Theoretical Population Biology 75 (2009) 320–330

Contents lists available at ScienceDirect

Theoretical Population Biology

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

A coalescent dual process in a Moran model with genic selection

A.M. Etheridge, R.C. Griffiths *

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK

ARTICLE INFO

Article history:
Received 31 December 2008
Available online 31 March 2009

Keywords:
Ancestral Selection Graph
Coalescent process
Genic selection
Harmonic measure
Moran model
Wright–Fisher diffusion

ABSTRACT

A coalescent dual process for a multi-type Moran model with genic selection is derived using a generator approach. This leads to an expansion of the transition functions in the Moran model and the Wright–Fisher diffusion process limit in terms of the transition functions for the coalescent dual. A graphical representation of the Moran model (in the spirit of Harris) identifies the dual as a strong dual process following typed lines backwards in time. An application is made to the harmonic measure problem of finding the joint probability distribution of the time to the first loss of an allele from the population and the distribution of the surviving alleles at the time of loss. Our dual process mirrors the Ancestral Selection Graph of [Krone, S. M., Neuhauser, C., 1997. Ancestral processes with selection. Theoret. Popul. Biol. 51, 210–237; Neuhauser, C., Krone, S. M., 1997. The genealogy of samples in models with selection. Genetics 145, 519–534], which allows one to reconstruct the genealogy of a random sample from a population subject to genic selection. In our setting, we follow [Stephens, M., Donnelly, P., 2002. Ancestral inference in population genetics models with selection. Aust. N. Z. J. Stat. 45, 395–430] in assuming that the types of individuals in the sample are known. There are also close links to [Fearnhead, P., 2002. The common ancestor at a nonneutral locus. J. Appl. Probab. 39, 38–54]. However, our methods and applications are quite different. This work can also be thought of as extending a dual process construction in a Wright–Fisher diffusion in [Barbour, A.D., Ethier, S.N., Griffiths, R.C., 2000. A transition function expansion for a diffusion model with selection. Ann. Appl. Probab. 10, 123–162]. The application to the harmonic measure problem extends a construction provided in the setting of a neutral diffusion process model in [Ethier, S.N., Griffiths, R.C., 1991. Harmonic measure for random genetic drift. In: Pinsky, M.A. (Ed.), Diffusion Processes and Related Problems in Analysis, vol. 1. In: Progress in Probability Series, vol. 22, Birkhäuser, Boston, pp. 73–81].

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Consider a population in which each individual is labelled according to a type taken from the set $[d] = \{1, 2, \ldots, d\}$. We write $X_j(t)$ for the frequency of type j individuals in the population at time $t \geq 0$ and $X(t) = \left(X_j(t)\right)_{j\in[d]}$ and model the evolution of the population using a multi-type Wright–Fisher diffusion process. In the simplest setting, there is no selection, and mutation between types is parent independent. That is, each individual mutates to type j at rate $\theta_j/2 \geq 0$, *independent* of its current type. The generator of the diffusion process is then

$$\mathcal{L} = \frac{1}{2}\sum_{i,j\in[d]} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j} + \frac{1}{2}\sum_{i\in[d]}(\theta_i - |\theta|x_i)\frac{\partial}{\partial x_i}, \tag{1}$$

where we use the notation $|a| = \sum_j a_j$ for the sum of elements in a vector. If $\theta > 0$ (meaning that $\theta_j > 0$ for all $j \in [d]$) the Wright–Fisher diffusion has a Dirichlet stationary distribution

$$\mathcal{D}(x, \theta) = \frac{\Gamma(|\theta|)}{\Gamma(\theta_1)\cdots\Gamma(\theta_d)} x_1^{\theta_1 - 1}\cdots x_d^{\theta_d - 1}$$

for $x > 0$ and $|x| = 1$. More generally, one can allow some of the θ_j to vanish, in which case we obtain a generalized Dirichlet distribution in which the corresponding frequencies x_j vanish with probability one. Ethier and Griffiths (1993) show that the transition distribution of the diffusion can be expressed as a mixture,

$$\mathbb{P}(t, x, \cdot) = \sum_{k=0}^{\infty} q_k^{|\theta|}(t) \sum_{|l|=k} \mathcal{M}(l; k, x)\mathcal{D}(\cdot, \theta + l), \tag{2}$$

where $\mathcal{M}(l; k, x)$ denotes the multinomial distribution,

$$\mathcal{M}(l; k, x) = \binom{k}{l} x_1^{l_1}\cdots x_d^{l_d}, \quad |l| = k,$$

and $q_k^{|\theta|}(t)$ are the transition functions of a (dual) pure death process which we denote by $\{L(t), t \geq 0\}$. This process should be thought of as evolving in backwards time. Lineages are lost through coalescence, through which $k \to k - 1$ at rate $k(k - 1)/2$, and

* Corresponding author.
E-mail addresses: etheridge@stats.ox.ac.uk (A.M. Etheridge), griff@stats.ox.ac.uk (R.C. Griffiths).

0040-5809/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2009.03.004

mutation, resulting in $k \to k-1$ at an additional rate $k|\theta|/2$. We suppose that $L(t)$ starts from infinity (although it will be finite at any $t > 0$). If $|\theta| = 0$ this dual process is the number of blocks in the famous Kingman coalescent (Kingman, 1982). The expansion (2) still holds in this case, except that now we will have $L(t) \geq 1$ for all $t \geq 0$ and so the summation is over $k \geq 1$. There is an explicit expression for the transition functions,

$$q_k^{|\theta|}(t) = \sum_{j=k}^{\infty} \rho_j^{|\theta|}(t)(-1)^{j-k} \frac{(2j+|\theta|-1)(j+|\theta|)_{(k-1)}}{j!(j-k)!},$$

where $\rho_j^{|\theta|}(t) = e^{-j(j+|\theta|-1)t/2}$, (Griffiths, 1980; Tavaré, 1984; Griffiths, 2006). To understand the expansion (2), one can think of the infinite number of individuals that make up $L(0)$ as the 'leaves' in a forest of trees. Each tree either grows from a 'founder' at time t (which corresponds to time zero in the diffusion process) or its root arose through a new mutation. This subdivides the leaves into 'families' and leads to the Dirichlet mixture. If there are k founder lineages, then their types are determined by sampling k individuals from the diffusion at time zero, and hence the probability that the numbers of founder lineages of types $1, \ldots, d$ are given by $l = (l_1, \ldots, l_d)$ with $|l| = k$ is just $\mathcal{M}(l; k, x)$. Let $U = (U_1, \ldots, U_k)$ be the relative family sizes of these founder families in the leaves of the tree, and $V = (V_1, \ldots, V_d)$ be the frequencies of families derived from new mutations on the tree edges in $(0, t)$. Then $U \oplus V = (U_1, \ldots, U_k, V_1, \ldots, V_d)$ has a $\mathcal{D}(u \oplus v, (1, \ldots, 1) \oplus \theta)$ distribution. The term $\mathcal{D}(\cdot, \theta + l)$ in (2) is obtained by combining families of individuals of the same type, corresponding to adding the parameters in the Dirichlet distribution. If $\theta = 0$ the process is one of pure random drift. There is an analogous mixture representation for the transition function of a Fleming–Viot process. In that setting the finite set $[d]$ is replaced by an infinite type space. At each time t, $X(t)$ is now a probability measure on the type space and the Dirichlet distribution is replaced by a Poisson–Dirichlet distribution (Ethier and Griffiths, 1993). This is a canonical representation because the d-type diffusion can be obtained by taking the measure that determines the type after mutation in the Fleming–Viot process to be atomic with atoms $\{\theta_j/|\theta|; \theta_j > 0, j \in [d]\}$. The transition distribution (2) first appeared in Griffiths and Li (1983) and Tavaré (1984) with an interpretation based on Griffiths (1980) lines of descent. Donnelly and Tavaré (1987) discuss (2) and give a probabilistic explanation. Watterson (1984) derived an analogous representation for the distribution of old and new allele families in a neutral Moran model.

Of course one can obtain the multi-type Wright–Fisher diffusion with selection as an infinite population limit of Moran models (with weak selection). For a population of size N and types from $[d]$, each individual of type i gives birth at rate $\lambda_i = \lambda(1 + \sigma_i/N)$ (to an offspring of the same type) and an individual is selected at random from the population to die (thus maintaining constant population size). It is convenient to suppose that the constants $\{\sigma_i\}_{i\in[d]}$ are negative and we write σ for $\max\{\sigma_i - \sigma_j : i, j \in [d]\}$. In addition, mutation changes an individual of type i to an individual of type j at rate μp_{ij}. If we take $\lambda = N/2$ and let $N \to \infty$ we recover the multi-type Wright–Fisher diffusion. Krone and Neuhauser (1997) and Neuhauser and Krone (1997) exploited the graphical representation of a Moran model with selection (which can be thought of as a biased voter model on a complete graph) to write down the Ancestral Selection Graph (ASG). This graph is traced out by a branching–coalescing system of lineages and has embedded within it the genealogy of a random sample from the population. Passing to a weak selection limit they obtain the same duality in the diffusion setting. In that limit, if there are currently j edges in the graph then $j \to j-1$ (through coalescence) at rate $\binom{j}{2}$ and $j \to j+1$ at rate $\sigma j/2$. These branching events correspond to 'potential selective events'. In order to extract the genealogy of a sample of size n, one starts with n edges in the graph and traces back until the first (almost surely finite) time when there is only one edge. This is the 'ultimate ancestor'. The type of the ultimate ancestor is chosen (by sampling from the population at that time) and then one works back through the graph using the rule that 'the fitter type always wins' whenever one arrives at a point corresponding to a branching event. This allows us to prune the graph to recover the genealogical tree (and the types in the sample). Mano (in press) found an explicit, though complicated, expression for the transition functions of the number of ancestors in the ASG as we trace backwards in time by considering the ASG as a moment dual in a two-allele Wright–Fisher diffusion process with selection and without mutation. Donnelly and Kurtz (1999) use their 'modified lookdown' approach to construct, simultaneously, the Fleming–Viot process with selection and the ancestral selection graph that encodes the genealogy. Stephens and Donnelly (2002) and Fearnhead (2002) consider the case when the types of individual in the sample are known and construct an ASG with 'typed' lines. Stephens and Donnelly (2002) deal with general diploid selection (in an infinite population limit) whereas Fearnhead (2002) considers the genic selection that interests us here. His results too are valid in the infinite population limit. Both papers deal only with parent-independent mutation. The transition rates in their typed ASGs are similar to those of (24) in this paper once we specialize to a diffusion model with genic selection and parent-independent mutation.

Barbour et al. (2000) derive a transition density expansion for a Wright–Fisher diffusion with genic selection in terms of the transition functions of a dual process. The generator of the diffusion process $\{X(t), t \geq 0\}$ with selection coefficients $\{\sigma_j \leq 0, j \in [d]\}$ is

$$\mathcal{L} = \frac{1}{2} \sum_{i,j\in[d]} x_i(\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} + \frac{1}{2} \sum_{i\in[d]} (\theta_i - |\theta| x_i + x_i(\sigma_i - s(x))) \frac{\partial}{\partial x_i}, \tag{3}$$

where $s(x) = \sum_{j\in[d]} \sigma_j x_j$. The transition distribution in the diffusion can be written as a mixture,

$$\mathbb{P}(t, x, \cdot) = \sum_{l\in\mathbb{Z}_+^d} h_{xl}(t) \mathcal{D}(\cdot, l + \theta, \sigma). \tag{4}$$

Here $\mathcal{D}(y, l + \theta, \sigma)$ is a weighted Dirichlet distribution whose density is weighted by $\exp\left(\sum_{j\in[d]} \sigma_j y_j\right)$. In general, if ξ has a Dirichlet distribution $\mathcal{D}(\cdot, \theta + l)$ we write

$$v(\theta + l) = \mathbb{E}\left[\exp\left\{\sum_{j\in[d]} \sigma_j \xi_j\right\}\right]$$

for the normalizing constant in $\mathcal{D}(\cdot, l + \theta, \sigma)$. The functions $\left\{h_{xl}(t), t \geq 0\right\}$ are transition functions of a multi-type birth and death process started from an infinite number of individuals whose types have frequencies $x = (x_1, \ldots, x_d)$. (In fact, showing that one can construct the birth–death process from this entrance boundary at infinity is rather involved.) The non-zero entries in the αth row of the Q matrix for the multi-type birth and death process are

$$q(\alpha, \alpha + e_j) = \frac{1}{2} \frac{|\sigma_j||\alpha|(\alpha_j + \theta_j)}{|\alpha| + |\theta|} \cdot \frac{v(\theta + \alpha + e_j)}{v(\theta + \alpha)}$$

$$q(\alpha, \alpha - e_j) = \frac{1}{2} \alpha_j(|\theta| + |\alpha| - 1) \cdot \frac{v(\alpha + \theta - e_j)}{v(\theta + \alpha)}$$

$$q(\alpha, \alpha) = -\frac{1}{2}\left[\sum_{j\in[d]} \alpha_j|\sigma_j| + |\alpha|(|\alpha| + |\theta| - 1)\right]. \tag{5}$$

Download English Version:

https://daneshyari.com/en/article/4502656

Download Persian Version:

https://daneshyari.com/article/4502656

Daneshyari.com