Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 194202, 8 pages
http://dx.doi.org/10.1155/2014/194202

Research Article

On Coalescence Analysis Using Genealogy Rooted Trees

Ao Yuan,[1] Gengsheng Qin,[2] Wenqing He,[3] and Qizhai Li[4]

[1] *Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC 20057, USA*
[2] *Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, USA*
[3] *Department of Statistics and Actuarial Science, University of Western Ontario, London, ON, Canada N6A 5B7*
[4] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Ao Yuan; yuanao@hotmail.com

Received 10 August 2013; Accepted 10 January 2014; Published 23 February 2014

Academic Editor: Henggui Zhang

DNA sequence data are now being used to study the ancestral history of human population. The existing methods for such coalescence inference use recursion formula to compute the data probabilities. These methods are useful in practical applications, but computationally complicated. Here we first investigate the asymptotic behavior of such inference; results indicate that, broadly, the estimated coalescent time will be consistent to a finite limit. Then we study a relatively simple computation method for this analysis and illustrate how to use it.

1. Introduction

In the past decades, considerable progress has been made in the field of population genetics. One of the main goals is to infer the coalescence time of the population under study, that is, to infer the time since their most recent common ancestor (MRCA) and its distribution based on the observed data.

In genetics, coalescent theory is a retrospective of population genetics that traces all genes in a sample from a population to a single ancestral copy shared by all the members of the population. The coalescent time of a population is the time of their most recent common ancestor. The inheritance relationship among the genes is typically represented as a gene genealogy, similar to a phylogenetic tree. The goal of coalescent analysis is to infer the coalescent time of a sample of n individuals independently sampled from a population of size N, based on their observed DNA sequence diversity. Unlike parameter inference for independent and identically distributed (iid) data, for which asymptotic limit can be used conveniently to characterize the estimator when the data size is large, various existing studies indicate that the estimated MRCA, in unit of N generations, is unclear as whether it will concentrate as the data sample size increases without bound. In contrast, in the estimation of mutation rate in the same setting, the estimate is consistent and asymptotically normal [1], although at a much slower rate of $\log^{1/2}(n)$, compared to the rate of $n^{1/2}$ for i.i.d. data. Also, different from usual parameters, the MRCA changes with n, the number of sequences. This prompts us to the investigate the asymptotic behavior of the estimated coalescent time. We want to know whether such estimator will be asymptotically consistent and in what sense if it does. Conditioning on the total number of segregating sites, we find that such estimators converge or not to some nonnegative finite limits in posterior mean, depending on the behavior of the number of mutations on all the branches of the rooted trees constructed from the observed data. Also, analysis of this problem with this type of data is often computationally extensive and complicated; we study a relatively simple simulation method for this problem. We first study the asymptotic behavior of this method in Section 3, and then describe and illustrate our method for this problem in Section 4.

In coalescence inference, mitochondrial DNA (mtDNA) data plays an important role. Mitochondria is one of the few genes existing outside the cell nucleus, and for mammalian it is only maternally inherited. Human mtDNA is a double-stranded molecule sequence about 16,500 base pairs in length. It is outside the cell nuclear, and it is known that the mutation

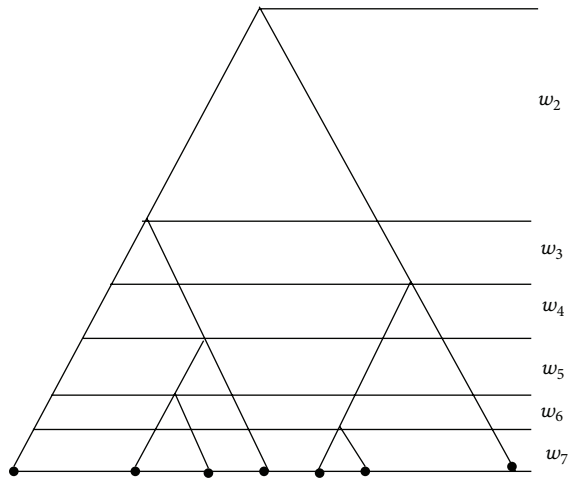

Figure 1: Coalescent tree for a sample of seven individuals.

rate in mtDNA is about 10 times that of the nuclear genes, and that on one section of the mitochondria, its control region, the mutation rate is even one order higher. The simple inheritance pattern and high variability make mtDNA an important source in the study of human evolutionary history. Each site on the DNA strand has one of the four bases A, C, G, or T. As the molecule evolves, mutations occur in the form of base substitutions. The change between purines (A,G) or pyrimidines (C,T) is called transition; that between a purine and pyrimidine is transversion. The former type of substitution is much more common than the latter.

We focus on the control region of the mitochondrial data in Griffiths and Tavaré [2], which is part of the data in Ward et al. [3]. They are from a segment of the control region, with 352 base pairs (sites), out of which 159 are purine sites and 193 are pyrimidine sites. This data contains 63 sequences sampled from a North American Indian tribe, the Nuu-Chah-Nulth, from Vancouver Island. After eliminating sequences with multiple mutations on some single sites, so that the assumption of at most one mutation each site is met, the remaining data has 55 sequences, with 14 distinct sequences (called lineages) in the data. Site at which not all the observed sequences have the same base is a *segregating site*. The whole sequences are long, but only the segregating sites are informative for the analysis; the other sites are ignored. The mentioned data has 18 segregating sites and is presented in Table 1, with the frequency (or multiplicity) of each lineage.

2. Brief Review of Background and Related Methods

The coalescent is a model for the genealogical tree of a random sample of n DNA sequences from a large population. An example of such a tree of sample size $n = 7$ is given in Figure 1.

For more detailed reviews of this topic, see Hudson [4] and Donnelly and Tavaré [5].

In coalescence inference one has the following.

Basic Assumptions. The population size N is large, remains unchanged for many generations into the past, and is known, or can be estimated from other sources; the data is a random sample from the population; the number of births in each generation follows the Wright-Fisher model (since the population is of constant size, the number of deaths also follows the similar model); mutation (substitution) at any nucleotide site can occur only once in the ancestry and is irreversible; mutations that occur in different time intervals are independent; the time point at which mutation occurs follows a Poison distribution with rate $\theta/2$ to be defined latter, independently in each branch of the genealogy tree, where θ is known, or can be estimated from other methods or sources.

The inference of coalescence time t_n of a sample population of size n has two steps. The first step is modeling the distribution of t_n without any data, the *predata* distribution; then in the second step, update the predata distribution, using the observed data, to the *postdata* distribution, based on which the formal inference is conducted. The predata distribution is pioneered by Kingman [6, 7]; he showed that, in time units of N generations,

$$t_n = \sum_{j=2}^{n} w_j, \tag{1}$$

where the w_j's are independent waiting times. w_j is the time from $j-1$ common ancestors of the sample to j common ancestors. A quick reference on this can be found in Tavaré [8]. Here w_j is distributed as exponential $\mathrm{Exp}(j(j-1)/2)$, with $E(w_j) = 2/(j(j-1))$. The w_js can be represented graphically as a coalescent tree as in Figure 1; then t_n is the height of the tree. Define the tree length as

$$l_n = \sum_{j=2}^{n} j w_j; \tag{2}$$

then (Kingman)

$$\begin{aligned} E(t_n) &= 2\left(1 - \frac{1}{n}\right), \\ \mathrm{Var}(t_n) &= 8\sum_{j=2}^{n}\frac{1}{j^2} - 4\left(1 - \frac{1}{n}\right)^2; \\ E(l_n) &= 2\sum_{j=1}^{n-1}\frac{1}{j}, \\ \mathrm{Var}(l_n) &= 4\sum_{j=1}^{n-1}\frac{1}{j^2}. \end{aligned} \tag{3}$$

The time unit is transformed to years by the relationship $t_n NY$, where Y is the average years of each generation, which is usually taken as 20–25. Here we see that, as an initial analysis without the observed data, the coalescent time of a random sample of size n from a population of size N is roughly $2N$ generations, as long as $n(\leq N)$ is moderately large.

Table 1: Nucleotide position in control region.

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Lineage
			Purines									Pyrimidines							Freqs.
Lineage																			
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	5
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

Each row of the table represents a DNA sequence lineage. In this data, there are transitions but no transversion observed.

Thus, the coalescent time of a sample from a subpopulation is roughly the same as that of the population (as long as the sample size is moderately large). This phenomenon is further investigated by Watterson [9], who showed that

$$P\left(A_N\left(t_n\right)=1\right)=\frac{(n-1)\,(N+1)}{(n+1)\,(N-1)}, \tag{4}$$

where $A_N(t_n)$ is the number of ancestors, at t_n generations ago, of the population with size N from which the data sample of size n is drawn. Here the sample must be a random draw from the population; otherwise the result may not be reliable. For example, the sample of size n is drawn from a subpopulation of size $N_1 < N$ from a population of size N; then by (3), the predata estimated of the coalescent time t_n of this sample is roughly $2N_1$ generations, but also it is roughly $2N$ generations since the sample is also from the whole population. The paradox arises from the sampling scheme. If the sample is drawn from the subpopulation of size N_1, one can only use $2N_1$ as the time scale, not $2N$, since the samples drawn from the subpopulation are expected to have smaller genetic variation than from the whole population.

For mutation, the common assumption is that the times at which mutation occurs follow a Poison process with constant rate $\theta/2$, so that, in any branch of length l from the tree, the number of mutations on that branch has a Poison distribution with mean $l\theta/2$, independently of the mutations on the other branches. For the time scale mentioned before, usually $\theta = 2N\mu$, where μ is the probability of a mutation that occurs per sequence per generation. For DNA sequences, μ is the sequence length (number of bases) times the mutation rate per site per generation and is often available from other sources. Since the coalescent time of a sample with moderate size is approximately $2N$ generations, θ can be approximately interpreted as the cumulative (since the time of MRCA) mutation rate (number of mutations) per sequence. Also, since the population size is N, $\theta/2$ can also be interpreted as the mutation rate of the whole population per generation.

Thus, given the mutation rate θ and the tree length l_n, the number of mutations s_n in a sample of n individuals from the given population follows the Poison distribution $\mathrm{Po}(\theta l_n/2)$ [10]

$$\begin{aligned} P\left(s_n=k\mid l_n=l\right) &= e^{-\theta l/2}\frac{(\theta l/2)^k}{k!} \\ &:= \mathrm{Po}\left(k,\frac{\theta l}{2}\right) \quad k=0,1,2,\ldots \end{aligned} \tag{5}$$

Note that this probability does not depend on n, but on k, l, and θ. Why $\theta l_n/2$?. Take $n = 2$; then $\theta l_n/2 = \theta t_n \approx \theta$, which is the expected number of cumulative mutations since t_n generations ago in a sequence. So $\theta l_n/2$ is a reasonable choice of the parameter in the Poison distribution. But if we model the number k of cumulative mutations per sequence since t_n generations ago, for moderately large n, we should use $\mathrm{Po}(k, t_n N\mu) \approx \mathrm{Po}(k, 2N\mu) = \mathrm{Po}(k, \theta)$.

The key in the coalescence inference is to evaluate the postdata distribution of t_n, which is much more involved than its predata distribution, it depends heavily on the mutation distribution in the data. For example, if more mutations occur in the earlier stage of the genealogy tree, then the estimated t_n will be bigger. Although under the assumption that mutation can only occur at most once at each site and mutation is irreversible, the total number of mutations in the observed data is just the number of segregating sites. But how the mutations distribute in the branches of the genealogy tree is unknown. Such distribution is crucial in the inference of t_n, which depends on how much data information being used and on the actual methods. This is our focus from now on. Denoting by D_n the observed data, the estimated coalescent

time $\widehat{t}_n$ of the sample is given by the postdata distribution mean of t_n as

$$\widehat{t}_n = E\left(t_n \mid D_n\right). \tag{6}$$

The inference can be viewed as a Bayesian procedure, with the predata and postdata distributions that correspond to the prior and posterior distributions in a Bayesian framework. But unlike the common Bayes setting, here the parameter t_n varies with the sample size n, and the data cannot be modeled i.i.d. with this parameter. That is the reason the inference of t_n cannot be made arbitrarily accurate, in the sense that the variance of the postdata distribution cannot be arbitrarily small, as the sample size increases without bound. Also, generally the postdata distribution is not in closed form and has to be evaluated by sampling methods. Tavaré et al. [10] derived the postdata distribution based on only the number of segregating sites in the sample. This method is very convenient to use, but does not use the DNA sequences structural information. The well known method in Griffiths and Tavaré [2], hereafter GT, is based on the full data information represented by a set of rooted trees. This method is one of the basic tools in coalescent inference using full data information, but is computationally complicated.

To evaluate the postdata coalescent distribution, GT used the probabilities recursion formula, derived in Ethier and Griffiths [11]. The method is not easy to fully understand and correctly use for many geneticists. Also these probabilities are computationally prohibitive; the postdata distribution of t_n is computed by a Markov chain Monte Carlo sampling and is quite involved.

Here we study a relatively simple approximate method using the full data information; in this method, instead of computing the tree probabilities as in GT, we just set the postdata tree probabilities as uniform for the $s+1$ rooted trees and use a simulation method to compute the coalescent distribution; thus, getting round of the complicated evaluations of the tree probabilities, it is easy to understand and much simpler in computation.

The rooted tree plays an important role in the analysis, which is not uniquely determined from the data. The data is equivalent to an unrooted tree, which is equivalent to a set of unrooted trees. Each rooted tree has a 0-1 valued matrix representation which is convenient for some computations, but not any 0-1 valued matrix corresponds to a rooted tree. In the following, we give more details about them and their relationships.

Rooted Tree. A rooted tree consists of a system of branches, subbranches, and so forth. The tip of each branch or subbranch represents a known lineage. The observed mutations in the sample are represented as dots in the branches, subbranches, and so forth at specified positions. The observed multiplicity of each lineage is represented as leaves at the tip of each branch or subbranch, and so forth.

The presentation of a rooted tree is unique up to the relative positions of its branches, subbranches, and so forth. A rooted tree has several levels of randomness. If we only know the sample size n, then the rooted tree has a total of n leaves; apart from that, the shape of the tree, how to split, how to allocate the leaves, how many mutations, and the distribution of the mutations are all random. If the data and the number of mutations are given, then the tree can only take a few shapes. Different from GT and other related literatures, here we put the observed lineage frequencies (multiplicities) as leaves in the corresponding tips of branches, subbranches, and so forth of the rooted tree.

Different from a coalescent tree which has a complete time ordering of the splitting points of branches, a rooted tree has only partial time orderings of these splits and mutations. We only know that splits of branch(es) occurred before those of its subbranches, but do not know the ordering of splits of different branches. We know that mutation(s) on the branch occurred before those on its subbranch(es), but do not know the order of ones on the same branch, same subbranch(es), or on different subbranches. For a given sequence data, it may correspond to more than one different rooted tree. For the observed data in Table 1, all the columns are for segregating sites, and there is no transversion. Under the assumption that mutation can only occur at most once at each site and mutation is irreversible, at each segregating site, one and only one of the base types is mutant; the other type is ancestral. So if we know the mutation status at each segregating site, the mutation statuses are said to be labelled, and we can use a 0-1 valued matrix $\mathbf{X} = (x_{ij})$ to denote the observed data, where $x_{ij} = 1$, if the base type of lineage i at site j is mutant, and $x_{ij} = 0$ otherwise. Such 0-1 matrix representation of the data is convenient in the analysis. It is easy to see that each rooted tree uniquely determines a 0-1 valued matrix $\mathbf{X}$, but an arbitrary 0-1 valued matrix may not correspond to a rooted tree. It must satisfy some conditions to corresponds a rooted tree. There are abundant methods and algorithms on how to judge if a given 0-1 values matrix is a valid representation of a rooted tree, and if so how to build the rooted tree (e.g., [12–16]). We find the method that appeared in a number of articles and is stated as Lemma 1 in Gusfiled [16] is easy to use. Given a valid 0-1 valued matrix $\mathbf{X}$ (means it satisfies the condition for representing a tree), one can uniquely draw a rooted tree corresponding to it. Here, uniqueness means the genealogy relationships, including which lineages are in the same branch or subbranch, and so forth and which mutation sites are on which section of which branch or subbranch and so forth, are determined, but the particular shape of the tree, such as some branch put on the left or right side, the angle of branches, their lengths, and so forth, are irrelevant. Thus, there is a 1-1 correspondence between a rooted tree and a valid 0-1 valued matrix. Given the observed data, the mutation statuses at the sites are usually unknown. For data with s segregating sites, there are 2^s different ways to labelling the mutation statues, but most of the labeling matrices do not qualify to be representations of a rooted tree; it is known that there are only $s + 1$ different rooted trees, and hence $s + 1$ different labellings (matrices) correspond to the data, and there are existing algorithms to construct the rooted trees and their corresponding matrices (e.g., [16, 17]). However, we find that the method in GT is convenient. By this method, one first needs to construct one rooted tree from the data or its valid 0-1 valued matrix. For example, start from the least shared mutations labeling that, on each column (site)

of the data, label the less common base type as mutant (the other as ancestral). It is easy to check the conditions for its validity using Lemma 1 mentioned above. Construct the rooted tree corresponding to this matrix and convert it to an unrooted tree as in GT; that is, absorb those subbranches without mutations into their branch(es), and then straighten the branches, subbranches, and so forth. The unrooted tree is uniquely determined from any of the $s+1$ rooted trees.

Then, based on this unrooted tree, one can get all the other rooted trees as in Griffiths and Tavaré [18]; that is, alternatively put the tree root point near each of the vertexes that stretch out that vertex, then arrange the branches, subbranches, and so forth into the desired shapes; if there are more than one mutation between two adjacent vertexes, put the tree root point in the middle of two such adjacent mutations, alternatively for all such pairs of mutations, and shape the tree as above. This way we get all the rooted trees from the unrooted tree. In fact, given any rooted tree, all the other s rooted trees can be constructed in the same way above, without using the unrooted tree. Once the rooted trees are constructed, the corresponding matrix representations are at hand.

3. Asymptotic Behavior of MRCA Estimate

For parameter inference with independent and identically distributed data and sample size n, it is known that the estimator is asymptotically consistent and asymptotically normal with rate $\sqrt{n}$. But for inference of MRCA, the data D_n are not independent and identically distributed, and existing studies indicated that the distribution of the estimated MRCA $t_n \mid D_n$ will not concentrate, even if $n \to \infty$. In the case of estimating the mutation rate with the same data, the estimator is found to be consistent and asymptotically normal with rate $\log^{1/2}(n)$ [1]. This motivates us to investigate the asymptotic behavior of $\bar{t}_n = E(t_n \mid D_n)$ as a commonly used point estimator of the coalescent time. We want to know whether this estimator has similar asymptotic behavior as the mutation rate estimator. We find that such estimators are not consistent almost surely. To describe the result, we consider the data set in three different commonly used forms. The first type of data we consider is in the form of a coalescent tree as in Figure 1. This type of data is often not practical, as for most real data we do not have the information to construct such tree. But as a starting point it will provide us some guide on the result. There are $n-1$ nodes (splitting points) in the tree numbered 2 to n in their time order. Recall the definition of the ith coalescent time w_i. Between the $(i-1)$th and ith node there are exactly i segments, denote them as $w_{i1}, \ldots, w_{ii}$ from left to right, each has length w_i. Assume the number of mutations k_{ij} on segment w_{ij} is known. Let $\mathbf{w} = \{w_{ij} : i = 2, \ldots, n; j = 1, \ldots, i.\}$, $\mathbf{k} = \{k_{ij} : i = 2, \ldots, n; j = 1, \ldots, i.\}$ be the mutation distribution corresponding to $\mathbf{w}$ and $k_i = \sum_{j=1}^{i} k_{ij}$. Here this type of data is fully represented by $\mathbf{k}$. When we do not have $\mathbf{w}$, $\mathbf{k}$ is not uniquely determined. But given each rooted tree $\mathbf{T}_r$, $\mathbf{w}$ and the location information of the mutations, a mutation vector $k_r = \{k_{r,i} : i = 2, \ldots, n; k_{r,i} = \sum_{j=1}^{i} k_{r,ij}\}$ can be constructed by a random manner (to be detailed in Section 4) corresponding to $\mathbf{T}_r$. Denote $\pi(\mathbf{T}_r) = \pi(\mathbf{T}_r \mid D_n) = 1/(s+1)$ $(r = 1, \ldots, s+1)$ be our prior on the rooted tree $\mathbf{T}_r$'s, that is, without additional knowledge we treat each rooted tree as equally likely from the observed data. Here our $\pi(\mathbf{T}_r)$'s have different meaning from the probabilities $p^0(\mathbf{T}_r, \mathbf{n})$'s as in GT (the latter do not sum up to one, but to the probability of obtaining the unrooted tree from the observed data). We have (Appendix)

$$E(t_n \mid D_n, \theta) = \sum_{r=1}^{s+1} E\left[E(t_n \mid \mathbf{k}_r, D_n, \theta)\right] \pi(\mathbf{T}_r \mid D_n) = \frac{1}{s+1} \sum_{r=1}^{s+1} \sum_{i=2}^{n} \frac{E[k_{r,i}] + 1}{i(i+\theta-1)}. \tag{7}$$

The commonly available data is in the form of Table 1, which is equivalent to $s+1$ rooted trees; here $s = |\mathbf{k}| := \sum_{i,j} k_{ij} = |\mathbf{k}_r|$ $(r = 1, \ldots, s+1)$.

The last method is to estimate t_n only by the number of mutations s, without using the information in the rooted trees.

We have the following result (proof in Appendix).

Proposition 1. *(i) One has*

$$E(t_n \mid \mathbf{k}, s, \theta) = 2\sum_{i=2}^{n} \frac{k_i + 1}{i(i+\theta-1)}; \tag{8}$$

consequently, the above estimator will diverge almost surely, if $\mathbf{k}$ *is treated as random.*

(ii) One has

$$E(t_n \mid D_n, s, \theta) = \frac{2}{s+1} \sum_{r=1}^{s+1} \sum_{i=2}^{n} \frac{E[k_{r,i}] + 1}{i(i+\theta-1)}, \tag{9}$$

and $\widehat{t}_n$ *will converge or not depending on that of the series above.*

(iii) One has

$$E(t_n \mid s, \theta) = \frac{2}{s+1} \sum_{|\mathbf{k}|=s} \sum_{i=2}^{n} \frac{E[k_i] + 1}{i(i+\theta-1)}, \tag{10}$$

and the asymptotic behavior of the above estimator depends on the series above.

Remark 2. The above result tells us that $\widehat{t}_n$ cannot be characterized by an asymptotic deterministic quantity, even for large data size. The estimator is dominated by the number of mutations in the first few coalescent times. Hence, the only practical way to infer the coalescent time is via numerical methods, as the postdata coalescent distribution has no closed form even asymptotically. In contrast, the predata mean $E(t_n) = 2(1 - 1/n) \to 2$ is convergent but is inaccurate as an estimator of the coalescence time for the population under study.

4. The Proposed Method

The method is to construct the mutation vector $k_r = (k_{r,2}, \ldots, k_{r,n})$ and compute the data probability directly from

the genealogy rooted tree $\mathbf{T}_r$'s. Suppose that there are s segregating sites in the sequence data, which is exactly the total number of mutations occurred in the history of the n sampled individuals, then there are $s+1$ different rooted trees $\mathbf{T}_r$'s compatible with the data. Each of the rooted trees is a fixed genealogy structure, with the multiplicities as the leaves, but the number of mutations among the tree segments is random, subject to the total number of mutations being s. The structure consists of the tree branches, subbranches within each branches, sub-subbranches, and so on,, and the leaves. These are the fixed features of a rooted tree. Given the data, the rooted tree is a display of how the s mutations are distributed along the lineages, but there is no time scale in the tree, so (5) cannot be used to compute the mutation probabilities. Each rooted tree tells us a partial ordering of the mutations. For example, in the rooted tree, we know mutations at sites 4, 6, and 14 occurred before the split of lineages a, b, e, and f, thus occurred before the mutations at sites 1, 5, and 10. But we do not know which of 4, 6, and 14 occurred first. We know mutation 1 occurred before 10, but we do not know the order of 1 and 5, and so forth. If we have the full data $(\mathbf{k}_r, \mathbf{w})$ corresponding to all the rooted trees, $\mathbf{T}_r$'s, we can compute $\widehat{t}_n = E(t_n \mid D_n, \theta)$ as in Proposition 1(ii). But $\mathbf{w}$ and the $\mathbf{k}_r$s are not directly available; however, $\mathbf{w}$ can be easily simulated by the prior exponential distribution, and each rooted tree $\mathbf{T}_r$ has an initial mutation distribution on its branch segments. Denote by $s_{ij\ldots}$ the $(i, j, \ldots)$th segment (the order is arbitrary, e.g., we can lable them from upper to lower and left to right locations), and let $|s_{ij\ldots}|$ be the number of mutations on it (many of them are zeros; we can concentrate on the segments with nonzero mutations). Denote $\mathbf{s} = \{s_{ij\ldots}\}$. Given $(\mathbf{w}, \mathbf{s})$, k_r can be sampled from $\mathbf{T}_r$ (to be detailed latter). Let $E_{(\mathbf{w},\mathbf{k}_r)}$ be the expectation with respect to $(\mathbf{w}, \mathbf{k}_r)$. The above motivates us to estimate t_n by

$$\widehat{t}_n = E\left(t_n \mid D_n, \theta\right) = \frac{2}{s+1} \sum_{r=1}^{s+1} \sum_{i=2}^{n} E_{(\mathbf{w},\mathbf{k}_r)} \left[\frac{k_{r,i} + 1}{i\,(i+\theta-1)} \right]. \quad (11)$$

The above expectation is not easy to compute directly since we do not know the joint distribution of $(\mathbf{w}, \mathbf{k}_r)$. Instead we use simulation method. For this, we sample $\mathbf{w}^{(1)} \cdots \mathbf{w}^{(M)}$ independently and generate $k_r^{(m)} = \{k_{r,i}^{(m)}\}$ (see below) corresponding to $\mathbf{w}^{(m)}$ and $\mathbf{T}_r$ for each r then approximate $\widehat{t}_n$ as

$$\widehat{t}_n \approx \frac{2}{M} \sum_{m=1}^{M} \frac{1}{s+1} \sum_{r=1}^{s+1} \sum_{i=2}^{n} \frac{k_{r,i}^{(m)} + 1}{i\,(i+\theta-1)}. \quad (12)$$

Now we consider generating $k_r^{(m)}$. After $w^{(m)} = (w_2^{(m)}, \ldots, w_n^{(m)})$ is allocated among the branches of $\mathbf{T}_r$, we only need to consider each segment $s_{ij\ldots}$ with nonzero number t_r of mutations in them. Each length of $s_{ij\ldots}0$ in $\mathbf{T}_r$ is the summation of some $d = d_{ij\ldots}$ of the $w_j^{(m)}$'s. For simplicity of exposition and notation, suppose that they are $w_2^{(m)}, \ldots, w_{d+1}^{(m)}$; then given the t_r mutations in $[0, w_2^{(m)} + \cdots + w_{d+1}^{(m)}]$ and using (5), it is easy to see that the number of mutations $\widetilde{k}_r = (\widetilde{k}_{r,1}, \ldots, \widetilde{k}_{r,d})$ in each of the d intervals $[0, w_2^{(m)}]$, $[w_2^{(m)}, w_2^{(m)} + w_3^{(m)}]$,..., $[w_2^{(m)} + \cdots + w_{d-1}^{(m)}, w_2^{(m)} + \cdots + w_d^{(m)}]$ follows the multinomial distribution

$$\begin{aligned} P\left(\widetilde{k}_r = (k_{r,1}, \ldots, k_{r,d}) \mid t_r\right) &= M\left(k_{r,1}, \ldots, k_{r,d}; t_r, q_1, \ldots, q_d\right) \\ &= \frac{t_r!}{k_{r,1}! \cdots k_{r,d}!} q_1^{k_1} \cdots q_d^{k_d}, \end{aligned} \quad (13)$$

where $q_j = w_{j+1}/(w_2 + \cdots + w_{d+1})$ $(j = 1, \ldots, d)$. After all the nonzero $t_r = |s_{ij\ldots}|$'s are allocated in the corresponding intervals, we have

$$k_{r,i}^{(m)} = \sum_{(i)} \widetilde{k}_{r,l}, \quad (i = 2, \ldots, n), \quad (14)$$

where the summation is for all $\widetilde{k}_{r,l}$'s that fall in $[w_2^{(m)} + \cdots + w_i^{(m)}, w_2^{(m)} + \cdots + w_{i+1}^{(m)}]$.

Specifically, the simulation method is as below. For $m = 1, \ldots, M$, do the following steps.

(i) Sample $w^{(m)} = (w_2^{(m)}, \ldots, w_n^{(m)})$ from the coalescent distribution as in (1); that is, the $w_i^{(m)}$'s are independent, with $w_i^{(m)} \sim \exp(i(i-1)/2)$. Or equivalently, sample $u \sim U(0,1)$ and set $w_i^{(m)} = -2/(i(i-1))\ln(1-u)$.

(ii) For each fixed $1 \le r \le s+1$, allocate $w^{(m)}$ to the $n-1$ coalescent events of the n sequences based on each rooted tree $\mathbf{T}_r$. See illustration below for details.

(iii) Allocate the $\widetilde{k}_r^{(m)}$ mutations in the corresponding segments according to (13). Then get the $k_{r,i}^{(m)}$'s as in (14).

After all the M iterations, evaluate (12) until convergence, which can be assessed by relative error, for example.

Illustration: Allocate $\mathbf{w}^{(m)}$ to the $n-1$ coalescent events of the n sequences based on rooted tree. We use the backward method; that is, first allocate $w_n^{(m)}$, then $w_{n-1}^{(m)}$, ..., and last $w_2^{(m)}$. Consider the rooted tree, for example. There are $n = 55$ sequences, with frequencies $(3, 1, 19, 2, 2, 1, 5, 1, 1, 1, 4, 8, 8, 3)$ for lineages $(m, n, e, b, a, f, k, c, h, g, i, j, l, d)$. Note that sequences (leaves) in each lineage (branch) only coalescence within each branch (if the branch has more than one leaves), and branch with a single leaf coalescences only at MRCA $w_2^{(m)}$. We first decide $w_{55}^{(m)}$ goes to which branch or pairs of single branches. Since it is the latest coalescent time, it can only go to a pair of leaves in some branch with multiple leaves. Since branch n has only 1 leaf; it is excluded at this step. The remaining branches $(m, \langle e, b, a, f\rangle, k, \langle c, h, g, i, j, l\rangle, d)$ all have multiple leaves with a total of 54. We assign $w_{55}^{(m)}$ to one of these branches with weights proportional to their number of leaves, that is, with probabilities $(3, 24, 5, 19, 3)/54$. Suppose that $w_{55}^{(m)}$ is assigned to $\langle e, b, a, f\rangle$; we need to decide which subbranch it goes to. We have three candidate subbranches (e, b, a) with number of leaves $(19, 2, 2)$. We randomly assign $w_{55}^{(m)}$ to them with weights $(19, 2, 2)/23$. Suppose it is assigned

to branch e; then $w_{55}^{(m)}$ will go to a pair within this branch, and which pair is irrelevant. But the pair will be treated as a single leaf in assigning the rest $w_j^{(m)}$'s. So after this step, we reassign the number of leaves in e as 18.

Now we assign $w_{54}^{(m)}$. The procedure is the same as above; the only difference is now e has 18 leaves. The candidate branches are still $(m, \langle e, b, a, f\rangle, k, \langle c, h, g, i, j, l\rangle, d)$ with weights $(3, 23, 5, 19, 3)/53$. Suppose that $w_{54}^{(m)}$ is also allocated to e of branch $\langle e, b, a, f\rangle$; then e has 17 leaves now.

We now allocate $w_{53}^{(m)}$ to candidates $(m, \langle e, b, a, f\rangle, k, \langle c, h, g, i, j, l\rangle, d)$ with weights $(3, 22, 5, 19, 3)/52$. Suppose that $w_{53}^{(m)}$ is allocated to $\langle c, h, g, i, j, l\rangle$; we need to decide which of the 4 subbranches it will go to. c has only 1 leaf and is excluded. So we allocate subbranches $(\langle h, g\rangle, \langle i, j\rangle, l)$ with weights $(2, 12, 4)/18$. Supposing it goes to $\langle h, g\rangle$, since it has only one pair of leaves, then h and g are merged as one leaf after this assignment.

Continue this way, until $w_2^{(m)}$ is allocated. Then all the branches in this rooted tree have lengths as the $w_i^{(m)}$'s allocated to them. After this step, the length of each segment $s_{ij\ldots}$ of $\mathbf{T}_r$ is a summation of some $w_i^{(m)}$'s. Since $|s_{ij\ldots}|$ is known from each $\mathbf{T}_r$, we can allocate each of the $\widetilde{k}^{(m)}$'s by (13), then get the $k_i^{(m)}$'s by the formula that follows it. Then compute (12).

The assumption that the population size N is constant can be relaxed the same way as in GS and Tavaré et al. [10].

Appendix

Proof of the Proposition

(i) Recall that the k_{ij}'s are independent, the k_i's are independent, and the w_i's are independent with $w_i \sim \exp(i(i-1)/2)$, $E(w_i) = 2/(i(i-1)/2)$, $k_{ij} \mid w_i \sim \mathrm{Po}(\cdot, w_i\theta/2)$, $k_i|w_i \sim \mathrm{Po}(\cdot, iw_i\theta/2)$, and so $E(k_i) = E(E(k_i \mid w_i)) = \theta/(i-1)$. Observe

$$\begin{aligned} P(w_i \mid D_n, \theta) &= P(w_i \mid \mathbf{k}, \theta) = P(w_i \mid k_i, \theta) \\ &\propto P(w_i)\, P(k_i \mid w_i, \theta) \\ &= \frac{i(i-1)}{2} \exp\left(-\frac{i(i-1)\, w_i}{2}\right) \\ &\quad \times \mathrm{Po}\left(k_i, \frac{iw_i\theta}{2}\right) \\ &\propto (w_i\theta)^{k_i} \exp\left(-\frac{i(i+\theta-1)\, w_i}{2}\right). \end{aligned} \tag{A.1}$$

The right-hand side above is the density of $\Gamma(k_i+1, 2/(i(i+\theta-1))$ distribution, up to an normalizing constant. It has mean $E(w_i \mid \mathbf{k}, \theta) = 2(k_i+1)/(i(i+\theta-1))$ and variance $\mathrm{Var}(w_i \mid \mathbf{k}, \theta) = 4(k_i+1)/(i^2(i+\theta-1)^2)$. Since $t_n = \sum_{i=2}^n w_i$, we have

$$\begin{aligned} E(t_n \mid \mathbf{k}, \theta) &= \sum_{i=1}^n E(w_i \mid \mathbf{k}, \theta) \\ &= 2\sum_{i=2}^n \frac{k_i+1}{i(i+\theta-1)} \\ &= S_n \sum_{i=2}^n a_{n,i} x_i + b_n, \end{aligned} \tag{A.2}$$

where $x_i = (i-1)k_i$, the x_i's are i.i.d with $E(x_i) = \theta$, $a_{n,i} = a_{n,i}(\theta) = S_n^{-1}(\theta)2/[i(i-1)(i+\theta-1)]$, $S_n = S_n(\theta) = 2\sum_{i=2}^n 1/[i(i-1)(i+\theta-1)]$, and $b_n(\theta) = 2\sum_{i=2}^n 1/[i(i+\theta-1)]$. Note that $\{S_n(\theta)\}$ and $\{b_n(\theta)\}$ are convergent sequences with

$$\begin{aligned} S_n(\theta) &\longrightarrow S(\theta) := \sum_{i=2}^{\infty} \frac{2}{i(i-1)(i+\theta-1)} < \infty, \\ b_n(\theta) &\longrightarrow b(\theta) := \sum_{i=2}^{\infty} \frac{2}{i(i+\theta-1)} < \infty. \end{aligned} \tag{A.3}$$

Note also that $\sum_{i=2}^n a_{n,i} = 1$, that is, $\{a_{n,i}\}$ is a weight sequence for each fixed n. Since $\lim_n a_{n,i} > 0$ for fixed i, by the results for weighted sum of i.i.d. random variables (see [19], for a review of such results), a necessary condition for $\sum_{i=2}^n a_{n,i}x_i$ to converge (a.s.) is that $\lim_n a_{n,i} = 0$ for all fixed i, or no term will be dominant as $n \to \infty$. Since this condition is not satisfied, the first few terms are dominant and we have

$$\begin{aligned} &\sum_{i=2}^n a_{n,i} x_i \text{ diverges (a.s.)}, \\ &\text{or equivalently } E(t_n \mid \mathbf{k}, \theta) \text{ diverges (a.s.)}. \end{aligned} \tag{A.4}$$

The proofs of part (ii) and (iii) are similar to that of part (i) and omitted.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] A. Yuan, Z. Zhang, G. Liu, and G. Chen, "On the estimation of mutation rate based on coalescence genealogy," *Journal of Advanced Bioinformatics Applications and Research*. In press.

[2] R. C. Griffiths and S. Tavaré, "Ancestral inference in population genetics," *Statistical Science*, vol. 9, no. 3, pp. 307–319, 1994.

[3] R. H. Ward, B. L. Frazier, K. Dew-Jager, and S. Pääbo, "Extensive mitochondrial diversity within a single Amerindian tribe," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 19, pp. 8720–8724, 1991.

[4] R. R. Hudson, "Gene genealogies and the coalescent process," in *Oxford Surveys in Evolutionary Biology*, D. Futuyma and J. Antonovics, Eds., pp. 1–44, Oxford University Press, New York, NY, USA, 1991.

[5] P. Donnelly and S. Tavaré, "Coalescents and genealogical structure under neutrality," *Annual Review of Genetics*, vol. 29, pp. 401–421, 1995.

[6] J. F. C. Kingman, "On the genealogy of large populations," *Journal of Applied Probability*, vol. 19, pp. 27–43, 1982.

[7] J. F. C. Kingman, "Exchangeability and the evolution of large populations," in *Exchangeability in Probability and Statistics*, G. Koch and F. Spizzchino, Eds., pp. 97–112, North-Holland, Amsterdam, The Netherlands, 1982.

[8] S. Tavaré, "Ancestral inference from DNA sequence data," in *Case Studies in Mathematical Modeling: Ecology, Physiology and Cell Biology*, H. G. Othmer, F. R. Adler, M. A. Lewis, and J. Dallon, Eds., chapter 5, pp. 81–96, Prentice Hall, New York, NY, USA, 1997.

[9] G. A. Watterson, "Mutant substitutions at linked nucleotide sites," *Advances in Applied Probability*, vol. 14, no. 2, pp. 206–224, 1982.

[10] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly, "Inferring coalescence times from DNA sequence data," *Genetics*, vol. 145, no. 2, pp. 505–518, 1997.

[11] S. N. Ethier and R. C. Griffiths, "The infinitely-many-sites model as a measure valued diffusion," *Annals of Probability*, vol. 15, no. 2, pp. 515–545, 1987.

[12] J. Camin and R. Sokal, "A method for deducing branching sequences in phylogeny," *Evolution*, vol. 19, no. 3, pp. 311–326, 1965.

[13] J. S. Farris, "Inferring phylogenetic trees from chromosome inversion data," *Systematic Zoology*, vol. 27, pp. 275–284, 1967.

[14] K. S. Booth and G. S. Lueker, "Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms," *Journal of Computer and System Sciences*, vol. 13, no. 3, pp. 335–379, 1976.

[15] J. Felsenstein, "Numerical methods for inferring evolutionary trees," *The Quarterly Review of Biology*, vol. 57, no. 4, pp. 379–404, 1982.

[16] D. Gusfield, "Efficient algorithms for inferring evolutionary trees," *Networks*, vol. 21, no. 1, pp. 19–28, 1991.

[17] R. C. Griffiths, "An algorithm for constructing genealogical trees," Statistics Research Report 163, Department of Mathematics, Monash University, Melbourne, Australia, 1987.

[18] R. C. Griffiths and S. Tavaré, "Unrooted genealogical tree probabilities in the infinitely-many-sites model," *Mathematical Biosciences*, vol. 127, no. 1, pp. 77–98, 1995.

[19] N. H. Bingham, "Extensions of the strong law," *Advances in Applied Probability*, pp. 27–36, 1986.