Theoretical Population Biology 112 (2016) 126–138

Contents lists available at ScienceDirect

Theoretical Population Biology

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

A coalescent dual process for a Wright–Fisher diffusion with recombination and its application to haplotype partitioning

Robert C. Griffiths [a], Paul A. Jenkins [b,c,*], Sabin Lessard [d]

[a] Department of Statistics, University of Oxford, United Kingdom
[b] Department of Statistics, University of Warwick, United Kingdom
[c] Department of Computer Science, University of Warwick, United Kingdom
[d] Département de Mathématiques et de Statistique, Université de Montréal, Montréal, Canada

ARTICLE INFO

Article history:
Received 14 April 2016
Available online 1 September 2016

Keywords:
Coalescent
Wright–Fisher diffusion
Recombination
Duality

ABSTRACT

Duality plays an important role in population genetics. It can relate results from forwards-in-time models of allele frequency evolution with those of backwards-in-time genealogical models; a well known example is the duality between the Wright–Fisher diffusion for genetic drift and its genealogical counterpart, the coalescent. There have been a number of articles extending this relationship to include other evolutionary processes such as mutation and selection, but little has been explored for models also incorporating crossover recombination. Here, we derive from first principles a new genealogical process which is dual to a Wright–Fisher diffusion model of drift, mutation, and recombination. The process is reminiscent of the *ancestral recombination graph*, a widely-used multilocus genealogical model, but here ancestral lineages are typed and transition rates are regarded as being conditioned on an observed configuration at the leaves of the genealogy. Our approach is based on expressing a putative duality relationship between two models via their infinitesimal generators, and then seeking an appropriate test function to ensure the validity of the duality equation. This approach is quite general, and we use it to find dualities for several important variants, including both a discrete L-locus model of a gene and a continuous model in which mutation and recombination events are scattered along the gene according to continuous distributions. As an application of our results, we derive a series expansion for the transition function of the diffusion. Finally, we study in further detail the case in which mutation is absent. Then the dual process describes the dispersal of ancestral genetic material across the ancestors of a sample. The stationary distribution of this process is of particular interest; we show how duality relates this distribution to haplotype fixation probabilities. We develop an efficient method for computing such probabilities in multilocus models.

1. Introduction

The concept of duality is a powerful technique for inferring the properties of one Markov process by looking at another related process, usually (as in this paper) discovered by considering the dynamics of the former in reverse time (see Jansen and Kurt, 2014, for recent review). The idea has found many applications in population genetics, playing for example a central role in the constructions of the ancestral selection graph (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997) and the ancestral influence graph (Donnelly and Kurtz, 1999). One particularly well known duality is between the Wright–Fisher diffusion describing pure genetic drift and Kingman's coalescent (Kingman, 1982). To illustrate the idea, consider a single neutral locus with two alleles. The Wright–Fisher diffusion $(X_t)_{t\geq 0}$ is the process on $[0, 1]$ describing the evolution of the frequency of one allele, with infinitesimal generator

$$\mathscr{L}f(x) = \frac{1}{2}x(1-x)f''(x) \tag{1}$$

and domain $\mathscr{D}(\mathscr{L}) = C^2([0, 1])$. The corresponding dual is the pure death process $(L_t)_{t\geq 0}$ on $\mathbb{N} = \{0, 1, \ldots\}$ with infinitesimal generator

$$\mathscr{K}f(n) = \binom{n}{2}[f(n-1) - f(n)], \tag{2}$$

which describes the dynamics of the *ancestral*, or *block-counting*, process of Kingman's coalescent.

* Correspondence to: Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom.
E-mail address: p.jenkins@warwick.ac.uk (P.A. Jenkins).

http://dx.doi.org/10.1016/j.tpb.2016.08.007
0040-5809/

The two processes are dual with respect to the function $F : [0, 1] \times \mathbb{N} \to \mathbb{R}$ defined by $F(x, n) = x^n$ (i.e. *moment duals*): for each $x \in [0, 1]$, $n \in \mathbb{N}$ and $t \geq 0$,

$$\mathbb{E}[F(X_t, n) \mid X_0 = x] = \mathbb{E}[F(x, L_t) \mid L_0 = n]. \tag{3}$$

We note for later use that this implies

$$\mathscr{L}F(\cdot, n)(x) = \mathscr{K}F(x, \cdot)(n), \quad x \in [0, 1],\ n \in \mathbb{N}, \tag{4}$$

and for general $\mathscr{L}$, $\mathscr{K}$, and F, the converse is also true under certain conditions on F (Jansen and Kurt, 2014). We also emphasise that, in this example and all others encountered in this paper, this duality is obtained via time-reversal, so that the time indices in the two processes run in different directions. Were we to run the two processes on a joint probability space, running X_t from time 0 to T would correspond to running L_t *backwards* from time T to 0.

There have been numerous extensions to the models captured by (4). For example, Ethier and Griffiths (1990a) describe a birth–death process which is dual to a two-locus Wright–Fisher diffusion with recombination between the two loci, and use it to prove an ergodic theorem for the diffusion. Mano (2013) uses the same process to derive a method to compute the transient moments of the diffusion. Generalising further, Ethier and Kurtz (1993) describe a duality relationship between a Fleming–Viot process with very general mutation, selection, and recombination operators and a function-valued dual process analogous to the block-counting process of the coalescent. Here, the function changes state as a jump process reminiscent of (2) due to genetic drift, selection, and recombination, while mutation contributes a deterministic component evolving the function continuously between jumps. Dualities in which mutation is either deterministic or absent can be used to compute some quantities of interest in the two models, but they are not the most general available. In this paper our purpose is different: it is to develop a coalescent dual for the Wright–Fisher diffusion in which mutation contributes to the *random* evolution of the dual process. This type of duality is important because the dual process describes the posterior genealogical dynamics of a sample, conditional on the allelic configuration observed in the present day. This is precisely the process of interest when one wishes to perform statistical inference under a coalescent model given some sample of genetic variation taken from a contemporary population (see Stephens, 2007, for an introduction). For example, a careful approximation of these dynamics provides a suitable proposal process in an importance sampling algorithm (examples for multilocus models include Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001; Larribe et al., 2002; Griffiths et al., 2008; Larribe and Lessard, 2008; Jenkins and Griffiths, 2011; Kamm et al., 2016). This duality is also important because it provides a way of obtaining an expression for the transition function of the underlying diffusion (Griffiths, 1979; Donnelly and Tavaré, 1987; Ethier and Griffiths, 1993).

Dualities of this latter form have been developed for a number of models extending (1) and (2). These include models of mutation (Griffiths, 1980; Donnelly and Tavaré, 1987), natural selection (Barbour et al., 2000; Fearnhead, 2002; Stephens and Donnelly, 2003; Etheridge and Griffiths, 2009), and Λ-coalescent dynamics (Etheridge et al., 2010), as well as dualities for the Moran model which is a prelimit of the corresponding diffusion (Etheridge and Griffiths, 2009; Etheridge et al., 2010). Hitherto, there has not been described a corresponding dual process for models incorporating both mutation and recombination (by which we mean homologous, meiotic, crossover). [The existence of one such process is implicit in Fearnhead and Donnelly (2001) and Griffiths et al. (2008), but there the focus was on inference rather than any description of the process.] The goal of this paper is to derive such a duality relationship from first principles: in particular, we identify a genealogical dual for the Wright–Fisher diffusion with recombination which is similar to the ancestral recombination graph (ARG) of Griffiths and Marjoram (1997); the key differences being that here the lineages are typed, and jumps in the genealogical process are to be understood in an *a posteriori* sense. We obtain results both for a finite-locus model with general mutation structure and for its limit with continuous breakpoint distribution and infinitely-many-sites mutation. Our key object of study is a generalisation of the generator $\mathscr{L}$ defined in (1) and the duality identity (4). As applications of our approach we recover systems of recursive equations for the sampling distribution of the models (usually obtained more toilsomely by direct coalescent arguments), and we also obtain the first transition function expansion for a diffusion model incorporating recombination. Finally, we study the case of no mutation in further detail and develop an efficient method for computing the distribution of how ancestral genetic material is dispersed across the ancestors of a contemporary population (the so-called *partitioning process*). Using duality, these distributions also yield fixation probabilities for haplotypes in multilocus models.

The paper is structured as follows. In Section 2 we illustrate our approach with a known example of a K-allele system at a single locus. We then extend this in Section 3 to an L-locus model. In Section 4 we apply these results to develop a series expansion for the transition function of the diffusion. In Section 5 we generalise the model further, to a continuous model of a gene in which mutation and recombination rates are modelled by a probability density function. In Section 6 we return to the L-locus model and study in further detail the dual process of a Wright–Fisher diffusion without mutation, and Section 7 concludes with a brief discussion.

2. Warm up: K-alleles at one locus

To illustrate the main idea and to clarify some notation, we first consider an extension of (4) to incorporate K-alleles with parent-independent mutation (PIM) at one locus. The key step is to make a judicious choice of duality function F so that, when we apply to it the infinitesimal generator of the underlying diffusion as an operator on the first variable of F, we *recognise* the resulting expression as the action of another generator acting on the second variable. Further applications of this idea can be found in Ethier and Griffiths (1993), Barbour et al. (2000), and Etheridge and Griffiths (2009).

Denote the finite type space of the locus by $E = \{1, \ldots, K\} =: [K]$. The mutation model is specified by a rate parameter $\theta > 0$ and a distribution $(P_i)_{i\in E}$ over the type of a mutant offspring (independent of the parental allele). Within this framework, the Wright–Fisher diffusion $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\geq 0}$ has state space

$$\Delta_E = \left\{ \boldsymbol{x} = (x_i)_{i\in E} \in [0, 1]^E : \sum_{i\in E} x_i = 1 \right\} \tag{5}$$

and generator

$$\mathscr{L}f(\boldsymbol{x}) = \frac{1}{2}\sum_{i\in E}\sum_{j\in E} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}) + \frac{\theta}{2}\sum_{i\in E}(P_i - x_i)\frac{\partial}{\partial x_i} f(\boldsymbol{x}), \tag{6}$$

where δ_{ij} denotes the Kronecker delta, and $\mathscr{D}(\mathscr{L}) = C^2(\Delta_E)$. Motivated by the choice of $F(x, n)$ we encountered above, let us evaluate $\mathscr{L}F(\boldsymbol{x}, \boldsymbol{n})$ for $F : \Delta_E \times \mathbb{N}^{|E|} \to \mathbb{R}$ defined by

$$F(\boldsymbol{x}, \boldsymbol{n}) = \frac{1}{m(\boldsymbol{n})}\prod_{i\in E} x_i^{n_i}, \tag{7}$$