Theoretical Population Biology 108 (2016) 36–50

Contents lists available at ScienceDirect

Theoretical Population Biology

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

The multivariate Wright–Fisher process with mutation: Moment-based analysis and inference using a hierarchical Beta model

Asger Hobolth [a,*], Jukka Siren [b]

[a] Bioinformatics Research Center, Aarhus University, Denmark
[b] Department of Biosciences, University of Helsinki, Finland

ARTICLE INFO

Article history:
Received 10 June 2015
Available online 29 November 2015

Keywords:
Allele frequency
Diffusion
Dirichlet model
Hierarchical Beta
Moments
Multivariate Wright–Fisher

ABSTRACT

We consider the diffusion approximation of the multivariate Wright–Fisher process with mutation. Analytically tractable formulas for the first-and second-order moments of the allele frequency distribution are derived, and the moments are subsequently used to better understand key population genetics parameters and modeling frameworks. In particular we investigate the behavior of the expected homozygosity (the probability that two randomly sampled genes are identical) in the transient and stationary phases, and how appropriate the Dirichlet distribution is for modeling the allele frequency distribution at different evolutionary time scales. We find that the Dirichlet distribution is adequate for the pure drift model (no mutations allowed), but the distribution is not sufficiently flexible for more general mutation models. We suggest a new hierarchical Beta distribution for the allele frequencies in the Wright–Fisher process with a mutation model on the nucleotide level that distinguishes between transitions and transversions.

1. Introduction

Present day data sets for studying genetic variation within and between species often consist of millions of markers and hundreds to thousands of individuals. The huge number of individuals makes tree-based analyses (e.g. based on phylogenetics or coalescent theory) difficult because the number of possible trees increases very fast with the number of individuals. This difficulty is pronounced when studying closely related species, where incomplete lineage sorting or deep coalescence events can distort phylogenetic analyses (Maddison, 1997). The discrepancy between species trees and gene trees can be taken into account by using multispecies coalescence methods (Degnan and Rosenberg, 2009; Heled and Drummond, 2010). However, this more detailed framework is computationally more challenging because the unknown gene trees need to be marginalized out from the model in order to carry out species tree inference. In some special cases the gene trees can be marginalized out using dynamic programming techniques (e.g. Bryant et al., 2012), but in general it is necessary to perform large-scale Monte Carlo simulations to do the integration (Heled and Drummond, 2010).

An attractive alternative to tree-based methodology is to model the allele frequencies over time in terms of a diffusion process, which is derived as an infinite population limit of the Wright–Fisher model. Unfortunately the transition density for the diffusion process corresponding to the basic Wright–Fisher model with a general mutation model remains unknown; the solution to the Fokker–Planck equation is not available (Ewens, 2004, Chapter 5). We emphasize, however, that Griffiths (1979) provides an expression for the transition density for the multivariate Wright–Fisher diffusion process in a mutation model where the mutation rate q_{ij} from allele i to allele j only depends on j, i.e. $q_{ij} = q_j$ (the so-called parent-independent mutation model). The expression in Griffiths (1979) is in terms of orthogonal polynomials. Griffiths and Spanó (2010) provide an overview of spectral expansions of the transition density for the general Wright–Fisher process in two dimensions and the parent-independent mutation model in more than two dimensions.

Numerical approximations have also been proposed to approximate the transition density, but they are limited to a small number of populations or species due to computational complexity (e.g. Gutenkunst et al., 2009). The numerical solutions also assume that each site has experienced at most one mutation and consequently

* Correspondence to: Bioinformatics Research Center, Aarhus University, C.F. Møllers Alle, Building 1110, DK-8000 Aarhus C, Denmark.
E-mail addresses: asger@birc.au.dk (A. Hobolth), jukka.p.siren@helsinki.fi (J. Siren).

http://dx.doi.org/10.1016/j.tpb.2015.11.001
0040-5809/

has at most two alleles, which restricts their usage to closely related samples. For more distantly related samples where multiallelic loci are expected to occur, it is important to generalize to the multivariate case (Jenkins et al., 2014).

An alternative strategy to numerically solve the Fokker–Planck equation is to approximate the transition density by a parametric distribution. This methodology has a long tradition in population genetics and computational phylogenetics starting from the seminal work by Edwards, Cavalli-Sforza and Felsenstein in the 1960s and 1970s (Edwards and Cavalli-Sforza, 1964; Cavalli-Sforza and Edwards, 1967; Felsenstein, 1973), and continuing topresent day (Nicholson et al., 2002; Gaggiotti and Foll, 2010; Sirén et al., 2011; Pickrell and Pritchard, 2012). However, most of the methods have been developed for situations where the time span is sufficiently short to ignore mutations and consider only pure drift. Furthermore, the parametric distributions have been either the Gaussian or Dirichlet distributions.

We derive the first- and second-order moments of the multiallelic Wright–Fisher process with mutation and use the moments to characterize genetic variation and to fit parametric models. Our approach generalizes the work by Sirén (2012) and Sirén et al. (2013) to arbitrary mutation models. In the first part of the paper (Section 2) we provide new analytically tractable formulas for the first- and second-order moments of the multivariate Wright–Fisher model with mutation. These new formulas allow us to characterize the expected mean and (co)variance of the frequency of an allele, and in particular we investigate in detail the expected homozygosity (Section 3). Furthermore we demonstrate how our formulas can be used to re-derive previous results for the various general symmetric models considered in Griffiths (1980). We emphasize that our mutation structure is completely unrestricted.

In the second part of the paper (Section 4) we use the expressions for the means and (co)variances of the allele frequencies to obtain insight into approximate models for the allele frequency distribution over time. In particular we find that while the Dirichlet model is a suitable approximate model for the allele frequency distribution in the Wright–Fisher process with no mutation (pure drift), it is not appropriate for the Wright–Fisher model with a mutation structure that corresponds to the Kimura model. Instead, we propose a novel hierarchical Beta model for the Wright–Fisher process with Kimura mutations. The paper ends with a brief summary of our main findings, and a discussion of similar methodology.

2. First- and second-order moments in the Wright–Fisher with mutation process

We consider a constant-sized haploid population with N individuals. We denote by $z(m) = (z_1(m), \ldots, z_K(m))$ the row-vector of the number of alleles $1, \ldots, K$ in generation m, and we let U be the $K \times K$ mutation probability matrix such that U_{ij} is the probability for a mutation from allele i to allele j in a generation. The Wright–Fisher model with mutation is then given by the multinomial distribution

$$z(m+1)|z(m) \sim \text{Mult}(N, x(m)U), \tag{1}$$

where $x(m) = z(m)/N$ is the allele frequency in generation m.

We are now in a position to formulate our main result:

Theorem 1 (*General Formulas for the Mean and Variance in the Wright–Fisher with Mutation Process*)**.** *Consider the K-allele Wright–Fisher model with mutation probability matrix U and with initial allele frequency $x(0)$. Define the rate matrix $Q = N(U - I)$. In the diffusion approximation the mean of the allele frequency is given by*

$$\mathrm{E}[x(t)|x(0)] = x(0)e^{Qt}, \tag{2}$$

and the variance is given by

$$\mathrm{Var}[x(t)|x(0)] = \int_0^t e^{-s}(e^{Qs})'\mathrm{diag}\{x(0)e^{Q(t-s)}\}(e^{Qs})ds - (e^{Qt})'x(0)'x(0)e^{Qt}(1 - e^{-t}). \tag{3}$$

Here we make use of a slight abuse of notation such that $x(t)$ is the allele frequency distribution in generation Nt.

Despite the huge interest in the Wright–Fisher process we believe the clean formula for the variance is a new result.

Proof. Repeated use of the law of total expectation gives the mean value

$$\begin{aligned}\mathrm{E}[x(m)] &= \mathrm{E}\big[\mathrm{E}[x(m)|x(m-1)]\big] = \mathrm{E}[x(m-1)U] \\ &= \mathrm{E}[x(m-1)]U = \cdots = x(0)U^m,\end{aligned}$$

where for ease of notation we have omitted the conditioning on $x(0)$. We approximate U^m as follows

$$\begin{aligned}U^m &= U^{tN} = \left[\left\{I + (U - I)\right\}^N\right]^t = \left[\left\{I + Q/N\right\}^N\right]^t \\ &\approx (e^Q)^t = e^{Qt},\end{aligned}$$

where we scale time as $m = tN$ and define $Q = N(U - I)$. Note that with this definition Q becomes a rate matrix where off-diagonal entries are non-negative and rows sum to zero. Thus we have, with a small abuse of notation,

$$\mathrm{E}[x(t)|x(0)] = x(0)e^{Qt}.$$

The proof of the variance is more involved, but the main idea is to make repeated use of the law of total variance. The proof can be found in Appendix A. □

Many procedures are available for calculating matrix exponentials (e.g. Moler and Van Loan, 2003), so a numerical calculation of the mean is straight forward. Calculating the variance is more difficult. In Appendix B we provide an analytical expression for the mean and variance in the case of a reversible mutation matrix. The expression is based on an eigenvalue decomposition of the rate matrix.

There is a long tradition for careful investigation of mutation models in phylogenetics (e.g. Felsenstein, 2004, Chapter 13). In this paper we consider in particular the pure drift model ($U = I$; see Corollaries 3 and 5), the Jukes–Cantor model ($U_{ij} = u, \ i \neq j$; see Corollaries 4 and 8), and the symmetric model ($U = U'$; see Theorems 7 and 9). We give special attention to the Kimura model (Felsenstein, 2004, page 196–200) with $K = 4$ and mutation probability matrix

$$U_{ij} = \begin{cases} \kappa u & \text{if mutation } i \to j \text{ is a transition} \\ u & \text{if mutation } i \to j \text{ is a transversion} \\ 1 - (\kappa + 2)u & \text{if } i = j \end{cases}$$

or, equivalently,

$$Q_{ij} = \begin{cases} N\kappa u & \text{if mutation } i \to j \text{ is a transition} \\ Nu & \text{if mutation } i \to j \text{ is a transversion} \\ -N(\kappa + 2)u & \text{if } i = j. \end{cases} \tag{4}$$

We parameterize the rate matrix using either $\alpha = N\kappa u$ (the rate for a transition) and $\beta = Nu$ (the rate for a transversion), or using $\kappa = \alpha/\beta$ (the ratio of the transition rate and transversion rate) and $\theta = \alpha + 2\beta = N(\kappa + 2)u$ (the mutation rate).