Theoretical Population Biology xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Theoretical Population Biology

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

An efficient algorithm for generating the internal branches of a Kingman coalescent

M. Reppell [a,*], S. Zöllner [b,c]

[a] Department of Human Genetics, University of Chicago, Chicago, IL, USA
[b] Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
[c] Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA

ARTICLE INFO

Article history:
Received 22 December 2016

Available online xxxx

Keywords:
Coalescent
Genealogical topology
Coalescent simulations

ABSTRACT

Coalescent simulations are a widely used approach for simulating sample genealogies, but can become computationally burdensome in large samples. Methods exist to analytically calculate a sample's expected frequency spectrum without simulating full genealogies. However, statistics that rely on the distribution of the length of internal coalescent branches, such as the probability that two mutations of equal size arose on the same genealogical branch, have previously required full coalescent simulations to estimate. Here, we present a sampling method capable of efficiently generating limited portions of sample genealogies using a series of analytic equations that give probabilities for the number, start, and end of internal branches conditional on the number of final samples they subtend. These equations are independent of the coalescent waiting times and need only be calculated a single time, lending themselves to efficient computation. We compare our method with full coalescent simulations to show the resulting distribution of branch lengths and summary statistics are equivalent, but that for many conditions our method is at least 10 times faster.

1. Introduction

In recent years the declining costs of human sequencing and genotyping have facilitated increasingly large studies. Sequencing experiments with tens of thousands of samples (Coventry et al., 2010; Nelson et al., 2012; Tennessen et al., 2012), and genotyping projects combining hundreds of thousands of samples (Teslovich et al., 2010; Morris et al., 2012; Berndt et al., 2013) are now common. The data observed in such large studies is frequently compared with simulated data generated according to theoretical models to test hypotheses about demography or disease architecture. Coalescent simulations are a common and widely used approach for generating such simulated data (Nelson et al., 2012; Ferreira et al., 2013; Gazave et al., 2014). The coalescent (Kingman, 1982) is a model which traces the ancestry of a present day sample backwards through time until reaching the most recent common ancestor of the entire sample. Researchers have expanded the coalescent to model a range of population histories and conditions (Kaplan et al., 1988; Takahata and Slatkin, 1990; Griffiths and Tavarè, 1994; Neuhauser and Krone, 1997). However, for large samples, coalescent simulations can become computationally burdensome, especially in Monte-Carlo applications when many datasets have to be generated. Here, we propose a sampling method that allow us to generate individual branch lengths and configurations in Kingman coalescent genealogies, and which allow us selectively generate limited portions of sample genealogies. This approach is particularly effective for research questions that need consider only limited portions of the full genealogies generated by coalescent simulations. For example, the study of very rare variants found to be abundant in human populations (Nelson et al., 2012; Tennessen et al., 2012), concerns only external or nearly external genealogical branches.

Our method relies on a set of equations that give the probability of a genealogical branch starting and ending at specific coalescent events. To derive these probabilities we first define the *length* and *size* of a branch. The structure of a coalescent genealogy is a bifurcating tree, with internal nodes that represent coalescent events where two lineages merge at a common ancestor. Therefore, a genealogical branch begins either at an external node along the tips of the tree or at a subsequent internal coalescent event, and then ends at a coalescent event closer to the root of the tree. The time between its beginning and ending events is the *length* of a branch. The *size* of a branch is a count of the number of external nodes in the final sample that it subtends (Fu, 1995). Branch size corresponds to the number of derived alleles that would appear in the final sample were a mutation event to occur along the branch's length. With a constant size population, where waiting

[*] Corresponding author.
E-mail address: mreppell@uchicago.edu (M. Reppell).

http://dx.doi.org/10.1016/j.tpb.2017.05.002
0040-5809/© 2017 Elsevier Inc. All rights reserved.

times between coalescent events are independent, the combination of these equations provides an explicit probability distribution function for individual branch lengths. Directly sampling from this explicit formula is computationally challenging. Here we introduce a recursive calculation that gives the distribution of the number of branches with a given size in a genealogy. We show that these recursive computations combined with storing reusable intermediate results and sampling from simple exponential distributions facilitate a rapid method for sampling selected portions of genealogies.

Our approach of developing an algorithm targeted at a specific feature of the coalescent has previously been taken in many contexts. Simulation methods grounded in coalescent theory and designed to efficiently handle recombination (McVean and Cardin, 2005; Marjoram and Wall, 2006), selection (Fearnhead, 2006), or the number of ancestral lineages remaining (Blum and Rosenberg, 2007; Jewett and Rosenberg, 2014) have all been proposed. Previous work on the distribution of internal branches of the Kingman coalescent was focused on their summed length or the proportion of a genealogy with a given size (Fu and Li, 1993; Fu, 1995; Griffiths and Tavarè, 1998; Wooding and Rogers, 2002; Polanski and Kimmel, 2003; Dahmer and Kersting, 2015). The proportion of a tree with a given size was of interest because under the infinite sites mutation model, the number of segregating sites observed in a sample with a given number of derived alleles is a function of the total length of branches with a size equal to the number of derived alleles. Fu and Li (1993) presented the expectation and variance of the total summed length of both external and internal branches along a Kingman coalescent without recombination. Griffiths and Tavarè (1998) expanded on this work to derive an expression for the probability of a mutation having a specific number of descendants in the final sample, even in samples from populations with variable past sizes. Jenkins and Song (2011) built on Griffiths and Tavaré's work by considering allele configurations with two separate mutation events, and they extended their work to variable size populations in 2014 (Jenkins et al., 2014). In related work Ferretti et al. (2016) was able to derive closed expressions for the joint frequency spectrum of two linked sites. Fu (1995) gave expressions for the expectations, variances, and covariances for a sample's frequency spectrum. Efficient methods for modeling the total time in a sample's genealogy with a given size have been developed (Wooding and Rogers, 2002; Polanski and Kimmel, 2003; Polanski et al., 2003). However, the methods of Wooding and Rogers (2002) and Polanski et al. (2003) fail to model individual branch lengths and their topology. The topology of a genealogy is where the correlation between observed mutations arises. These correlations can contain information about demography lacking from the frequency spectrum (Gutenkunst et al., 2009) and influence the outcome of tests for neutral evolution (Ledda et al., 2015). In the absence of recombination, this correlation is equivalent to the linkage disequilibrium between mutations, and it has been shown that patterns of linkage disequilibrium between very rare variants can provide information about departures from Wright–Fisher neutrality (Wall, 1999), including recent population growth rates (Reppell et al., 2014). With the focus in our work on individual genealogical branches rather than their summed length, we more closely build on the findings of Rosenberg (2006), which derived the expectation and variance for the number of internal branches with a specific size.

Here our calculations build a sampling framework that can quickly generate portions of a genealogy with a specific size. Considering all coalescent events on a tree, we integrate over all possible starts and ends for a branch of a given size. Conditional on the start and end of the branch we then calculate the probability that the branch has a given length. For a constant size population, we show that our work gives rise to an explicit probability distribution function for branch lengths. As this formula becomes computationally intractable as sample size grows, we introduce a computationally more efficient algorithm that recursively calculates all probabilities of start and end points and evaluates the conditional probability of branch length by Monte Carlo sampling. We compare our sampling method with full coalescent simulations for a range of sample sizes and demonstrate it performs up to 10 times faster, and show that as long as the ratio of branch size to sample size is moderate (<0.15) it produces branches with an equivalent length distribution and summary statistics.

2. Methods

In this section we first provide the full probability distribution function for genealogical branches under a model of constant population size, and then subsequently derive its components, notably in 2.2 and 2.3. In Section 2.4 the distribution of the number of branches with a given size in a genealogy is derived, which we combine with the proceeding work to propose a sampling method that can efficiently generate selected portions of genealogies. In Section 2.5 we combine the elements of the preceding sections into our proposed algorithm, which we label topology free sampling. Section 2.6 gives summary statistics we use to evaluate our method, and Section 2.7 gives details of the open source software implementation of our method and the simulations we use in this text.

2.1. A probability distribution function for coalescent branch lengths in a model with constant population size

The probability that a coalescent tree branch of size j has length ℓ is the product of three probabilities: the probability that the branch begins at specific coalescent event, then, conditional on its starting event, the probability that it ends at a specific coalescent event, and finally, conditional on its starting and ending events, the probability that the intervening coalescent times sum to ℓ. For the random variable L_j, the length of a branch with size j:

$$P(L_j = \text{Length } \ell) = \sum_{Start}\sum_{End} P(\text{Length } \ell|Start, End)P(End|Start)P(Start|Size = j) \tag{1}$$

$P(End|Start)$ and $P(Start|Size = j)$ are given in Sections 2.2 and 2.3, respectively. For a constant size population, the length of a branch follows a hypoexponential distribution: it is a sum of coalescent waiting times, each an exponential random variable with a unique rate. The rates that define the hypoexponential distribution are conditional on a branch's starting and ending coalescent events which define the number of ancestral lines remaining during the branch's duration. If we label coalescent events $k \in 1, 2, \ldots, n-1$ such that at event k, $n-k+1$ ancestral branches are reduced by 1 to $n-k$ ancestral branches we can write the exact probability distribution of branch lengths with size j as

$$P(L_j = \ell) = \sum_{k=1}^{n-2} P_{Start}(k|j) \times \sum_{b=k+1}^{n-1}\left[P_{End}(b|k,j)\sum_{z=k+1}^{b}\frac{e^{-\binom{n-z+1}{2}l}\prod_{v=k+1}^{z}\binom{n-v+1}{2}}{\prod_{v=k+1,v\neq z}^{b}\left(\binom{n-v+1}{2}-\binom{n-z+1}{2}\right)}\right] \tag{2}$$

where $P_{Start}(k|j)$ is the probability a branch with size j begins at coalescent event k (Eq. (4)), $P_{End}(b|k,j)$ is the conditional probability a branch with size j that began at event k, ends at event b

Please cite this article in press as: Reppell, M., Zöllner, S., An efficient algorithm for generating the internal branches of a Kingman coalescent. Theoretical Population Biology (2017), http://dx.doi.org/10.1016/j.tpb.2017.05.002.