Theoretical Population Biology 112 (2016) 22–32

Contents lists available at ScienceDirect

Theoretical Population Biology

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

An approximate stationary solution for multi-allele neutral diffusion with low mutation rates

Conrad J. Burden [a,b,*], Yurong Tang [a]

[a] Mathematical Sciences Institute, Australian National University, Canberra, Australia
[b] Research School of Biology, Australian National University, Canberra, Australia

ARTICLE INFO

Article history:
Received 17 March 2016
Available online 2 August 2016

Keywords:
Multi-allele Wright–Fisher
Neutral evolution
Forward Kolmogorov equation

ABSTRACT

We address the problem of determining the stationary distribution of the multi-allelic, neutral-evolution Wright–Fisher model in the diffusion limit. A full solution to this problem for an arbitrary $K \times K$ mutation rate matrix involves solving for the stationary solution of a forward Kolmogorov equation over a $(K-1)$-dimensional simplex, and remains intractable. In most practical situations mutations rates are slow on the scale of the diffusion limit and the solution is heavily concentrated on the corners and edges of the simplex. In this paper we present a practical approximate solution for slow mutation rates in the form of a set of line densities along the edges of the simplex. The method of solution relies on parameterising the general non-reversible rate matrix as the sum of a reversible part and a set of $(K-1)(K-2)/2$ independent terms corresponding to fluxes of probability along closed paths around faces of the simplex. The solution is potentially a first step in estimating non-reversible evolutionary rate matrices from observed allele frequency spectra.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

The rapidly reducing cost of high throughput sequencing now allows for the acquisition of genome-wide data for detecting nucleotide allele frequencies extracted from multiple alignments within a population across large numbers of genomic sites (Pool et al., 2010). The existence of such data raises the possibility of estimating not only specific mutation rates, but complete evolutionary rate matrices from the current observed state of allele frequencies with the genome.

In a recent paper Vogl (2014) has developed a general algorithm and, in the limit of slow scaled mutation rates, a maximum likelihood estimate, of the two parameters defining the scaled instantaneous rate matrix for the case of bi-allelic neutral evolution. The estimator is similar in style to Watterson's estimator for the infinite allele case (Watterson, 1975), and assumes the data to consist of a site-frequency spectrum (or allele-frequency spectrum) obtained from genotyping a finite number of individuals at a relatively large number of independent sites whose evolution is subject only to genetic drift and identical-rate mutations. It is derived by assuming the data has a beta-binomial distribution as a result of being sampled from the well-known beta-distribution solution to the diffusion limit of the neutral Wright–Fisher model (Wright, 1931). The method is extended to include selection and the analysis of the low mutation rate limit developed further by Vogl and Bergman (2015).

A necessary first step in generalising the Vogl estimator to the multi-allele case, and in particular to the 4-allele case relevant to genomic rate matrices, is the generalisation of Wright's stationary beta distribution to higher dimensions. This involves finding a stationary solution to the multi-allelic forward Kolmogorov equation (see Eq. (3) in the next section). There is no known general solution to this partial differential equation for an arbitrary instantaneous rate matrix.

However, physical mutation rates are extremely slow on the scale relevant to the diffusion limit, and therefore we argue that for practical purposes it is not necessary to solve the forward Kolmogorov equation in its entirety over the full volume of the 3-dimensional simplex on which its solution is defined. Consider for instance the numerical stationary solution to the discrete Wright–Fisher defined by Eqs. (1) to (2), shown in Fig. 1. For the purposes of illustration we have simulated this solution using the popular Hasegawa–Kishino–Yano matrix (HKY85) (Hasegawa et al., 1985) with a small population in order to render the simulation numerically tractable, and mutation rates which are unrealistically high by at least two orders of magnitude to enable

* Corresponding author at: Mathematical Sciences Institute, Australian National University, Canberra, Australia.
E-mail addresses: conrad.burden@anu.edu.au (C.J. Burden), yurong.tang@anu.edu.au (Y. Tang).

http://dx.doi.org/10.1016/j.tpb.2016.07.005
0040-5809/© 2016 Elsevier Inc. All rights reserved.

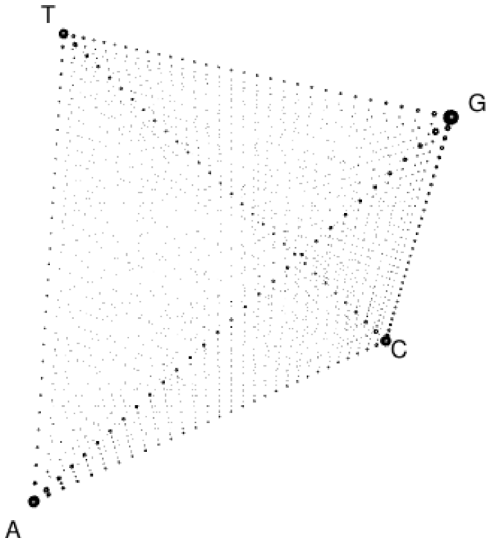

Fig. 1. Stationary distribution of allele frequencies for the HKY85 model for a haploid population of size $N = 30$ with parameters $\alpha = 0.2$, $\beta = 0.1$, $\pi_A = \pi_T = 0.2$ and $\pi_C = \pi_G = 0.3$, using the parameterisation defined in Hasegawa et al. (1985). The corners labelled A, C, G and T correspond to allele frequencies $\mathbf{i} = (N, 0, 0, 0)$, $(0, N, 0, 0)$, $(0, 0, N, 0)$ and $(0, 0, 0, N)$ respectively, and the volume of the sphere at each coordinate point is proportional to the probability mass function.

the distribution to be visible over the entire simplex on the scale of the plot.

The distribution is clearly dominated by the corners of the tetrahedron, indicating that the majority of genomic sites are not polymorphisms (SNPs). This effect is explained in Vogl and Bergman (2015) in the context of the 2-allele Moran model as a strong dominance of genetic drift over mutations for polymorphic sites. Most of the remaining support of the distribution lies on the edges of the tetrahedron, which correspond to 2-allele SNPs. The interiors of the four faces, corresponding to 3-allele SNPs, and the interior volume of the tetrahedron, corresponding to 4-allele SNPs, account for only a small fraction of the total probability. Consistent with observation of the human genome (Hodgkinson and Eyre-Walker, 2010; Cao et al., 2015; Phillips et al., 2015), the multi-allele neutral Wright–Fisher model predicts that 3- and 4-allele SNPs are extremely rare when scaled mutation rates are low. In fact, when tri-allelic SNPs are observed, the least frequent allele is generally observed in only 1% or 2% of the population (see Table S1 of Hodgkinson and Eyre-Walker (2010)), corresponding to points very close to an edge of the tetrahedron.

Zeng (2010) has demonstrated that it is feasible to estimate mutation rates and selection parameters from site-frequency data via numerical solution of the multi-allelic discrete Wright–Fisher model by assuming the stationary distribution to be restricted to the corners and edges of the simplicial lattice. However an analytic solution to the continuum diffusion limit would facilitate far more computationally efficient maximum likelihood estimate of parameters, and also provide physical insight into the dynamics of mutation.

Below we present an approximate analytic solution to the neutrally evolving multi-allelic forward Kolmogorov equation in the form of a set of line densities defined on the edges of the solution simplex for the general case of K alleles. The basis of our solution is a novel parameterisation of the most general form of the instantaneous rate matrix Q. The parameterisation consists of writing Q as the sum of a time-reversible part (Tavaré, 1986) plus a non-reversible part parametrised by $(K-1)(K-2)/2$ 'probability fluxes' corresponding to a set of independent closed triangular paths following edges of the solution simplex. Note that in this paper the terms "reversible" and "non-reversible" are used in the sense of a continuous-time Markov chain. A non-reversible rate matrix allows mutations among all states, but the system is not in detailed balance, that is to say the mutation rate from allele-1 to allele-2 need not balance the mutation rate from allele-2 to allele-1 at equilibrium. The assumption that rate matrices are reversible is popular in the phylogenetics literature because the pulley principle (Felsenstein, 1981) simplifies calculations. However there is no biochemical justification for this assumption. We find that in the limit of low mutation rates, and if neutral evolution is assumed, asymmetry in the allele frequency spectrum along edges of the solution simplex can only be explained by the non-reversible part of Q. Equivalently, if Q is reversible, the allele frequency spectrum is symmetric along each edge.

The structure of this paper is as follows. Section 2 contains a review of the multi-allelic neutral Wright–Fisher model and sets out the statement of the problem. Section 3 reviews the $K = 2$ solution to the forward Kolmogorov equation with a focus on non-standard boundary conditions. Sections 4–6 contain our approximate solutions for the $K = 3$, 4 and arbitrary K cases respectively. Section 7 discusses the strand-symmetric case. Conclusions are summarised in Section 8. Appendix A is devoted to deriving the asymptotic behaviour of the solution to Eq. (3) near the simplex boundary in the limit of low mutation rates. Appendix B is devoted to technical details of obtaining marginal distributions of the stationary K-allele solution in terms of effective 2-allele models.

2. Review of the multi-allelic neutral Wright–Fisher model

We consider the neutral evolution Wright–Fisher model for K alleles, labelled $A_1 \ldots A_K$ (see, for example, Section 4.1 of Etheridge (2011)). Given a haploid population of size N (or diploid population of size $N/2$), let the number of individuals of type A_a at time step τ be $Y_a(\tau)$ for discrete times $\tau = 0, 1, 2, \ldots$. Also, let u_{ab} be the probability of an individual making a transition from A_a to A_b in a single time step, where $u_{ab} \geq 0$ and $\sum_{b=1}^{K} u_{ab} = 1$. Writing $\mathbf{Y}(\tau) = (Y_1(\tau), \ldots, Y_K(\tau))$, the multi-allele neutral Wright–Fisher model is defined by the transition matrix from an allele frequency $\mathbf{i} = (i_1, \ldots, i_K)$ to an allele frequency $\mathbf{j} = (j_1, \ldots, j_K)$ in the population given by the multinomial distribution

$$\text{Prob}\left(\mathbf{Y}(\tau+1) = \mathbf{j} | \mathbf{Y}(\tau) = \mathbf{i}\right) = \frac{N!}{\prod_{a=1}^{K} j_a!} \prod_{a=1}^{K} \left(\sum_{b=1}^{K} \frac{i_b}{N} u_{ba} \right)^{j_a}. \tag{1}$$

This transition matrix defines a finite state Markov chain with a state space of dimension $\binom{N+K-1}{K-1}$.

The usual diffusion limit is obtained by defining random variables $X_a(t) = Y_a(\tau)/N$ equal to the relative proportion of type-A_a alleles within the population at continuous time $t = \tau/N$. The limit $N \to \infty$ and $u_{ab} \to 0$ for $a \neq b$ is taken in such a way that the $K \times K$ instantaneous rate matrix Q, whose elements are defined by

$$Q_{ab} = N(u_{ab} - \delta_{ab}), \tag{2}$$

remains finite. Here δ_{ab} is the Kronecker delta, equal to 1 if $a = b$ and 0 otherwise. This limit gives the forward Kolmogorov equation

$$\frac{\partial f}{\partial t} = -\sum_{a=1}^{K-1} \frac{\partial}{\partial x_a} \sum_{b=1}^{K} x_b Q_{ba} f + \frac{1}{2} \sum_{a,b=1}^{K-1} \frac{\partial^2}{\partial x_a \partial x_b} \left\{ (\delta_{ab} x_a - x_a x_b) f \right\}, \tag{3}$$

Download English Version:

https://daneshyari.com/en/article/6372281

Download Persian Version:

https://daneshyari.com/article/6372281

Daneshyari.com