Bioinformatics, 32(5), 2016, 795–797
doi: 10.1093/bioinformatics/btv627
Advance Access Publication Date: 9 November 2015
Applications Note

Genetics and population analysis

SpectralTDF: transition densities of diffusion processes with time-varying selection parameters, mutation rates and effective population sizes

Matthias Steinrücken[1,†], Ethan M. Jewett[2,†] and Yun S. Song[2,3,4,5,6,*]

[1]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA 01003, USA, [2]Department of Statistics, [3]Department of EECS, [4]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA, [5]Department of Mathematics and [6]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Associate Editor: Janet Kelso

Received on April 30, 2015; revised on October 5, 2015; accepted on October 22, 2015

Abstract

Motivation: In the Wright–Fisher diffusion, the transition density function describes the time evolution of the population-wide frequency of an allele. This function has several practical applications in population genetics and computing it for biologically realistic scenarios with selection and demography is an important problem.
Results: We develop an efficient method for finding a spectral representation of the transition density function for a general model where the effective population size, selection coefficients and mutation parameters vary over time in a piecewise constant manner.
Availability and implementation: The method, called SpectralTDF, is available at https://sourceforge.net/projects/spectraltdf/.
Contact: yss@berkeley.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The transition density function (TDF) of the Wright–Fisher diffusion describes the time evolution of the frequency of an allele (Ewens, 2004). The TDF is useful for understanding the effects of demography, mutation and selection on genetic variation, and it is a key component of a number of methods for inferring selection coefficients (Bollback *et al.*, 2008; Steinrücken *et al.*, 2014; Williamson *et al.*, 2004), predicting allele fixation times (Waxman, 2011) and computing population genetic statistics such as the site frequency spectrum (Živković *et al.*, 2015).

Most existing approaches for computing the TDF assume either restrictive models of dominance (Kimura, 1955, 1957) or selective neutrality (Griffiths, 1979; Shimakura, 1977; Vogl, 2014) or are computationally slow for selection strengths commonly observed in biological data (Barbour *et al.*, 2000). However, Song and Steinrücken (2012) and Steinrücken *et al.* (2013) recently developed a numerically stable and computationally efficient method for finding a spectral representation of the TDF for a general selection model in the case of constant parameters (population size, mutation rates and selection coefficients). Despite the utility of this new approach, assuming that model parameters remain constant over time is often too restrictive for biological applications (Siepielski *et al.*, 2009).

Živković *et al.* (2015) have extended the spectral method of Song and Steinrücken (2012) to handle piecewise-constant population size functions. However, their approach requires a restricted model of selection in which the fitness of a homozygote is twice that of a heterozygote (i.e. additive or genic selection). Furthermore, selection parameters are assumed to remain constant over time, and the model does not allow for recurrent mutations.

Here, we present the first method for computing the TDF under arbitrary models of dominance and recurrent mutation while allowing selection parameters, mutation rates and effective population sizes to change over time in a piecewise constant manner.

2 Approach

We consider a biallelic locus with two alleles, A_0 and A_1, evolving in a single panmictic population. In the corresponding Wright–Fisher diffusion, X_t denotes the frequency of allele A_1 at time t, measured continuously in units of generations. We assume that either X_0 is given or the distribution of X_0 is specified. The effective population size, mutation rates and selection parameters are assumed to be constant within each of K disjoint epochs. As illustrated in Figure 1, the kth epoch has effective size N_k (diploid individuals) and duration τ_k. Epoch boundaries are denoted by $t_0, t_1, \dots, t_K$, with $t_k = \sum_{i=1}^{k} \tau_i$.

Within the kth epoch, the per-generation probability that a copy of allele A_0 mutates to allele A_1 is a_k, and the per-generation probability that a copy of allele A_1 mutates to allele A_0 is b_k. In addition, selection acts in such a way that the relative fitness of an individual carrying i copies of allele A_1 is $1 + s_{ki}$ $(i = 1, 2)$.

The TDF $p_k(t; x, y)$ in epoch k is defined by $p_k(t; x, y)dy = \mathbb{P}(y \le X_{t_{k-1}+t} < y + dy | X_{t_{k-1}} = x)$, where $t_{k-1} + t < t_k$. The TDF $p_k(t; x, y)$ satisfies the partial differential equation $\partial p_k(t; x, y)/\partial t = \mathcal{L}_k p_k(t; x, y)/2N_k$, where $\mathcal{L}_k$ is the diffusion generator given by

$$\begin{aligned}\mathcal{L}_k = &\frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2} + \frac{1}{2}[\alpha_k - (\alpha_k + \beta_k)x]\frac{\partial}{\partial x} \\ &+ 2x(1-x)[\sigma_{k1}(1-2x) + \sigma_{k2}x]\frac{\partial}{\partial x}.\end{aligned} \tag{1}$$

See Song and Steinrücken (2012) for discussion on the appropriate boundary conditions. In Equation (1), the parameters $\alpha_k = 4N_k a_k$, $\beta_k = 4N_k b_k$, $\sigma_{k1} = N_k s_{k1}$ and $\sigma_{k2} = N_k s_{k2}$ are the population-scaled versions of the mutation and selection parameters.

Within each epoch, k, a spectral representation of the TDF $p_k(t; x, y)$ can be obtained by employing the framework of Song and Steinrücken (2012), who developed an efficient algorithm for finding the eigenvalues and the eigenfunctions of the diffusion generator $\mathcal{L}_k$. The challenge in computing the TDF for the full model with K epochs lies in knitting together the expressions for the densities $p_k(t; x, y)$ across the different epochs. The method we implement involves an efficient and numerically stable algorithm for carrying out this knitting procedure using a polynomial interpolation method, which is detailed in Supplementary Methods.

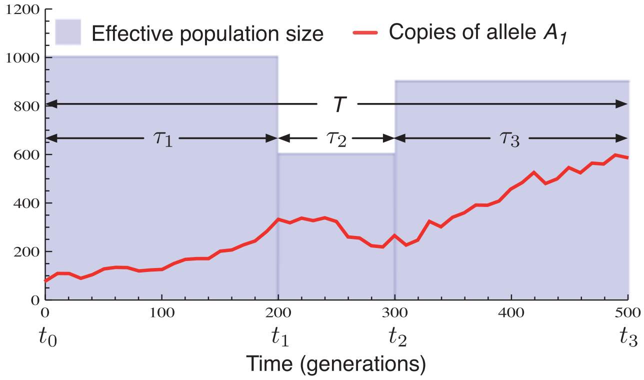

Fig. 1. Diagram of the model. A population has constant size in each of K epochs ($N_1 = 1000$, $N_2 = 600$, $N_3 = 900$). An allele, A_1, at a locus of interest evolves over time, subject to pressures of mutation and selection that are constant within each epoch

3 Implementation

Our algorithm has been implemented in JAVA. The inputs to the program are the effective population sizes (number of diploid individuals) $\boldsymbol{N} = (N_1, \dots, N_K)$; epoch durations $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$; per-generation mutation rates $\boldsymbol{a} = (a_1, \dots, a_K)$ and $\boldsymbol{b} = (b_1, \dots, b_K)$; selection parameters $\boldsymbol{s}_1 = (s_{11}, \dots, s_{K1})$ and $\boldsymbol{s}_2 = (s_{12}, \dots, s_{K2})$; initial allele frequency X_0 and the time $t \in [0, T]$ at which the TDF will be evaluated. A plot of the TDF evaluated at each epoch boundary point ($t = \tau_1, \tau_1 + \tau_2$ and T) in Figure 1 is shown in Figure 2. The full command options are detailed in the user manual distributed with the software.

4 Discussion

Our implementation provides a fast and numerically stable method for computing the TDF for a general model with piecewise-constant population sizes and a broad range of time-varying mutation and selection parameters. It also allows for a variety of initial conditions, including a specified initial frequency and stationary distributions under mutation-selection balance or mutation-drift balance.

The JAVA implementation is designed to be used either as a stand-alone application or in combination with other methods. For example, the code can be easily incorporated into the method of Steinrücken *et al.* (2014), allowing the inference of selection parameters from time series data sampled from populations with time-varying demographic and selection parameters. In general, the method we present provides a flexible and efficient tool for studying the evolution of allele frequencies over time under complex evolutionary scenarios.

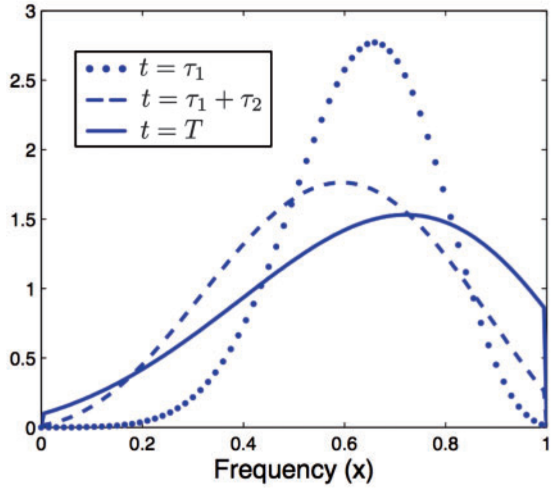

Fig. 2. Plot of the TDF for the model shown in Figure 1 with the parameters specified in the example in Section 3, evaluated at the times t_1, t_2 and T

Acknowledgement

We thank Daniel Živković and Anand Bhaskar for helpful discussions and collaboration on earlier related work.

Funding

This research was supported, in part, by NIH grants R01-GM094402 (M.S. and Y.S.S.) and R01-GM109454 (E.M.J.), and by a Packard Fellowship for Science and Engineering (Y.S.S.).

Conflict of Interest: none declared.

References

Barbour,A. *et al*. (2000) A transition function expansion for a diffusion model with selection. *Ann. Appl. Probability*, **10**, 123–162.

Bollback,J. *et al*. (2008) Estimation of $2N_es$ from temporal allele frequency data. *Genetics*, **179**, 497–502.

Ewens,W. (2004) *Mathematical Population Genetics: I*, 2nd edn. Springer-Verlag, New York.

Griffiths,R. (1979) A transition density expansion for a multi-allele diffusion model. *Adv. Appl. Probability*, **11**, 310–325.

Kimura,M. (1955) Stochastic processes and distribution of gene frequencies under natural selection. In: Warren,K.B. (ed), *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. 20. Cold Spring Harbor Laboratory Press, Baltimore, MD, pp. 33–53.

Kimura,M. (1957) Some problems of stochastic processes in genetics. *Ann. Math. Stat.*, **28**, 882–901.

Shimakura,N. (1977) Equations différentielles provenant de la génétique des populations. *Tohoku Math. J. Second Ser.*, **29**, 287–318.

Siepielski,A. *et al*. (2009) Its about time: the temporal dynamics of phenotypic selection in the wild. *Ecol. Lett.*, **12**, 1261–1276.

Song,Y.S. and Steinrücken,M. (2012) A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics*, **190**, 1117–1129.

Steinrücken,M. *et al*. (2013) An explicit transition density expansion for a multi-allelic Wright-Fisher diffusion with general diploid selection. *Theor. Popul. Biol.*, **83**, 1–14.

Steinrücken,M. *et al*. (2014) A novel spectral method for inferring general diploid selection from time series genetic data. *Ann. Appl. Stat.*, **8**, 2203–2222.

Vogl,C. (2014) Biallelic mutation-drift diffusion in the limit of small scaled mutation rates. *arXiv*, 1409.2299.

Waxman,D. (2011) A unified treatment of the probability of fixation when population size and the strength of selection change over time. *Genetics*, **188**, 907–913.

Williamson,S. *et al*. (2004) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA*, **102**, 7882–7887.

Živković,D. *et al*. (2015) Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics*, **200**, 601–617.