Theoretical Population Biology 81 (2012) 197–209

Contents lists available at SciVerse ScienceDirect

Theoretical Population Biology

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

The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates

Claus Vogl [a,*], Florian Clemente [b]

[a] Institute of Animal Breeding and Genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria

[b] Institut of Population Genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria

ARTICLE INFO

Article history:
Received 17 March 2011
Available online 13 January 2012

Keywords:
Moran model
Mutation bias
Genic selection
Small mutation rates
Small samples
Drosophila melanogaster

ABSTRACT

We analyze a decoupled Moran model with haploid population size N, a biallelic locus under mutation and drift with scaled forward and backward mutation rates $\theta_1 = \mu_1 N$ and $\theta_0 = \mu_0 N$, and directional selection with scaled strength $\gamma = sN$. With small scaled mutation rates θ_0 and θ_1, which is appropriate for single nucleotide polymorphism data in highly recombining regions, we derive a simple approximate equilibrium distribution for polymorphic alleles with a constant of proportionality. We also put forth an even simpler model, where all mutations originate from monomorphic states. Using this model we derive the sojourn times, conditional on the ancestral and fixed allele, and under equilibrium the distributions of fixed and polymorphic alleles and fixation rates. Furthermore, we also derive the distribution of small samples in the diffusion limit and provide convenient recurrence relations for calculating this distribution. This enables us to give formulas analogous to the Ewens–Watterson estimator of θ for biased mutation rates and selection. We apply this theory to a polymorphism dataset of fourfold degenerate sites in *Drosophila melanogaster*.

1. Introduction

In the limit of relatively high recombination and small mutation rates, each polymorphic site can be considered independent from the rest of the genome. The distribution of allele frequencies at a large number of such loci has been called the "allele-frequency spectrum" or "site-frequency spectrum". In a classical manuscript, Wright (1931) introduced a bi-allelic equilibrium model and derived the equilibrium allele frequency distribution, up to a constant of proportionality. Most recent treatments of similar models, however, assume irreversible mutations (*e.g.,* Sawyer and Hartl, 1992; Hartl et al., 1994; Bustamante et al., 2001; Griffiths, 2003; Ewens, 2004; Evans et al., 2007). If mutation rates are low and an outgroup is available to infer the ancestral state, *i.e.,* if states can be polarized, theory assuming irreversibility allows inference of selection coefficients for polymorphic sites. The quality of polarization and thus the quality of inference under this model depends on the relative age of outgroups: if outgroups are too closely related, polymorphism shared among species is problematic; if outgroups are too distantly related, double mutations may obscure patterns. Thus, for real data analysis, a model allowing for back mutations may be better suited. Furthermore, if mutation parameters are to be estimated in addition to the selection coefficient, an approach using reversible mutations is necessary. Relatively recently, McVean and Charlesworth (1999) reconnect to earlier work to derive some statistics for the allele-frequency spectrum and provide such an approach. Zeng and Charlesworth (2009, 2010) use the Wright–Fisher model and forward simulations to infer parameters using sequence data with a reversible mutation model (Shapiro et al., 2007).

In population genetics theory, the Wright–Fisher model (Fisher, 1930; Wright, 1931) and later the Moran model (Moran, 1958) have received the most attention among the explicit models moving forwards in time. Many classic results were derived using diffusion theory (Fisher, 1930; Wright, 1931; Kimura, 1955a,b). A key parameter in population genetics is the population size N. In the limit of large N (usually a reasonable assumption), results from different models and approaches converge. Diffusion theory can be seen either as a model in its own right or as an approximation to the explicit models in the limit of large N. Since we are mostly interested in this limit, the mathematically most tractable approach has been used, usually diffusion theory (Ewens, 2004). The models and approaches discussed so far move forward in time. Since the 1980's, the coalescent (Kingman, 1982), an approach that looks backward in time, has been used to derive insights into the distribution of small samples and into the genealogic tree behind allelic distributions.

* Corresponding author.
E-mail address: claus.vogl@vetmeduni.ac.at (C. Vogl).

0040-5809
doi:10.1016/j.tpb.2012.01.001

Using a Moran model, Muirhead and Wakeley (2009) showed that exact equilibrium solutions (up to a constant of proportionality) can be obtained relatively easily for population genetic models with mutation, drift, and frequency-dependent selection, both for infinitely many and a finite number of K-alleles. Some of their results go beyond those readily available by diffusion theory. Baake and Bialowons (2008) and Etheridge and Griffiths (2009) use a Moran model where mutation, selection, and drift are decoupled. With this model, Etheridge and Griffiths (2009) derive formulas for mutation, drift, and genic selection and show that they converge to the usual diffusion derived formulas in the limit of large N. Furthermore, boundary conditions are rather difficult to incorporate into diffusion theory (*e.g.,* Evans et al., 2007). This argues for multiple approaches to population genetics problems, challenging the nearly exclusive focus on diffusion theory in forward models.

Starting from a decoupled Moran model (Baake and Bialowons, 2008; Etheridge and Griffiths, 2009), we concentrate particularly on small scaled mutation rates ($\theta_0 \ll 1$ and $\theta_1 \ll 1$) with directional selection. We derive theory analogous to a model without selection and apply it to a dataset of *Drosophila melanogaster* introns and fourfold degenerate sites (Shapiro et al., 2007).

2. Small θ without selection

In this section, we re-derive known results for the case without selection, i.e., the mutation-drift model. We show how results derived for the infinite sites model follow from the general case for small scaled mutation rates, i.e., with θ_0 and θ_1 small and of order $\theta \ll 1$.

Without selection, the mutation-drift equilibrium distribution of a locus with two alleles is known to be beta-binomially distributed in the diffusion limit and also in the decoupled Moran model (Baake and Bialowons, 2008; Etheridge and Griffiths, 2009), which we will introduce in more detail below. The probability of finding $y = i$ copies of allele one in a small sample of size n, with $0 < i < n$, is:

$$\Pr(y = i \mid n, \theta_0, \theta_1) = \frac{n!}{(\theta_0+\theta_1)_n} \frac{(\theta_0)_{n-i}}{(n-i)!} \frac{(\theta_1)_i}{i!}, \tag{1}$$

where $(a)_i$ is the rising factorial or Pochhammer function: $(a)_i = a(a+1)(a+2)\cdots(a+i-1)$ and $(a)_0 = 1$. For small θ, we have $(\theta)_i = \theta(i-1)! + O(\theta^2)$. Therefore, formula (1) becomes for $0 < i < n$:

$$\begin{aligned}\Pr(y = i \mid n, \theta_0, \theta_1) &= \frac{\theta_0\theta_1}{\theta_0+\theta_1} \frac{n}{i(n-i)} + O(\theta^2) \\ &= \frac{\theta_0\theta_1}{\theta_0+\theta_1} \left(\frac{1}{i} + \frac{1}{n-i}\right) + O(\theta^2). \end{aligned} \tag{2}$$

Here, $\theta_0\theta_1/(\theta_0+\theta_1)$ serves as an approximate constant of proportionality. For $y = 0$ and $y = n$, we have $\Pr(y = 0 \mid n, \theta_0, \theta_1) = \theta_0/(\theta_0+\theta_1) + O(\theta^2)$ and $\Pr(y = n \mid n, \theta_0, \theta_1) \approx \theta_1/(\theta_0+\theta_1) + O(\theta^2)$, respectively. For a sample of L loci, the expectation of the sum of all polymorphic loci then is to first order in θ:

$$L \sum_{i=1}^{n-1} \frac{\theta_0\theta_1}{\theta_0+\theta_1} \left(\frac{1}{i} + \frac{1}{n-i}\right) = L \frac{2\theta_0\theta_1}{\theta_0+\theta_1} \sum_{i=1}^{n-1} \frac{1}{i}. \tag{3}$$

This recapitulates formula (17) in RoyChoudhury and Wakeley (2010). It can be rearranged to give a method of moments estimator of polymorphism in a sample that extends the Ewens–Watterson estimator of molecular variation $\hat{\theta}_w$ (Ewens, 1974; Watterson, 1975) to biased mutation rates. If the mutation rates are balanced, *i.e.,* $\theta_0 = \theta_1 = \theta$, formula (3) reduces to $L\theta \sum_{y=1}^{n-1} 1/y$. This estimator has been derived with the infinite-sites model that assumes negligible scaled mutation rates θ.

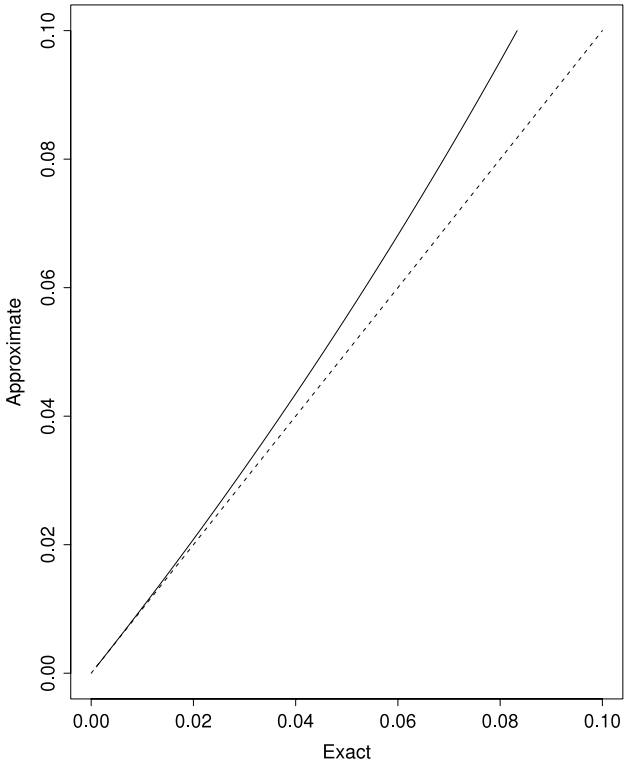

Fig. 1. Comparison of the exact versus the approximate probability of polymorphism in a sample of size $n = 2$ (solid line). The dashed line shows equality.

Obviously, the quality of the approximation depends on the amount of polymorphism: according to our simulations, $2\theta_0\theta_1/(\theta_0+\theta_1)$ should be below 0.05, or better 0.02 (compare also: Desai and Plotkin, 2008). In Fig. 1, we plot the exact versus the approximate probability of polymorphism in a sample of $n = 2$.

We note that in the case without selection, the same formulas also hold for $n = N$, *i.e.,* for the equilibrium distribution of the whole population with N haploid individuals. With selection, the case of small θ_0 and θ_1 has not been explored extensively. It is not known yet, if formulas similar to (1)–(3) can also be derived.

3. The decoupled Moran model with mutation, selection, and drift

In this section, we re-derive the equilibrium distribution of the decoupled Moran model, up to a constant, by showing that this distribution satisfies detailed balance. Baake and Bialowons (2008) and Etheridge and Griffiths (2009) use the same modified Moran model for their derivations. For the case of small mutation rates θ, we will derive a simple constant of proportionality and the allele-frequency spectrum, sojourn times, and divergence rates conditional on the ancestral and fixed allele.

3.1. Basic model

With the Moran model, generations overlap. It moves from step t to step $t+1$; between steps, exponentially distributed waiting times may be introduced. In the pure-drift case, a constant population of N haploid individuals is assumed. In a birth/death event, a random individual j dies and is replaced by the offspring of a randomly chosen individual i. The process repeats indefinitely. The lifespan of an individual is geometrically distributed with a