Theoretical Population Biology 114 (2017) 88–94

Contents lists available at ScienceDirect

Theoretical Population Biology

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

An alternative derivation of the stationary distribution of the multivariate neutral Wright–Fisher model for low mutation rates with a view to mutation rate estimation from site frequency data

Dominik Schrempf[a,b,*], Asger Hobolth[c]

[a] Institut für Populationsgenetik, Vetmeduni Vienna, Austria
[b] Vienna Graduate School of Population Genetics, Austria
[c] Bioinformatics Research Center, Aarhus University, Denmark

ARTICLE INFO

Article history:
Received 4 July 2016
Available online 29 December 2016

Keywords:
Wright–Fisher model
Moran model
Diffusion equation
Stationary distribution
Boundary mutation model

ABSTRACT

Recently, Burden and Tang (2016) provided an analytical expression for the stationary distribution of the multivariate neutral Wright–Fisher model with low mutation rates. In this paper we present a simple, alternative derivation that illustrates the approximation. Our proof is based on the discrete multivariate boundary mutation model which has three key ingredients. First, the decoupled Moran model is used to describe genetic drift. Second, low mutation rates are assumed by limiting mutations to monomorphic states. Third, the mutation rate matrix is separated into a time-reversible part and a flux part, as suggested by Burden and Tang (2016). An application of our result to data from several great apes reveals that the assumption of stationarity may be inadequate or that other evolutionary forces like selection or biased gene conversion are acting. Furthermore we find that the model with a reversible mutation rate matrix provides a reasonably good fit to the data compared to the one with a non-reversible mutation rate matrix.

1. Introduction

The incredible amount of available genome-wide sequence data (e.g., Mackay et al., 2012; Auton et al., 2015) drives the development of fast methods to infer evolutionary rate matrices. The Wright–Fisher model is the standard model in population genetics (Fisher, 1930; Wright, 1931). However, a full analytical or numerical treatment is usually intractable, especially when the population size is large. In this case, allele frequencies become semi-continuous quantities and may be modeled in terms of a diffusion process (Kimura, 1964; Ewens, 2004; Durrett, 2008). Wright was the first to investigate the solution for two alleles at stationarity, i.e., after the process has evolved for a very long time (Wright, 1931). Even if stationarity is assumed, mathematical limitations inhibit an analytical treatment of the multivariate case for general rate matrices (reviewed in Griffiths and Spanó, 2010), albeit solutions for parent independent mutation models exist (Griffiths, 1979).

The forthcoming arguments require the establishment of some notation. Consider a haploid population of constant size N and a single locus with K alleles $\{1, \dots, K\}$. The evolution of this population in the course of time can be described by a discrete-time Markov chain with discrete character-space; a lattice of $\binom{N+K-1}{K-1}$ points, each of which represents a specific assortment of the alleles within the population. The trajectory of a population within this lattice in space–time can be identified with a row vector of the number of alleles $\boldsymbol{z}(\tau) = (z_1(\tau), \dots, z_K(\tau))$ in generation τ. It only has $K-1$ independent elements but for convenience all K elements are kept together. Mutations are modeled by a time-homogeneous, $K \times K$ mutation probability matrix $\boldsymbol{U}$ with probabilities u_{ab} for a mutation from allele a to allele b in a single generation. Here and throughout this document a and b denote any of the alleles $\{1, \dots, K\}$. The diagonal elements $u_{aa} = 1 - \sum_{b:\, b\neq a} u_{ab}$ are the probabilities that allele a does not mutate.

The neutral, K-allelic Wright–Fisher model derives the distribution of the number of alleles in the next generation by sampling with replacement from the alleles in the present after mutation has taken place. In particular (e.g., Ewens, 2004),

$$\boldsymbol{z}(\tau+1) | \boldsymbol{z}(\tau) \sim \text{Mult}\left(N, \frac{\boldsymbol{z}(\tau)}{N}\boldsymbol{U}\right), \tag{1}$$

where $\text{Mult}(N, \boldsymbol{v})$ is the multinomial distribution with N draws and probability vector $\boldsymbol{v}$. An example of a numerically obtained stationary distribution of this process for $K = 4$, $N = 30$ and moderately large mutation rates (Table 1) is illustrated in Fig. 1. The parameter values are taken from Fig. 6 in Burden and Tang (2016),

* Corresponding author at: Institut für Populationsgenetik, Vetmeduni Vienna, Austria.
E-mail address: dominik.schrempf@vetmeduni.ac.at (D. Schrempf).

http://dx.doi.org/10.1016/j.tpb.2016.12.001

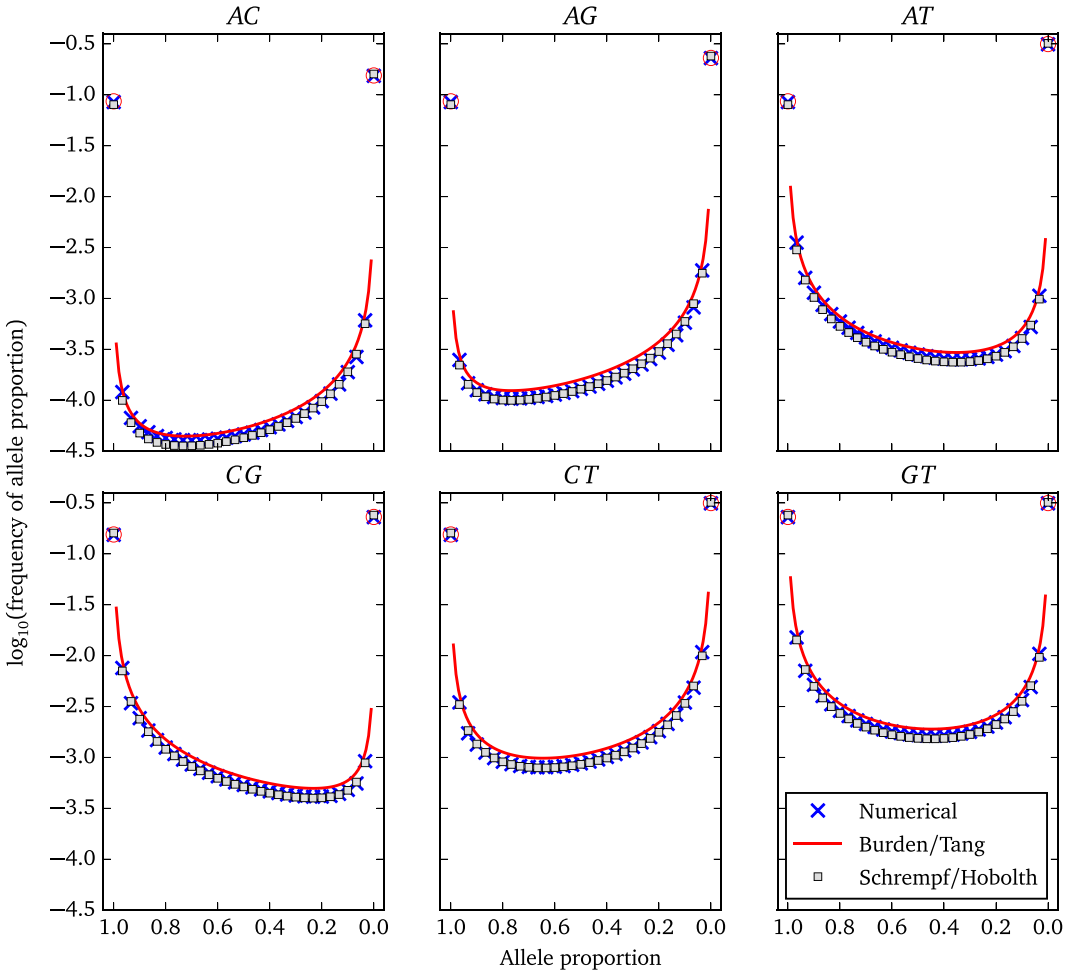

Fig. 1. Stationary distribution of the neutral Wright–Fisher model for $K = 4$ (nucleotides A, C, G and T), $N = 30$ and moderately large mutation rates (Table 1). Each subplot illustrates the distribution of allele frequencies when only two alleles are present (the x-axes denote the proportion of the first allele). Blue crosses are the results from a numerical Wright–Fisher simulation with full $\binom{N+K-1}{K-1} \times \binom{N+K-1}{K-1}$ transition matrix; i.e., small, non-zero probabilities corresponding to tri-allelic and tetra-allelic sites are omitted. The continuous approximation of Burden and Tang (2016) is in red. Gray squares denote the stationary distribution proposed in this manuscript.

Table 1
Parameters of the simulation (Fig. 1) and estimations with reversible and general mutation rate matrices for the great apes data (Fig. 4). The mutation rate matrix is $\boldsymbol{Q} = \boldsymbol{Q}^{\text{GTR}} + \boldsymbol{Q}^{\text{flux}}$, where $\boldsymbol{Q}^{\text{GTR}}$ and $\boldsymbol{Q}^{\text{flux}}$ are determined by Eqs. (6) and (7), respectively. The relation between $\boldsymbol{Q}$ and the mutation probability matrix $\boldsymbol{U}$ is given in Eq. (5). Note, that the mutation probabilities of the Moran model are twice as large as the corresponding mutation probabilities of the Wright–Fisher model. The order of ab is AC, AG, AT, CG, CT, GT.

	$(\pi_A, \pi_C, \pi_G, \pi_T)$	(r_{ab})	(ϕ_{ab})
Simulation	(0.1, 0.2, 0.3, 0.4)	$0.02 \cdot (1, 2, 3, 4, 5, 6)$	$0.02 \cdot (-0.75, -1.66, 1.625, 3.33, -2.6875, 1.25)$
Reversible	(0.22, 0.30, 0.24, 0.24)	$10^{-4} \cdot (4.0, 44, 2.2, 4.7, 27, 4.3)$	
General	(0.22, 0.30, 0.24, 0.24)	$10^{-4} \cdot (5.6, 42, 2.2, 4.7, 25, 5.5)$	$10^{-5} \cdot (29, -39, 2.2, 6.3, 19, -26)$

who provide an approximation for low mutation rates using diffusion limit arguments.

Here, we present a simple and intuitive derivation of the stationary distribution with adjusted normalization. We employ an approximation, which is nor of numerical nor of analytical type but alters the underlying model itself assuming low mutation rates (De Maio et al., 2015). In particular, we use the decoupled Moran model (Baake and Bialowons, 2008; Etheridge and Griffiths, 2009), which is mathematically most convenient, and limit mutations to monomorphic states. An application of our results to population data from great apes (Prado-Martinez et al., 2013) indicates that the assumption of stationarity is inadequate or that other evolutionary forces like selection or biased gene conversion are acting. Furthermore we find that the model with a reversible mutation rate matrix provides a reasonably good fit to the data compared to the one with a non-reversible mutation rate matrix.

2. Discrete multivariate boundary mutation model

2.1. Moran model

In the neutral, continuous-time Moran model with mutations, the rate from state $\boldsymbol{z}$ to $(\ldots, z_a - 1, \ldots, z_b + 1, \ldots)$ is (e.g., Durrett, 2008)

$$\frac{z_a z_b}{N} + z_a u_{ab}. \tag{2}$$

Importantly, the dynamics (and therefore also the stationary distribution) of the diffusion approximation of the Moran model with doubled mutation rates and the Wright–Fisher model have been shown to be equal (e.g., Wakeley, 2009, p. 58, and Durrett, 2008, Section 1.5).