i = N$, then the derived allele is declared to be the new ancestral allele at that site.

We let $\mu > 0$ be the mutation probability

$$\mu = \text{probability per individual and generation that an ancestral allele is replaced by the derived allele at a single mono-allelic site,}$$

and for each generation n and site i, we let J_n^i be independent, binomially distributed random variables, such that for $i = 1, \ldots, L$, $n \geq 1$,

$$J_n^i = \#\text{ of mutations in generation } n \text{ hitting a mono-allelic site } i \in \mathrm{Bin}(N, \mu).$$

In the limit of small mutation rate $\mu \to 0$, such that $N\mu$ is a small probability, we have

$$P(J_n^i = 0) = (1-\mu)^N = 1 - N\mu + o(N\mu)$$

as well as

$$P(J_n^i = 1) = N\mu + o(N\mu), \qquad P(J_N^i \geq 2) = o(N\mu).$$

Hence, given X_n^i in generation n, the random variable

$$J_{n+1}^i 1_{\{X_n^i = 0\}} = \#\text{ of mutations in site } i \text{ at generation } n+1$$

is approximately $\mathrm{Bin}(1, N\mu)$ distributed, for each i. It is the injection of new derived alleles into the population at mono-allelic sites, and the change-of-state of the Markov chain from 0 to 1,