Theoretical Population Biology 102 (2015) 3–15

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Genealogical histories in structured populations

Seiji Kumagai, Marcy K. Uyenoyama [*]

Department of Biology, Box 90338, Duke University, Durham, NC 27708-0338, USA

ARTICLE INFO

Article history:
Received 27 May 2013
Available online 11 March 2015

Keywords:
Population structure
Migration
Coalescence time
MRCA age
Generating functions

ABSTRACT

In genealogies of genes sampled from structured populations, lineages coalesce at rates dependent on the states of the lineages. For migration and coalescence events occurring on comparable time scales, for example, only lineages residing in the same deme of a geographically subdivided population can have descended from a common ancestor in the immediately preceding generation. Here, we explore aspects of genealogical structure in a population comprising two demes, between which migration may occur. We use generating functions to obtain exact densities and moments of coalescence time, number of mutations, total tree length, and age of the most recent common ancestor of the sample. We describe qualitative features of the distribution of gene genealogies, including factors that influence the geographical location of the most recent common ancestor and departures of the distribution of internode lengths from exponential.

1. Introduction

Under population structure, the rate of coalescence among genetic lineages depends upon the states of the lineages. The state may include, for example, whether lineages reside in the same or distinct individuals in a population undergoing inbreeding or in the same or distinct demes of a subdivided population. Crow and Maruyama (1971) addressed the effective number of alleles (inverse of the probability of identity between a pair of genes randomly sampled from the same deme), discussing its relationship to the time to fixation of a neutral mutation and the level of heterozygosity in a structured population. Much of the substantial body of work on structured populations has focused on analytical solutions for small samples (e.g., Nei and Feldman, 1972; Li, 1976; Griffiths, 1981; Strobeck, 1987; Takahata, 1988; Hudson, 1990; Nath and Griffiths, 1993; Wakeley, 1996; Rosenberg and Feldman, 2002; Innan and Watanabe, 2006; Wilkinson-Herbots, 2008).

In the context of an isolation-with-migration model (IM, Nielsen and Wakeley, 2001), Wang and Hey (2010) provided a detailed description of the nature of the coalescence process. Only lineages that reside in the same deme can coalesce in the immediately ancestral generation, with migration between demes inducing changes in the number residing in a given deme. Wang and Hey (2010) computed the likelihood as a convolution over the numbers of the various kinds of migration events and the time spent in various states (configuration of lineages among demes). Using a continuous time Markov chain (CTMC) framework, Hobolth et al. (2011) gave an analytical solution for the density of time since the most recent coalescence in a small sample in an IM model comprising two extant populations between which migration may have occurred since their divergence from an ancestral population. Zhu and Yang (2012) used this solution to develop a prior distribution of genealogical histories in their Markov Chain Monte Carlo (MCMC) sampler involving three extant populations. Mailund et al. (2011) incorporated the CTMC into a Hidden Markov Model for the estimation of divergence time and effective sizes of populations. Andersen et al. (2013) developed a CTMC framework to accommodate the IM model, deriving densities for coalescence time and mutation numbers.

This literature illustrates two related lines of research: developing a qualitative understanding of the effects of population structure on patterns of genetic variation and developing a computationally feasible approach to sampling-based inference frameworks. Our exploration of aspects of a sample derived from two demes falls at an intermediate point along this continuum. Our objectives include enhancing intuition about genealogies within structured populations as well as developing a model-based computational approach that may contribute to the determination of likelihoods of models and their parameters.

Building on previous methods (especially Takahata, 1988; Hudson, 1990; Uyenoyama and Takebayashi, 2004), we describe generating functions for total tree length and the number of segregating sites in a sample of arbitrary size. We explore the effect

[*] Corresponding author.
E-mail address: marcy@duke.edu (M.K. Uyenoyama).

http://dx.doi.org/10.1016/j.tpb.2015.01.003
0040-5809/© 2015 Elsevier Inc. All rights reserved.

of population structure on the density of time between successive coalescence events (internode length) within gene genealogies. To illustrate the qualitative behavior of the process, we conduct a sensitivity analysis of the location of the most recent common ancestor (MRCA) of a sample derived from a population comprising two demes.

2. Model

For a sample of n genes, the gene genealogy comprises $n-1$ coalescence events, demarcating $n-1$ levels, with level ℓ corresponding to the segment in which exactly ℓ lineages ancestral to the sample exist. At any point in the gene genealogy, each lineage may exist in various states. Within the framework of an IM model, for example, aspects of a configuration may include the type of each lineage (*e.g.*, deme of origin or location of descendants). We describe the spectrum of the states of all lineages at a level boundary as the configuration. We denote the configuration at the more recent boundary of a level as the entrance state and the configuration at the more ancient boundary the exit state. The exit state of a given level is identical to the entrance state of the next level into the past. A gene genealogy corresponds to a list of configurations, corresponding to the observed sample and every level boundary back to the MRCA, together with a list of ages of the level boundaries.

2.1. Continuous time

For any level of the gene genealogy, the process initiated at a given entrance state may visit various transient (non-coalescent) states en route to an absorbing (coalescent) state. We characterize the within-level process as a continuous-time, finite-state Markov chain, from the entrance state to the exit state. This framework appears to be essentially equivalent to that of Andersen et al. (2013).

Neuts (1995, Chapter 5) provides a lucid exposition of phase-type densities obtained from transition matrices of the form described in (5). Here, we place this general framework in the present context in order to facilitate presentation of our approach through generating functions (Section 2.2).

2.1.1. Transition probability matrix

Let $\boldsymbol{P}(t)$ denote the transition probability matrix, of which the ijth element represents the probability that a process exists in state j at time $t+s$ given its existence in state i at time s, for any s. With respect to entrance state i, we denote as transient any state j for which $\lim_{t\to\infty} \boldsymbol{P}_{ij}(t) = 0$. Termination of the level corresponds to absorption in exit state j ($\boldsymbol{P}_{jj}(t) = 1$ for all t). We restrict attention to processes in which all configurations can be classified as either transient or absorbing.

Under the Markov properties, $\boldsymbol{P}(t)$ satisfies the Chapman–Kolmogorov equations:

$$\boldsymbol{P}(t+s) = \boldsymbol{P}(t)\boldsymbol{P}(s). \tag{1}$$

In particular,

$$\boldsymbol{P}(t+dt) - \boldsymbol{P}(t) = \boldsymbol{P}(t)[\boldsymbol{P}(dt) - \boldsymbol{I}] = [\boldsymbol{P}(dt) - \boldsymbol{I}]\boldsymbol{P}(t), \tag{2}$$

for $\boldsymbol{I}$ the identity matrix and dt a small time increment, with instantaneous rates of change given by

$$\lim_{dt\to 0} \frac{\boldsymbol{P}(dt) - \boldsymbol{I}}{dt} = \boldsymbol{P}'(0) = \boldsymbol{A}. \tag{3}$$

From (2) we obtain

$$\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{P}'(0) = \boldsymbol{P}'(0)\boldsymbol{P}(t),$$

the solution of which gives

$$\boldsymbol{P}(t) = e^{\boldsymbol{A}t} = \boldsymbol{I} + \sum_{k=1}^{\infty} \frac{(\boldsymbol{A}t)^k}{k!} \tag{4}$$

(see, for example, Taylor and Karlin, 1998, Chapter VI, Section 6).

Given the entrance state of a level, the process may visit some number of transient states before absorption in an exit state. For a model of multiple demes, for example, the transient states may reflect different arrangements of the lineages among the demes and an exit state coalescence in one of the demes. For α the number of transient states accessible from the entrance state and β the total number of exit states accessible from any of the transient states, the instantaneous rates of change (3) correspond to

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \tag{5}$$

in which $\boldsymbol{U}$ (dimension $\alpha \times \alpha$) gives the instantaneous rates of within-level moves and $\boldsymbol{V}$ ($\alpha \times \beta$) between-level moves, with the lower blocks representing matrices of zeros of appropriate size ($\beta \times \alpha$ and $\beta \times \beta$).

Let λ_i represent the instantaneous rate of change from transient state i, the sum of the off-diagonal elements of row i. As the rows of $\boldsymbol{A}$ must sum to zero, the diagonal elements of $\boldsymbol{U}$ correspond to $-\boldsymbol{L}$, for $\boldsymbol{L}$ a diagonal matrix of the λ_i:

$$\boldsymbol{L} = \mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_\alpha). \tag{6}$$

From (4) and (5), we obtain the transition probability matrix

$$\boldsymbol{P}(t) = \begin{pmatrix} e^{\boldsymbol{U}t} & (e^{\boldsymbol{U}t} - \boldsymbol{I})\boldsymbol{U}^{-1}\boldsymbol{V} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}. \tag{7}$$

2.1.2. Hitting probabilities

That $\boldsymbol{U}$ represents rates of transition among transient states implies

$$\lim_{t\to\infty} e^{\boldsymbol{U}t} = \boldsymbol{0}.$$

From (7), this implies that the probability of absorption in exit state j from transient state i corresponds to the ijth element of

$$\boldsymbol{Q} = -\boldsymbol{U}^{-1}\boldsymbol{V}. \tag{8}$$

2.1.3. Density of internode length

In (7), the elements of $(e^{\boldsymbol{U}t} - \boldsymbol{I})\boldsymbol{U}^{-1}\boldsymbol{V}$ represent cumulative distribution functions, the ijth element of which gives the probability that the process reaches absorbing state j from transient state i by time t. Taking derivatives, we obtain the density of absorption time:

$$e^{\boldsymbol{U}t}\boldsymbol{V}. \tag{9}$$

A number of authors (*e.g.*, Takahata, 1988; Nath and Griffiths, 1993) have analyzed Laplace transforms of coalescence times, equivalent to moment generating functions (mgfs) of these non-negative variables. The mgf of duration of a level of the sample genealogy corresponds to

$$\boldsymbol{h}(b) = -[\boldsymbol{I}a + \boldsymbol{U}]^{-1}\boldsymbol{V}. \tag{10}$$

In general, the density may be recovered from the mgf using the inversion function

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ibt}\phi(b)db \tag{11}$$

in which $i = \sqrt{-1}$ and $\phi(b)$ represents the corresponding characteristic function,

$$\phi(b) = h(ib).$$

Whether numerical integration of (11) or computation of (9) poses greater challenges may depend on sample size and the process modeled.