Theoretical Population Biology 76 (2009) 92–104

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Structured coalescent processes from a modified Moran model with large offspring numbers

Bjarki Eldon [*]

Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, United States

ARTICLE INFO

Article history:
Received 6 February 2009
Available online 9 May 2009

Keywords:
Structured coalescent
Modified Moran model
Multiple mergers

ABSTRACT

Structured coalescent processes are derived for the finite island model under a migration mechanism that conserves the subpopulation sizes. The underlying population model is a modified Moran model in which the reproducing individual can have very many offspring with some probability. Convergence to a structured coalescent process results when assuming that migration follows a coalescent timescale which can be much shorter than the usual Wright–Fisher timescale. Three different limit processes are possible depending on the coalescent timescale, two of which allow multiple mergers of ancestral lines. The expected time to most recent common ancestor, and the expected total size of the genealogy, of balanced and unbalanced samples can be very similar, even when migration is low, if the coalescent process allows multiple mergers. The expected total size increases almost linearly with sample size in some cases. The results have implications for inference about genetic population structure.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

The structured coalescent (Takahata, 1988; Notohara, 1990; Herbots, 1997) describes the ancestral process of a finite sample of DNA sequences obtained from a population subdivided into subpopulations connected by migration. The migration mechanism is modeled in a way that keeps the subpopulation sizes constant (*conservative migration*; Nagylaki, 1980; Strobeck, 1987; Herbots, 1997). The coalescent on two subpopulations was considered by Takahata (1988), for any finite number of subpopulations by Notohara (1990), and placed in a rigorous framework by Herbots (1997). The underlying population model of the structured coalescent of Notohara (1990) and Herbots (1997) is the Wright–Fisher model (Fisher, 1930; Wright, 1931).

The Wright–Fisher model, as does any population model, has a coalescence timescale associated with it, which is proportional to the number of generations, on average, it takes for two individuals to coalesce in a single population. The timescale of a single haploid Wright–Fisher population of size N is N generations. By scaling (or 'speeding up') time by N, and assuming the population is really large (i.e. $N \to \infty$) one obtains convergence to the Kingman coalescent (Kingman, 1982a,b).

Pitman (1999) and Sagitov (1999) independently introduced the Λ-coalescent, in which any number of ancestral lines can coalesce to a single ancestor in the same coalescence event, as opposed to just two in the Kingman coalescent. In a Λ-coalescent, the rate at which each group of k ancestral lines out of n coalesce is

$$\lambda_{n,k} = \int_0^1 x^k (1-x)^{n-k} x^{-2} \Lambda(\mathrm{d}x) \tag{1}$$

where Λ is a finite nonnegative measure on Borel subsets of the interval [0,1] (Pitman, 1999). The Kingman coalescent is recovered from the Λ coalescent when Λ has unit mass at 0. Sagitov (1999) obtained convergence to a Λ-coalescent from a Cannings (1974) population model of non-overlapping generations. Eldon and Wakeley (2006) derive special cases of the Λ-coalescent from a modified Moran model of overlapping generations in which the reproducing individual can have very many offspring with some probability. The coalescence timescale in the model of Eldon and Wakeley (2006) can be much shorter than the $N^2/2$ time steps associated with the usual Moran model (Moran, 1958, 1962). One generation in the Wright–Fisher model is equivalent to N time steps in the Moran model.

Sweepstakes-style recruitment was proposed by Hedgecock (1994) and Beckenbach (1994) when considering data on Pacific oysters (*Crassostrea gigas*). Large offspring numbers may be found among marine organisms with high fecundity and high early mortality (Li and Hedgecock, 1998; Hedgecock, 1994; Beckenbach, 1994). Predictions about genetic diversity (Eldon and Wakeley, 2006), linkage disequilibrium (Eldon and Wakeley, 2008b), and estimates of migration rate based on F_{ST} (Eldon and Wakeley, 2009) have all been shown to be strongly influenced by large offspring numbers.

[*] Corresponding address: Organismic and Evolutionary Biology, Harvard University, 4100 Biological Laboratories, 16 Divinity Ave, Cambridge, MA 02138, United States. Tel.: +1 617 495 1568; fax: +1 617 384 5874.
E-mail address: eldon@fas.harvard.edu.

0040-5809/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2009.05.001

Limic and Sturm (2006) initiate the study of the Λ-coalescent in a spatial setting by deriving conditions for the process to "come down from infinity" (see also Schweinsberg, 2000b). Our work differs from that of Limic and Sturm (2006) in that we derive a limit process starting from a finite sample in a finite population, by enforcing certain conditions on migration. To derive conditions for a process to come down from infinity one necessarily starts with an infinite population and sample size.

We derive the rate matrix of a structured coalescent process under conservative migration for a finite sample from a finite number of subpopulations when individuals can have very many offspring with some probability. The underlying reproduction model is a modified Moran model introduced in Eldon and Wakeley (2006). Three different rate matrices are possible, depending on the probability of large reproduction events. The rate matrices differ in the rate of coalescence; one is the usual structured coalescent but on a Moran model timescale, while the other two allow multiple mergers of ancestral lines. We compare the different processes by calculating numerically the expected values and variances of the time to most recent common ancestor, and the total length of all the branches (total size) of the genealogy, for two different sample configurations. A key result is that, in some cases, the total size of the gene genealogy is essentially the same for the two sample configurations, even when migration is low.

2. General framework

We will follow the framework of Herbots (1997) in establishing convergence to a structured coalescent process, and in many cases adopt the same notation. In particular, we assume conservative migration (Nagylaki, 1980; Strobeck, 1987; Herbots, 1997), so the subpopulation sizes stay constant.

The number D of subpopulations in our model is finite. Let $c_i \in \{1, 2, \ldots\}$ be fixed and finite for all $i \in \mathcal{D} \equiv \{1, \ldots, D\}$ the set of subpopulation labels. Then $N_i \equiv c_i N$ is the population size of subpopulation i. We let $n_i(\tau)$ denote the number of ancestral lines present in subpopulation i τ timesteps into the past. One timestep in the usual Moran model corresponds to $1/N$ generations in the usual Wright–Fisher model. Then $\mathbf{n}_N(\tau) \equiv (n_1(\tau), \ldots, n_D(\tau))$ is the ordered finite set of the numbers of ancestral lines in each subpopulation at time τ. With $\mathbb{N} \equiv \{0, 1, 2, \ldots\}$, $\mathbf{n}_N(\tau) \in \mathbb{N}^{\mathcal{D}}$. The finite state space E of $\mathbf{n}_N(\tau)$ is

$$E \equiv \left\{ \mathbf{n} \in \mathbb{N}^{\mathcal{D}} : \sum_{i=1}^{D} n_i \leq n \right\}.$$

Herbots (1997) allows the number D of subpopulations to be infinite without rescaling time with D. However, in that case the expected time $\mathbb{E}T$ to most recent common ancestor of two ancestral lines in different subpopulations would be infinite, since $\mathbb{E}T$ is proportional to D (Nei and Feldman, 1972; Li, 1976; Griffiths, 1981).

Wakeley (1999, 2001) derives a structured coalescent process in which the number of demes $D \to \infty$, by applying Möhle's (1998) theorem of separation of timescales. In Wakeley's (1999) model, no more than two ancestral lines can be found in any subpopulation after a scattering phase, in which ancestral lines migrate to different subpopulations. If we did assume Wakeley's (1999) model of population subdivision with a population model of large offspring numbers, there would never be an opportunity of a coalescence event involving more than two ancestral lines. Eldon and Wakeley (2009) derive densities of coalescence times for two sequences sampled from a structured population as $D \to \infty$ by applying Möhle's (1998) theorem. Taylor and Véber (2009) study the infinitely many demes limit model with sporadic extinction and recolonization events, and find that the genealogy of a sample can have *simultaneous* multiple mergers (Ξ-coalescent; Schweinsberg, 2000a) in some cases.

We will establish convergence in distribution of the ancestral process $\mathbf{n}_N = \left\{ \mathbf{n}_N([tN_\gamma/2]) : t \geq 0 \right\}$ to a structured coalescent process $\{\mathbf{n}(t) : t \geq 0\}$, which is a continuous-time Markov chain, and $N_\gamma/2$ is the timescale of the process, as explained in the next section.

The infinitesimal generator of $\mathbf{n}(t)$ takes one of three forms given by Eqs. (49)–(51), depending on the coalescence timescale. As explained in Section 3, the coalescence timescale is proportional to $\min(N^\gamma, N^2)$, $0 < \gamma < \infty$, and so can be much shorter than the usual Wright–Fisher timescale of N timesteps (generations).

To explain the transitions of $\mathbf{n}(t)$, define the indicator function I_A as

$$I_A \equiv \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

For example, $I_{\gamma<2} = 1$ if $\gamma < 2$, and zero otherwise. Let $\boldsymbol{\epsilon}^i \in \mathbb{N}^{\mathcal{D}}$ denote the unit vector with elements ϵ_j such that $\epsilon_j \equiv I_{j=i}$. The transitions that change the value of $\mathbf{n}(t)$ are migration events and coalescence events. When an ancestral line migrates from subpopulation i to subpopulation j, $\mathbf{n}(t)$ changes value from $\mathbf{n}$ to $\mathbf{m} = \mathbf{n} - \boldsymbol{\epsilon}^i + \boldsymbol{\epsilon}^j$, $j \neq i$. In the case of a coalescence event in subpopulation i, $\mathbf{n}(t)$ changes value from $\mathbf{n}$ to $\mathbf{m} = \mathbf{n} - (x-1)\boldsymbol{\epsilon}^i$, with $2 \leq x \leq n_i$. In the usual structured coalescent, x is fixed at two.

The independence of migration and reproduction (Herbots, 1997) allows us to consider the coalescence and backwards migration processes separately. Let $\mathbf{P}_N$ denote the single timestep transition probability matrix of the ancestral process. Similarly, let $\mathbf{P}_N^{(m)}$ and $\mathbf{P}_N^{(c)}$ denote the single time step transition probability matrices of the backwards migration and coalescence processes, respectively. We can then write

$$\mathbf{P}_N = \mathbf{P}_N^{(m)} \cdot \mathbf{P}_N^{(c)}. \tag{3}$$

In proving convergence we write each matrix $\mathbf{P}_N^{(m)}$ and $\mathbf{P}_N^{(c)}$ in the form $\mathbf{I} + \mathbf{Q}_N/(N_\gamma/2) + \mathbf{R}_N$. The matrix $\mathbf{I}$ denotes the identity matrix. The matrix $\mathbf{Q}_N$ denotes the corresponding rate matrix, and holds terms from the corresponding single time step matrix that are of order $O(1/N_\gamma)$. The terms in $\mathbf{Q}_N$ include probabilities of coalescence involving ancestral lines in one subpopulation, or probabilities of migration involving one ancestral line. Finally, the matrix $\mathbf{R}_N$ holds all higher order terms, such as probabilities of coalescence of ancestral lines in more than one subpopulation, or of migration involving more than one ancestral line. Employing the dominated convergence theorem we conclude that with time rescaled in units of $N_\gamma/2$ time steps, the single time step ancestral process described by $\mathbf{P}_N$ will in a large population be approximated by a continuous-time Markov process with a infinitesimal generator whose exact form depends on γ (see Eqs. (49)–(51)).

3. Population model

Given n_i ancestral lines present in subpopulation i, the probability that x out of the n_i lines coalesce (an x-merger, $2 \leq x \leq n_i$) is

$$G_{n_i,x} = \sum_{u=2}^{c_iN} P_U(u) \frac{\binom{u}{x}\binom{c_iN-u}{n_i-x}}{\binom{c_iN}{n_i}}. \tag{4}$$

In the standard Moran model, $U = 2$ always, but we assume that U is a random variable with probability distribution P_U. We consider the simple distribution for U in Eq. (5) (in which $0 < \varepsilon_i < 1$):

$$P_U(u) = (1-\varepsilon_i)\, I_{u=2} + \varepsilon_i I_{u=\psi_i N_i}, \quad 2 \leq u \leq N_i. \tag{5}$$

The parameters ψ_i ($0 < \psi_i < 1$) have a clear biological meaning as the fraction of the population in subpopulation i replaced by the

Download English Version:

https://daneshyari.com/en/article/4502773

Download Persian Version:

https://daneshyari.com/article/4502773

Daneshyari.com