Statistical Science
1994, Vol. 9, No. 3, 307–319

Ancestral Inference in Population Genetics

R. C. Griffiths and Simon Tavaré

Abstract. Mitochondrial DNA sequence variation is now being used to study the history of our species. In this paper we discuss some aspects of estimation and inference that arise in the study of such variability, focusing in particular on the estimation of substitution rates and their use in calibrating estimates of the time since the most recent common ancestor of a sample of sequences.

Observed DNA sequence variation is generated by superimposing the effects of mutation on the ancestral tree of the sequences. For data of the type studied here, this ancestral tree has to be modeled as a random process. Superimposing the effects of mutation produces complicated sampling distributions that form the basis of any statistical model for the data. Using such distributions—for example, for maximum likelihood estimation of rates—poses some difficult computational problems. We describe a Monte Carlo method, a cousin of the popular "Markov chain Monte Carlo," that has proved very useful in addressing some of these issues.

Key words and phrases: Coalescent, ancestral inference, mitochondrial Eve, infinitely-many-sites, mitochondrial DNA, Markov chain Monte Carlo, Monte Carlo likelihoods.

1. INTRODUCTION

Recent advances in molecular biology have provided rapid and accurate methods for sequencing DNA from many different organisms. This has lead to the accumulation of enormous amounts of DNA sequence information from many different species. Comparison of homologous regions of DNA among species has been used to infer the ancestral relationships of the species, while comparison of sequence information from individuals within a species can be used to infer aspects of the evolutionary history of that species. These latter studies have been most prevalent in the context of human evolution, where one particular molecule, mitochondrial DNA, has had a profound influence.

Human mitochondrial DNA, first sequenced by Anderson et al. (1981), is a circular double-stranded molecule about 16,500 base pairs in length, containing genes that code for 13 proteins, 22 tRNA genes and 2 rRNA genes. Mitochondria live outside the nucleus of cells, where they play a role in oxidative phosphorylation and ATP synthesis. One part of the molecule, the control region (sometimes referred to as the D-loop), has received particular attention. This region is about 1,100 base pairs in length and contains promoters for transcription and the origin of replication for one of the DNA strands.

As the mitochondrial molecule evolves, mutations result in the substitution of one of the bases A, C, G or T in the DNA sequence by another one. Transversions, those changes between purines (A, G) and pyrimidines (C, T), are less frequent than transitions, the changes that occur between purines or between pyrimidines.

It is known that base substitutions accumulate extremely rapidly in mitochondrial DNA, occurring at about 10 times the rate of substitutions in nuclear genes. The control region has an even higher rate, perhaps an order of magnitude higher again. This high mutation rate makes the control region a useful molecule with which to study DNA variation over relatively short time spans, because sequence differences will be found among closely related individuals. In addition, mammalian mitochondria are almost exclusively maternally inherited, which makes these molecules ideal for studying the maternal lineages in which they arise. This simple mode

R. C. Griffiths is Reader, Department of Mathematics at Monash University, Clayton 3168, Australia. Simon Tavaré is Professor, Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089-1113.

of inheritance means that recombination is essentially absent, making inferences about molecular history somewhat simpler than in the case of nuclear genes.

Mitochondrial sequence variation has been used to study the history of our species. One such analysis led Cann, Stoneking and Wilson (1987) to suggest that all living mitochondria descend from a single ancestor who lived approximately 200,000 years ago in Africa. While the precise details of the hypothesis have been the subject of much heated debate [summarized, e.g., by Stoneking (1993)], there is general agreement about the power of the methodology. Another example focusing on human history concerns the origin of New World natives. Although it is generally believed that they have an Asian origin, there is still much debate about the number, time and composition of migrations into the New World. See Schurr et al. (1990), Torroni et al. (1992), Shields et al. (1993) and Ward et al. (1993). Archaeological and linguistic data have been used to study such migrations, but molecular data provide an evolutionary framework from which quantities such as times of divergence can be inferred.

In this paper, we focus on mitochondrial data sampled from a single North American Indian tribe, the Nuu-Chah-Nulth from Vancouver Island. Based on the archaeological record (cf. Dewhirst, 1978), it is clear that there is remarkable cultural continuity from earliest levels of occupation to the latest. This implies not only that there was no significant immigration into the area by other groups, but that the subsistence pattern and presumably the demographic size of the population has also remained roughly constant for at least 8,000 years. Based on the current size of the population that was sampled, there are approximately 600 women of childbearing age in the traditional Nuu-Chah-Nulth population.

The original data, appearing in Ward et al. (1991), comprised a sample of mtDNA sequences from 63 individuals. The sample approximates a random sample of individuals in the tribe, to the extent to which this can be experimentally arranged. Each sequence is the first 360 base pair segment of the control region, corresponding to positions 16,024–16,383 in the human reference sequence of Anderson et al. (1981). By convention, the characteristic attributes of the sequence data are defined in terms of the "light" strand. With reference to the light strand, this region comprises 201 pyrimidine sites and 159 purine sites; 21 of the pyrimidine sites are variable (or segregating), that is, not identical in all 63 sequences in the sample. In contrast, only five of the purine sites are variable. There are 28 distinct DNA sequences (hereafter called lineages) in the data. Because no transversions are seen in these data each DNA site is binary, having just two possible bases at that site. Furthermore, because there is no recombination each site in the sample has the same ancestral history.

The purpose of this paper is to describe some statistical approaches to understanding this history. In particular, we would like to estimate the rate at which substitutions accumulate in this region; to uncover evidence of population size fluctuations and possibly geographic subdivision; and to infer something about ancestral features of the population such as the distribution of the time to the most recent common ancestor of the sample.

In order to make statistical statements about these issues, we have to model several features of the data. First, and perhaps most important, we have to recognize that the sampling variation observed in the sequences comes from highly dependent data. This dependence comes from the fact that individuals in the sample are correlated because of their common ancestry. Second, this common ancestry is random in at least two respects: different samples produce different ancestries, each providing a snapshot of different parts of the ancestral tree linking all the individuals in the history of the population, and this population tree is itself one (and our only) run of the evolutionary process. Unfortunately, in population studies such as these we cannot observe the underlying ancestry of the sample, and so we must resort to a stochastic description of it. The natural time scale of the questions we study places the emphasis between real-time pedigrees in the human genetics arena and much longer time scale problems concerned with inferring ancestral relationships among distantly diverged species. The basic techniques we use come from population genetics.

Given a plausible model for the ancestral relationships among the molecules in the sample, we then have to superimpose the effects of mutation, the process that is ultimately responsible for the variation we see in the sample. This leads us to a statistical description of the variability that can be used to estimate population parameters, such as the substitution rate, and for ancestral inference.

To keep the presentation simple, we focus on one part of the data that seems to have a relatively simple mutation structure. As discussed later, we shall assume that substitutions at any nucleotide position can occur only once in the ancestry of the molecule. Hence we have eliminated lineages (i.e., distinct DNA sequences) in which substitutions are observed to have occurred more than once. The resulting subsample comprises 55 of the original 63 sequences, and 352 of the original 360 sites. Eight of the pyrimidine segregating sites were removed, resulting in a set of 18 segregating sites in all; 13 of these sites are pyrimidines, and 5 are purines. These

TABLE 1
*Nucleotide position in control region**

Position	1 0 6	1 9 0	2 5 1	2 9 6	3 4 4	8 8	9 1	1 2 4	1 4 9	1 6 2	1 6 6	1 9 4	2 3 3	2 6 7	2 7 1	2 7 5	3 1 9	3 3 9	
Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Lineage 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

*Mitochondrial data from Ward et al. (1991, Figure 1). Variable purine and pyrimidine positions in the control region. Position 69 corresponds to position 16,092 in the human reference sequence published by Anderson et al. (1981).

data are given in Table 1, subdivided into sites containing purines and pyrimidines. Each row of the table represents a distinct DNA sequence, and the frequencies of these lineages are given in the rightmost column of the table.

The layout of this paper is as follows. In Section 2 we give a description of the coalescent, a stochastic model often used by population geneticists to approximate the random ancestral relationships among a sample of molecules. In Section 3, we discuss the so-called infinitely-many-sites assumption that provides the simplest description of the molecular variability seen in a sample, and we show how these two features can be combined to give sampling theory which takes into account the effects of size variation in the ancestral populations. In Section 4, we describe a useful computational device, a type of Markov chain Monte Carlo method, for computing these sampling distributions. In Section 5, we use this method to estimate substitution rates in the region. In Section 6, we study the distribution of the time to the most recent common ancestor of the sample for plausible values of the substitution rate, and we infer something about the type of that ancestor. The paper closes with a discussion in Section 7.

2. THE COALESCENT

We have noted that mitochondrial DNA is maternally inherited and that mtDNA molecules therefore provide a way to study the ancestry of the females in which they arise. The first task is to model this ancestry. Suppose then that we have taken a random sample of size n from the present generation of individuals. Think of the sample as females; every male in the sample should be taken as a surrogate for his mother. Assume for the moment that they were sampled from a population (of females) that has been of approximately constant size N for many generations into the past. We impose reproductive neutrality by supposing that the joint distribution of the number of (female) offspring born to each individual is exchangeable, and identical in different generations. If we let ν denote the number of offspring born to a typical individual, then exchangeability guarantees that $\mathbb{E}\nu = 1$; we denote the variance of ν by σ_N^2. Notice that in the ancestral description of the population, many individuals in a given generation may share a common parent, and the structure of the ancestral process may therefore be very complicated. Kingman (1982a) introduced the coalescent as a continuous-time approximation, obtained in the limit of large population size, to this ancestral process.

Kingman (1982b) also provided a very useful invariance principle that shows that essentially all the exchangeable reproductive models can be approximated by this coalescent. Specifically, if time is measured in units of $\sigma^{-2}N$ generations, where $0 < \sigma^2 \equiv \lim_{N\to\infty} \sigma_N^2 < \infty$, then in the limit as $N \to \infty$ the ancestral process of the discrete model converges in distribution to the coalescent. In the population genetics literature it is common to assume that $\sigma^2 = 1$, corresponding to the approximation of the celebrated Wright–Fisher model.

The coalescent has a very simple structure. In this continuous-time approximation, ancestral lines

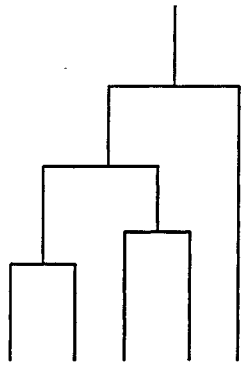

FIG. 1. *Coalescent tree.*

going backward in time coalesce when they have a common ancestor. Note that coalescences occur only between pairs of individuals. This process may also be thought of as generating a binary tree, with leaves representing the sample sequences and vertices where ancestral lines coalesce. The root of the tree is the most recent common ancestor (MRCA) of the sample sequences. Figure 1 illustrates a coalescent tree for five individuals.

It remains to describe the stochastic structure of this tree. This is made up of two parts: the time scale that determines the rate at which coalescences occur, and the topology that determines who is related to whom. Let $A_n(t)$ denote the number of distinct ancestors that the sample has t time units back in the past. The random process $\{A_n(t), t \geq 0\}$ is a pure death process that moves from state k to state $k-1$ at rate $k(k-1)/2$, and individuals are joined at random when coalescences occur.

The time T_j during which the sample has j distinct ancestors has an exponential distribution with parameter $j(j-1)/2$ and times for different j are independent. The time T_{MRCA} until the common ancestor of the sample is

$$T_{\text{MRCA}} = T_n + T_{n-1} + \cdots + T_2, \tag{1}$$

and hence

$$\mathbb{E}T_{\text{MRCA}} = 2\left(1 - \frac{1}{n}\right). \tag{2}$$

To allow for variable population sizes, suppose that, relative to the population size N at time 0 (the time of sampling), the size of the population time t units ago is $v(t)$. In this approximation we are assuming that all the past population sizes have been large. In the Wright–Fisher case, time is once more measured in units of N generations. The topology of the tree is just as before, but its time scale has to be changed to account for the fluctuations in population size. To do this, define

$$\Lambda(t) = \int_0^t \frac{ds}{v(s)}, \quad t \geq 0. \tag{3}$$

The process $\{\tilde{A}_n(t), t \geq 0\}$ giving the number of distinct ancestors of the sample t time units ago is then a nonhomogeneous death process whose distribution can be defined by

$$\tilde{A}_n(t) = A_n\big(\Lambda(t)\big), \quad t \geq 0. \tag{4}$$

If the population has been contracting as we look back into the past, then $v(t) \leq 1$ so that $\Lambda(t) \geq t$, from which it follows that $\tilde{A}_n(t) \leq A_n(t)$. Therefore, for any $k = 1, \ldots, n$ and for any $t \geq 0$, we have $\mathbb{P}(\tilde{A}_n(t) > k) \leq \mathbb{P}(A_n(t) > k)$. This stochastic ordering corresponds to the observation that it should take less time to find a common ancestor in a small population than in a large one.

3. SAMPLING DISTRIBUTIONS

Our first aim is to describe the process of substitutions that have occurred in the North American Indian mitochondrial sequences discussed in Section 1. To do this, we have to describe how mutations (in this case, substitutions of one base for another) can be superimposed on the ancestral coalescent tree. This is made up of two parts: one that records where mutations in the lineages occur; the second, what the effect of each mutation is. It is usual to assume that conditional on the ancestral tree, mutations occur at the points of Poisson processes of rate $\theta/2$, independently in each branch of the tree. In terms of the underlying discrete process, we are assuming that a mutation occurs with probability u per sequence per generation and that $\theta = 2Nu$, where N is the size of the population from which the sample was drawn.

The second part depends on the level of detail that is to be assumed about the effects of each mutation. Since we are just modeling substitutions, we need to specify the probabilities with which a mutation in a sequence changes a particular position, and the probabilities with which that particular base is changed to other bases. Despite the apparent simplicity of the model, it covers cases in which rates at different sites vary and so can be made to model the effects of hot spots and invariable sites, those positions which change very rapidly or not at all. More generally, the model can allow for complicated interactions between substitutions at different positions along the sequences and for recurrent mutations, those which occur at a particular site more than once in the history of the molecules.

In this article we describe the simplest possible substitution process, the so-called infinitely-many-sites model (cf. Watterson, 1975). In particular we assume that when a substitution occurs in a sequence it is at a location that has never seen a substitution before. This model arises as an approximation to the evolution of a sample of sequences of finite length, when it is assumed that sites at which recurrent mutations have occurred are rare and may be ignored. In practice most sequence data exhibit recurrent mutations, and we therefore have to select a subset of sites and individuals to which the simpler model applies. We saw earlier that in the full mitochondrial data set of 63 sequences, there are sites at which recurrent mutation has occurred, and we described how a reduced set (of 352 sites from 55 individuals) that is consistent with the infinitely-many-sites model was chosen. The variable sites in this reduced set of sequences are given in Table 1. In Section 5, we try to assess how this data selection has influenced our estimates of substitution rates.

What structure do these sites have? Because of the infinitely-many-sites assumption, the pattern of segregating sites tells us something about the mutations that have occurred in the history of the sample. Since mutations can occur only once at a given site, there is an ancestral type and a mutant type at each segregating site. For the moment assume we know which is which, and label the ancestral type as 0 and the mutant type as 1. To fix ideas, take each column of the data in Table 1 and label the most commonly occurring base as 0, the other as 1. The data can therefore be thought of as a matrix of 0's and 1's, with multiplicities for each distinct row (or lineage). This matrix can be represented as a rooted tree by labeling each distinct row by a sequence of mutations up to the common ancestor. These mutations are the vertices in the tree. This rooted tree is a condensed description of the coalescent tree with its mutations, and it has no time scale in it. It is convenient to label the root 0, even though it does not represent a mutation in the tree (in fact, it represents the first mutation occurring to the ancestors of the MRCA). The data sequences can also be thought of as the incidence matrix of the mutations occurring in the paths to the root. The information in the data is equivalent to the information in such a condensed tree. Algorithms for producing these trees are detailed in Griffiths (1987) and Gusfield (1991), for example. Felsenstein (1982) discusses where such trees arise in the systematics literature.

Figures 2 and 3 illustrate the connection between coalescent and condensed trees. Dots represent where mutations have taken place, and each lineage is represented by just a single line. For the site labeling mentioned above, the data in Table 1 are equivalent to the condensed tree shown in Figure 3. Since there is no time scale in these condensed trees, many topologically different coalescent trees may produce the given condensed tree. The coalescent tree shown in Figure 2 is one of many which produce the tree in Figure 3.

An alternative way of describing the condensed trees is by listing the mutation paths of each lineage backward in time to the root, together with the multiplicities of each lineage, as in Table 2. We think of a tree $(T, \mathbf{n})$ with d distinct sequences as being a listing of these paths together with multiplicities of types $\mathbf{n} = (n_1, \ldots, n_d)$.

Of course, in practice we never know which type at a site is ancestral. All that can be deduced then from the data is the number of segregating sites between each pair of sequences. In this case the data is equivalent to an *unrooted* tree whose vertices represent distinct lineages and whose edges are labeled by mutations between lineages. The ordering of mutations between lineages is not unique.

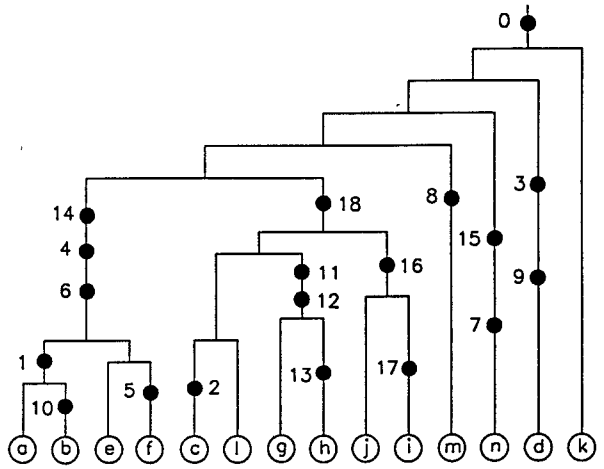

FIG. 2. *Possible coalescent tree for mtDNA data.*

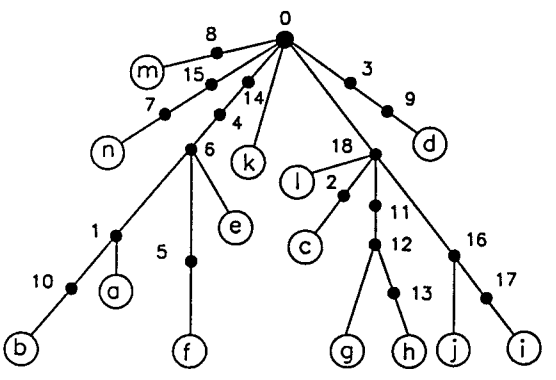

FIG. 3. *Rooted genealogical tree for mtDNA data.*

TABLE 2
*Mutation paths for the rooted tree in Figure 1**

Lineage							Lineage frequencies
a	1	6	4	14	0		2
b	10	1	6	4	14	0	2
c	2	18	0				1
d	9	3	0				3
e	6	4	14	0			19
f	5	6	4	14	0		1
g	12	11	18	0			1
h	13	12	11	18	0		1
i	17	16	18	0			4
j	16	18	0				8
k	0						5
l	18	0					4
m	8	0					3
n	7	15	0				1

*Numbers correspond to site labels in Table 1.

The unrooted tree can be constructed from any given rooted tree by reorganizing so that vertices represent lineages (rather than mutations), and mutations are along the edges. If the root lineage is not in the sample, it does not appear in the unrooted tree. The unrooted tree corresponding to the rooted tree in Figure 3 is shown in Figure 4. All possible rooted trees may be found from an unrooted tree by placing the root at a vertex or between mutations, then reading off mutation paths between lineages and the root. These paths are then paths from the leaves to the root in the rooted tree whose vertex labels are mutations. If there are s segregating sites, there will be $s + 1$ rooted trees that correspond to it. Each of these corresponds to a labeling of which type at each site is ancestral and which is mutant. For our example, there are 19 rooted trees that correspond to the unrooted tree in Figure 4. Some of the internal vertices in the unrooted tree may be inferred sequences, ones that are not represented in the sample. Further details about the construction of these trees may be found in Griffiths and Tavaré (1994b). The infinitely-many-sites assumption implies that either a rooted tree or an unrooted tree can be constructed uniquely from a collection of data sequences.

We have seen that the data may be represented as an unrooted tree $\mathbf{Q}$ with d vertices representing distinct sequences appearing in the sample, these vertices having multiplicities $\mathbf{n} = (n_1, \ldots, n_d)$. We would like to be able to calculate the probability $p^0(\mathbf{Q}, \mathbf{n})$ of the data under these models for a variety of parameter values. [The zero superscript in p^0 is used so that the notation agrees with that of Griffiths and Tavaré (1994b).]2 This would provide a way to use likelihood methods for parameter estimation and for inference about such parameters. The sampling distribution $p^0(\mathbf{Q}, \mathbf{n})$ can be found by first computing

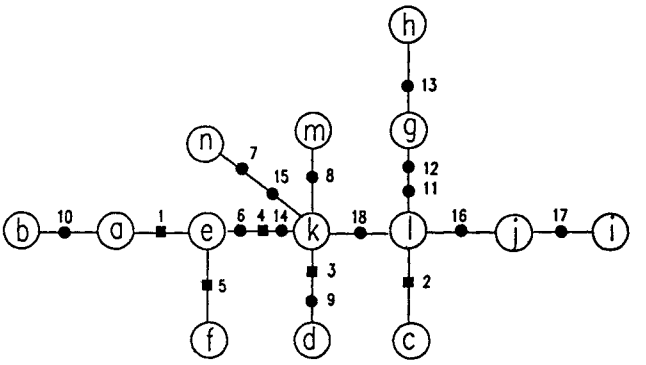

FIG. 4. *Unrooted genealogical tree for mtDNA data: • is pyrimidine site, ■ is purine site.*

the probability $p^0(T, \mathbf{n})$ of each of the $s + 1$ *rooted* trees T corresponding to the given unrooted tree, and summing:

$$p^0(\mathbf{Q}, \mathbf{n}) = \sum_T p^0(T, \mathbf{n}). \tag{5}$$

To calculate the rooted tree probabilities, we use the branching structure of the model to derive a recursion satisfied by these probabilities. The recursive nature of the method stems from the fact that we can look back up the coalescent tree to the most recent event in the past (it is either a coalescence or a mutation) and see what type of tree we must have had at that event to produce the tree that represents the data now. In the case of a constant-population-size process [$\Lambda(t) = t$], this produces the recursion

$$\begin{aligned} &n(n-1+\theta)p^0(T, \mathbf{n}) \\ &\quad = \sum_{k: n_k \geq 2} n(n_k - 1)p^0(T, \mathbf{n} - \mathbf{e}_k) \\ &\quad\quad + \theta \sum_{\substack{k: n_k = 1, x_{k0} \text{ distinct}, \\ \mathcal{S}\mathbf{x}_k \neq \mathbf{x}_j \,\forall\, j}} p^0(\mathcal{S}_k T, \mathbf{n}) \\ &\quad\quad + \theta \sum_{\substack{k: n_k = 1, \\ x_{k0} \text{ distinct}}} \sum_{j: \mathcal{S}\mathbf{x}_k = \mathbf{x}_j} (n_j + 1) \\ &\quad\quad\quad \cdot p^0\big(\mathcal{R}_k T, \mathcal{R}_k(\mathbf{n} + \mathbf{e}_j)\big). \end{aligned} \tag{6}$$

In equation (6), $\mathbf{e}_j$ is the jth unit vector, $\mathcal{S}$ is a shift operator which deletes the first coordinate of a path, $\mathcal{S}_k T$ deletes the first coordinate of the kth path of T, $\mathcal{R}_k T$ removes the kth path of T and "x_{k0} distinct" means that $x_{k0} \neq x_{ij}$ for all $(\mathbf{x}_1, \ldots, \mathbf{x}_d)$ and $(i, j) \neq (k, 0)$. The boundary condition is $p(T_1, \mathbf{e}_1) = 1$. Defining the degree of $(T, \mathbf{n})$ as $\{n - 1 +$ the number of vertices in $T\}$, we note that the system (6) is recursive in the degree of $(T, \mathbf{n})$.

The first term in (6) corresponds to the most recent event in the past being a coalescence [with probability $(n - 1)/(n - 1 + \theta)$], the second and

third to the event being a mutation [with probability $\theta/(n-1+\theta)$]. If the last event was a mutation, the lineage with this mutation is necessarily a singleton in the sample. In the second term, removing the last mutation from a lineage leaves the lineage still as a singleton in the data, but in the third term the lineage with the mutation removed is identical to another in the sample. For each singleton path in T with a distinct first coordinate, there is exactly one nonzero entry in the second and third terms.

A more detailed discussion and derivation appears in Griffiths and Tavaré (1994b). In the case of variable population size the recursion is no longer discrete, but rather has the form of an integro-recurrence (see Griffiths and Tavaré, 1994c).

4. A MARKOV CHAIN MONTE CARLO METHOD

For samples of the size we have here, computing probabilities like $p^0(Q, \mathbf{n})$ by solving recursions like (6) directly is computationally prohibitive. Instead we adopt a Monte Carlo approach that uses the recursion to construct a Markov chain in such a way that the probability of interest can be represented as the expected value of a functional along a sample path of this chain.

In general terms, the method is as follows. Any recursion like (6) can be written in the form

$$q(x) = \sum_{y \in \mathcal{A}} r(x, y) q(y) + \sum_{y \in \mathcal{B}} r(x, y) q(y), \quad x \in \mathcal{B}, \tag{7}$$

where $\mathcal{A}$ denotes the set of x values for which $q(x)$ is known; $\mathcal{B}$ denotes the set where it is unknown; and the kernel $r(x, y) \geq 0$. Let $p(x, y)$ be the transition matrix of a Markov chain X on $\mathcal{A} \cup \mathcal{B}$ satisfying $p(x, y) > 0$ whenever $r(x, y) > 0$, with the property that X visits $\mathcal{A}$ with probability 1 starting from any $x \in \mathcal{B}$. Define

$$h(x, y) = \frac{r(x, y)}{p(x, y)}, \quad x, y \in \mathcal{A} \cup \mathcal{B}.$$

Iterating the equation in (7), we see that, for $x \in \mathcal{B}$,

$$\begin{aligned} q(x) = & \sum_{y \in \mathcal{A}} p(x, y) h(x, y) q(y) \\ & + \sum_{y_1 \in \mathcal{B}} \sum_{y \in \mathcal{A}} p(x, y_1) p(y_1, y) h(x, y_1) h(y_1, y) q(y) \\ & + \sum_{y_1 \in \mathcal{B}} \sum_{y_2 \in \mathcal{B}} \sum_{y \in \mathcal{A}} p(x, y_1) p(y_1, y_2) p(y_2, y) h(x, y_1) \\ & \qquad \cdot h(y_1, y_2) h(y_2, y) q(y) + \cdots. \end{aligned}$$

It follows that

$$q(x) = \mathbb{E}_x q(X_\tau) \prod_{j=1}^{\tau} h(X_{j-1}, X_j), \tag{8}$$

where τ is the time when $\mathcal{A}$ is first visited by X. While there is some flexibility in the choice of $p(x, y)$, it is convenient to take

$$\begin{gathered} p(x, y) = \frac{r(x, y)}{f(x)}, \quad f(x) \equiv \sum_{y \in \mathcal{A} \cup \mathcal{B}} r(x, y), \\ h(x, y) = f(x). \end{gathered} \tag{9}$$

Multiple independent simulations of the Markov chain X starting from $X(0) = x$ may then be used to provide estimates of $q(x)$. This method is indeed a type of Markov chain Monte Carlo method, albeit of a rather different type than its better-known cousin.

For the case of (6), Griffiths and Tavaré (1994b) show how to construct the appropriate Markov chain $\{X(l),\ l = 0, 1, 2, \ldots\}$. The chain has a tree state space, with states $x = (T, \mathbf{n})$, and makes transitions as follows:

$$(T, \mathbf{n}) \to \begin{cases} (T, \mathbf{n} - \mathbf{e}_k), \\ (\mathcal{S}_k T, \mathbf{n}), \\ (\mathcal{R}_k T, \mathcal{R}_k(\mathbf{n} + \mathbf{e}_j)), \end{cases} \tag{10}$$

$$\text{with probability } \begin{cases} \dfrac{n_k - 1}{f(T, \mathbf{n})(n + \theta - 1)}, \\ \dfrac{\theta}{f(T, \mathbf{n}) n (n + \theta - 1)}, \\ \dfrac{\theta(n_j + 1)}{f(T, \mathbf{n}) n (n + \theta - 1)}, \end{cases}$$

for $k = 1, 2, \ldots, d$. The scaling factor is

$$f(T, \mathbf{n}) = \sum_{k=1}^{d} \frac{(n_k - 1)}{(n + \theta - 1)} + \frac{\theta m}{n(n + \theta - 1)},$$

where m is defined by

$$\begin{aligned} m = & \left| \{k\colon n_k = 1,\ x_{k,0} \text{ distinct}, \mathcal{S}\mathbf{x}_k \neq \mathbf{x}_j\ \forall\, j\} \right| \\ & + \sum_{k\colon n_k = 1,\ x_{k,0} \text{ distinct}} \ \sum_{j\colon \mathcal{S}\mathbf{x}_k = \mathbf{x}_j} (n_j + 1). \end{aligned}$$

The X process starts from an initial tree $(T, \mathbf{n})$ and runs backward in time until the time τ at which there is a tree $(T_1, \mathbf{e}_1)$ corresponding to a single root sequence. The process always moves toward $(T_1, \mathbf{e}_1)$, in the sense that the degree decreases by 1 at each move. The representation of $p^0(T, \mathbf{n})$ follows from (8) and (9) in the form

$$p^0(T, \mathbf{n}) = \mathbb{E}_{(T, \mathbf{n})} \left[\prod_{l=0}^{\tau - 1} f\Big(\big(T(l), \mathbf{n}(l)\big) \Big) \right], \tag{11}$$

where $X(l) \equiv (T(l), \mathbf{n}(l))$ is the tree at time l.

Equation (11) may be used to produce an estimate of $p^0(T, \mathbf{n})$ by simulating g independent copies of the tree process $\{X(l),\ l = 0, 1, \ldots\}$ and computing for each run j the value F_j of the functional along the sample path:

$$F_j \equiv \left[\prod_{l=0}^{\tau-1} f\Big(\big(T(l), \mathbf{n}(l)\big)\Big)\right]. \tag{12}$$

The average of the F_j over all g runs is then an unbiased estimator of $p^0(T, \mathbf{n})$. An estimate of $p^0(\mathbf{Q}, \mathbf{n})$ can then be found by summing over all rooted trees. The technique can be modified by changing the stopping time τ. Stopping the chain earlier and computing the probability $p^0(T, \mathbf{n})$ at that point explicitly results in both time and variance reduction (see Griffiths and Tavaré, 1994b).

The method can be modified by using importance sampling to compute $p^0(T, \mathbf{n})$ for fixed $(T, \mathbf{n})$ as a function of θ, from a single realization of the process $\{X(l),\ l = 0, 1, \ldots\}$. This produces a Monte Carlo approximant to the likelihood surface of interest. We proceed as follows: simulate the chain $\{X(l),\ l = 0, 1, \ldots\}$ with a particular value θ_0 as parameter, and obtain the likelihood surface for other values of θ using the representation

$$\begin{aligned} p^0_\theta(T, \mathbf{n}) &= \mathbb{E}^{\theta_0}_{(T,\mathbf{n})}\left[\prod_{l=0}^{\tau-1} h\Big(\big(T(l), \mathbf{n}(l)\big), \big(T(l+1), \mathbf{n}(l+1)\big)\Big)\right], \end{aligned} \tag{13}$$

where $(T(l), \mathbf{n}(l))$ is the tree at time l, and h is defined by

$$h\big((T, \mathbf{n}), (T, \mathbf{n} - \mathbf{e}_k)\big) = f_{\theta_0}(T, \mathbf{n})\frac{n + \theta_0 - 1}{n + \theta - 1}$$

and

$$h\big((T, \mathbf{n}), (T', \mathbf{n}')\big) = f_{\theta_0}(T', \mathbf{n}')\frac{\theta(n + \theta_0 - 1)}{\theta_0(n + \theta - 1)}$$

for the second type of transition in (10), when $(T', \mathbf{n}') = (\mathcal{S}_k T, \mathbf{n})$, and for the third type, when $(T', \mathbf{n}') = (\mathcal{R}_k T, \mathcal{R}_k(\mathbf{n} + \mathbf{e}_j))$.

The analogs of this scheme in the variable population size case can be found in Griffiths and Tavaré (1994c). As noted in the Introduction, the population size of the Nuu-Chah-Nulth is not thought to have fluctuated widely over the last 8,000 years, but the picture before that is less clear. With this in mind, we assume a constant population size in the analysis that follows. Further discussion of this point appears in Section 7.

5. ESTIMATING SUBSTITUTION RATES

The first issue we address is how to estimate and compare rates of substitution across the molecule. In particular, we ask whether the rates of substitution per site are the same in the purine and the pyrimidine regions.

We begin by describing how the infinitely-many-sites model approximates the evolution of sequences of finite length. Suppose then that the sequences have m_Y pyrimidine sites and m_R purine sites, in a total length of $m = m_Y + m_R$ sites. In our data, $m_R = 159$ and $m_Y = 193$. Let μ_R and μ_Y be the *per base* purine and pyrimidine substitution rates; μ_R and μ_Y measure the rate at which substitutions that change a site occur in the purine and pyrimidine regions, respectively. The overall substitution rate is

$$\theta = m_R\mu_R + m_Y\mu_Y,$$

and the marginal rates are

$$\theta_R = m_R\mu_R, \quad \theta_Y = m_Y\mu_Y, \tag{14}$$

respectively. The infinitely-many-sites process with multiple rates arises by supposing that m_R and m_Y tend to infinity in such a way that θ_R and θ_Y remain fixed. There are two equivalent ways in which the resulting process can be described. Suppose we have a realization of a coalescent tree with mutations superimposed on it. One description is to imagine mutations occurring according to the Poisson mechanism with rate $\theta/2$ in each branch, and then independently to label them a Y-site substitution with probability $p_Y = \theta_Y/\theta$ or an R-site substitution with probability $p_R = \theta_R/\theta$. The other is that mutations at the two types of site are laid down according to independent Poisson processes with rates $\theta_Y/2$ and $\theta_R/2$, respectively. In this model, because of the homogeneous Poisson process along the edges of the tree, it is not important whether mutations are laid down in a forward or backward direction. A more complex mutation process, where mutations are laid down in a Markov scheme from the MRCA type forward in time is considered in Griffiths and Tavaré (1994a).

In order to assess whether the per site substitution rates are equal, notice that, conditional on the total number of segregating sites (say, s), the number that occur in the R-positions is binomially distributed with parameters s and p_R. When $\mu_R = \mu_Y$, p_R reduces to $p_R = m_R/m = 159/352 = 0.45$, and the hypothesis of equal rates can be tested in the obvious way. For our data, the observed number of segregating sites is $s = 18$, and five of those occurred in the R-positions. The probability of seeing five or fewer successes is approximately 0.11, suggesting no differences in the rates at the two types of site. Indeed, the

same method can be used to find a very rough confidence interval for the ratio of the rates in the region. A 95% confidence interval for p_R is (0.07, 0.48), which translates into a 95% confidence interval for μ_Y/μ_R of (0.88, 10.81). We see that very wide fluctuations in rates are not inconsistent with these data.

We can use the Monte Carlo likelihood method described in Section 4 to estimate the overall substitution rate θ by maximum likelihood. For the unrooted tree in Figure 4, we ran $g = 200{,}000$ repetitions of the surface simulation algorithm to find the probability of each of the 19 rooted trees, summed these to find the probability of the unrooted tree and from this obtained an MLE of $\widehat{\theta} = 4.8$. The standard deviation (sd) can be estimated heuristically from the likelihood curve by first computing the second derivative at the maximum. This gave a value of $\mathrm{sd}(\widehat{\theta}) \approx 1.48$. As is typical in this field, the rate cannot be estimated very precisely (although in principle it can be estimated consistently as the sample size increases).

In estimating θ this way, we are not assuming anything about the type of each site. We could modify our algorithm to estimate simultaneously the overall rate θ and the relative rate in the R and Y classes. We do not pursue this here. Rather, we see what information can be found by analyzing the purine and pyrimidine sites separately, as though they were the whole data set. The unrooted subtrees can be found from the unrooted tree in Figure 4. Mutations labeled with ■ symbols correspond to the purine sites, those labeled with • symbols to pyrimidine sites. The frequencies of each subset of vertices can be calculated from the data in Table 1. Corresponding to the purine sites there are six rooted trees, and to the pyrimidine sites there are 14 rooted trees. Using the algorithm in (13) once more, we estimated the rates as $\widehat{\theta}_R = 1.22 \pm 0.61$ and $\widehat{\theta}_Y = 3.31 \pm 1.14$, the plus-or-minus figure being the estimated standard deviation.

Since each mutation produces a new segregating site, the number of segregating sites N_n^R among the purines and N_n^Y among pyrimidines in a sample of n sequences is precisely the number of mutations of each type, so that

$$\mathbb{E}N_n^Y = \theta_Y H_n, \quad \mathbb{E}N_n^R = \theta_R H_n, \tag{15}$$

where $H_n = 1 + 1/2 + \cdots + 1/(n-1)$. We can use (15) to construct moment estimators $\widetilde{\theta}$, $\widetilde{\theta}_R$ and $\widetilde{\theta}_Y$ of the rates θ, θ_R and θ_Y, respectively (cf. Watterson, 1975). These estimates satisfy $\widetilde{\theta} = \widetilde{\theta}_R + \widetilde{\theta}_Y$. For our data, we obtain $\widetilde{\theta} = 3.93, \widetilde{\theta}_R = 1.09$ and $\widetilde{\theta}_Y = 2.84$. The situation is very similar for the MLE, where we found $\widehat{\theta} = 4.76, \widehat{\theta}_R = 1.22$ and $\widehat{\theta}_Y = 3.31$.

Lundstrom, Tavaré and Ward (1992) used the full data set of 63 sequences and 360 sites to estimate rates, assuming a model with finitely many sites and uniform rates across each of the regions. They assumed that, conditional on the coalescent tree, substitutions occur independently at each site, with rate matrix R given by $R = (\nu/2s)(P - I)$. Here P is a matrix with identical rows (π_0, π_1), π_0 being the stationary frequency of the base labeled 0, and π_1 the stationary frequency of the other base. The overall rate at which substitutions that change the type at a site occur is then $\mu = 2\pi_0\pi_1\nu/s$. For large s this process is approximated by the infinitely-many-sites model with rate θ given by $\theta = s\mu$, just as in (14). This correspondence allows us to compare estimates across different models using different parts of the data. For the pyrimidine region, Lundstrom, Tavaré and Ward estimated $\mu_Y = 0.05 \times 0.48 = 0.024$, and this gives the estimate $\theta_Y \approx 193 \times 0.024 = 4.63$. This estimate is in good agreement with the rate $\widehat{\theta}_Y$ found here for the infinitely-many-sites model. For the purine region, Lundstrom, Tavaré and Ward estimated $\mu_R = 0.02 \times 0.42 \approx 0.008$, so that $\theta_R \approx 159 \times 0.008 = 1.27$, once more in good agreement with the estimate $\widehat{\theta}_R = 1.22$ found here. We conclude that our data selection method has had little influence on estimates of substitution rates in the region.

Estimating rates is an important problem because they are used to calibrate evolutionary clocks, and so to infer properties of the ancestry of the sample. Some examples are discussed in the next section.

6. ANCESTRAL INFERENCE

We discuss two examples of ancestral inference: identifying the ancestral lineage, and finding the conditional distribution of the time to the most recent common ancestor given the observed data $(\mathbf{Q}, \mathbf{n})$. The first question can be addressed by comparing the likelihoods (at the MLE $\widehat{\theta}$) of the different rooted trees that correspond to the unrooted tree in Figure 4. Each of these corresponds to a particular labeling of types in each site as ancestral or mutant. The p^0 probabilities of the 19 trees vary between 1.0×10^{-19} and 1.2×10^{-25}. The most likely is the one given in Table 2, which has the root at lineage k. This tree corresponds to the most frequent base at each site being labeled as ancestral. The next four most likely trees have likelihoods of 3.6×10^{-20}, 3.5×10^{-20}, 1.1×10^{-20} and 9.1×10^{-21}. These correspond to the root being placed between the mutations labeled 4 and 14, between 4 and 6, at lineage l and at lineage e in Figure 4, respectively. The relative likelihood that the root is at one of these locations, given the data, is about 97%. Looking at the topology of Figure 4, we find quite a tight concentration for the possible roots.

The second issue concerns inferences about the distribution of the time to the MRCA of the sample. In the absence of any data at all, T_{MRCA} has the distri-

bution given in (1). For a sample of size $n = 55$ this distribution has mean and standard deviation given by

$$\text{(16)} \quad \mathbb{E}(T_{\text{MRCA}}) = 1.96, \quad \text{sd}(T_{\text{MRCA}}) = 1.08.$$

How is this distribution changed when we condition on the data? To assess this, we calculate the conditional distribution $\mathbb{P}(T_{\text{MRCA}} \le w \mid (\mathbf{Q}, \mathbf{n}))$. We can use the Monte Carlo approach to address this, because

$$\text{(17)} \quad \mathbb{P}\big(T_{\text{MRCA}} \le w \mid (\mathbf{Q}, \mathbf{n})\big) = \frac{\mathbb{P}\big(T_{\text{MRCA}} \le w, (\mathbf{Q}, \mathbf{n})\big)}{\mathbb{P}\big((\mathbf{Q}, \mathbf{n})\big)},$$

a ratio of probabilities each of which satisfies a recursion. The denominator is $p^0(\mathbf{Q}, \mathbf{n})$, which we know how to compute already. It remains to calculate the numerator. First suppose that we want to find the analogous probability for a rooted tree $(T, \mathbf{n})$. To do this we modify the Markov chain scheme in Section 4 by keeping a more detailed description of the sample path. At each move of the chain $X(l) = (T(l), \mathbf{n}(l))$, record the time taken for that move. If the chain is currently in a state which corresponds to a sample of size m, then the time to the next move is exponential with parameter $m(m + \theta - 1)/2$; these waiting times can readily be simulated. For the jth of g simulated trajectories, we have to do the following:

1. watch the chain back until the state $(T_1, \mathbf{e}_1)$ is reached, and compute the functional

$$F_j(T) = \prod_{l=0}^{k-1} f\big(T(l), \mathbf{n}(l)\big),$$

   where k is the number of steps until $(T_1, \mathbf{e}_1)$ is hit;
2. record the total time $\tau_j(T)$ that the run lasts.

An estimator of $\mathbb{P}(T_{\text{MRCA}} \le w, (T, \mathbf{n}))$ is then

$$\frac{1}{g} \sum_{j=1}^{g} F_j(T) \mathbf{1}\big(\tau_j(T) \le w\big),$$

and $\mathbb{P}(T_{\text{MRCA}} \le w, (\mathbf{Q}, \mathbf{n}))$ can be estimated by summing these estimators over all the rooted trees T. An estimator of the conditional distribution function in (17) is therefore

$$\text{(18)} \quad \frac{\sum_T \sum_{j=1}^{g} F_j(T) \mathbf{1}\big(\tau_j(T) \le w\big)}{\sum_T \sum_{j=1}^{g} F_j(T)}.$$

If $\tau_{(j)}$ denotes the jth smallest of all the $\tau_j(T)$, and $F_{(j)}$ is the corresponding $F_j(T)$, then this empirical distribution has jumps of height $F_{(j)}/\sum_l F_{(l)}$ at time $\tau_{(j)}$. In practice, these distribution functions are approximated by binning the observations in the usual way.

TABLE 3
Summary statistics of time to MRCA

Data	θ	$\mathbb{E}(T_{\text{MRCA}} \mid \text{data})$	$\text{sd}(T_{\text{MRCA}} \mid \text{data})$
None (16)		1.96	1.08
R, *Y*-sites	3.3	1.40	0.55
R, *Y*-sites	4.8 (= $\hat{\theta}$)	1.20	0.39
R, *Y*-sites	6.3	0.96	0.12
Y-sites	3.3 (= $\hat{\theta}$)	1.26	0.41
R-sites	1.2 (= $\hat{\theta}$)	1.54	0.65

As summary statistics of these conditional distributions, we record in Table 3 the conditional mean $\mathbb{E}(T_{\text{MRCA}} \mid (\mathbf{Q}, \mathbf{n}))$ and the conditional standard deviation $\text{sd}(T_{\text{MRCA}} \mid (\mathbf{Q}, \mathbf{n}))$ for comparison with the unconditional figures given in (16). We see that the mean time to the MRCA is substantially reduced when information in the data is taken into account. For the full data set comprising both purine and pyrimidine sites this amounts to a reduction of about 40%.

As noted earlier, this population appears to have been relatively constant in size for the last 6,000 years, prompting us to assume a constant-population-size model in our analysis. To translate our estimates of ancestral times into real time units therefore requires an estimate of the current population size N of childbearing age. Ward et al. (1991) estimate this size at between 400 and 800 individuals; we use a value of $N = 600$ in what follows. If we take a generation to be 20 years, then the unconditional mean time corresponds to $1.96 \times 600 \times 20 \approx 23,500$ years. Conditional on the sample configuration, this mean is reduced to 14,400 years, a figure in remarkably good agreement with the estimate of 13,000 years given by Ward et al. (1993) using an entirely different approach.

Notice also that the standard deviation is substantially reduced, reflecting the fact that with more information the conditional distribution of the time to the MRCA should be much more concentrated about its mean (although this distribution will have nonzero variance even for an infinitely large sample).

The reduction in conditional mean time to the MRCA given the data, compared to the unconditional mean time, is not a general feature for all data sets. Intuitively, for a fixed value of θ, data showing more (less) segregating sites than the expected number should have increased (decreased) expected conditional time given the data. This may not be precisely true, because the expectations depend on the detailed structure of the data.

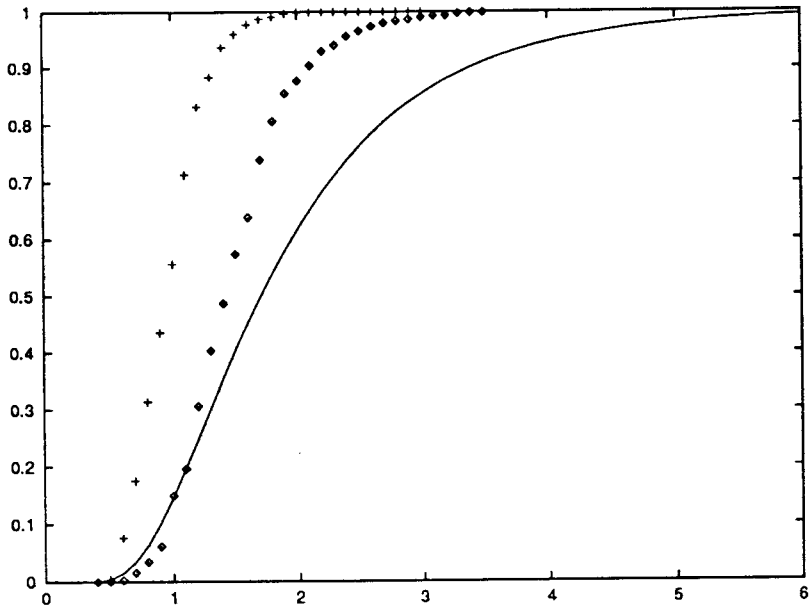

FIG. 5. *Empirical distribution of time to MRCA*: (*solid line*) *unconditional distribution*; (◇) *conditional distribution* ($\theta = 3.3$); (+) *conditional distribution* ($\theta = 6.3$).

To assess the variability in these estimates that arises from uncertainty in our estimates of θ, we calculated the corresponding quantities for two other plausible values of θ, namely, 3.3 and 6.3 (these being $\widehat{\theta} \pm 1$ sd). These results are also given in Table 3. Qualitatively, we see the same sort of behavior: the conditional mean is much reduced, as is the standard deviation. As should be expected, the mean time to the MRCA will be smaller the larger the assumed value of the clock rate, θ.

Since there is no recombination in these molecules, different regions of the molecule have shared an identical evolutionary history. In particular, each of the purine and pyrimidine regions provides a separate estimate of the distribution of the time to the MRCA. It is therefore of interest to know whether these regions give us similar ancestral signals. In Table 3, we give summary statistics for the analysis based on just the purine sites (R), and the pyrimidine sites (Y). We see that, while the conditional means get closer to their unconditional values, the two regions give estimates for the time to the MRCA comparable with that for the region as a whole.

We turn now to the distribution of the time to the MRCA. In Figure 5 we plot the empirical conditional distribution function (df) of T_{MRCA} given the data, obtained from (18), for values of θ bracketing the MLE. This is compared to the distribution of the unconditional time determined by (1). The estimated conditional df for $\theta = 4.8$, the MLE, lies between the two conditional df's shown in Figure 5. The marked reduction in variance is clearly shown in the very steep slope of the df's relative to the unconditional df. As might be expected, the conditional times are stochastically ordered in θ.

7. DISCUSSION

The methods described in this paper should be taken as illustrative of techniques currently being developed to study DNA sequence variation within populations. Our approach derives directly from the recognition of the central importance of ancestry encapsulated in Kingman's coalescent process. Accessible reviews about the coalescent appear in Hudson (1991) and Tavaré (1994). This way of thinking about population genetic data has revolutionized statistical approaches to the study of such variation. We have exploited this approach to derive sampling distributions for data, from which estimates of genetic parameters such as substitution rates might be found and from which hypotheses about the populations under study might be tested.

Our analysis of the mitochondrial sequence data from the Nuu-Chah-Nulth should be regarded as exploratory. There are several features that need further investigation, among them a deeper study of the issues of population expansion, admixture and migration. Assumptions of demographic stability are tenuous and, with the exception of humans, essentially untestable. Because of extensive archaeological data from many parts of the world, we do have some information on the demographic history of many human populations, and we also have some

record of whether extensive migration has occurred in the past. The archaeological data for the Nuu-Chah-Nulth happens to be extensive, and it allows us to identify something about the population at the time the area was settled. It is thought that the Nuu-Chah-Nulth population arose from a split of a much larger population at about that time. With Dr. Ryk Ward, we are currently studying the sort of effect such a bottleneck would have on estimates of the distribution of the time to the MRCA and thus on inferences about the colonization of the Americas. Monte Carlo methods for inference in the case of variable population size are described in Griffiths and Tavaré (1994c). In addition, we based our analysis on a simplified picture of the evolution of mitochondrial DNA molecules. Our data selection involved choosing a subset of sites and individuals compatible with the simple infinitely-many-sites model, and we saw that estimates of substitution rates obtained from this approach agree closely with analyses based on more detailed models. Our choice of the infinitely-many-sites model is based in part on computational expediency, and the belief that recurrent mutation has not occurred at many sites. Analogous methods have been devised for models in which recurrent mutation can occur (see Griffiths and Tavaré, 1994a) but these are much more computer-intensive and cannot (at the time of writing) handle sequences of the size described here.

Our statistical approach is based on Monte Carlo methods for solving certain sorts of recursion or integro-recurrence equations that derive from the coalescent. Thus far, population geneticists have emphasized the study of simpler summary statistics of the data, such as the number of segregating sites or the distribution of pairwise differences (obtained by comparing pairs of sequences and recording the number of pairs with $0, 1, \ldots$ differences), and have used these summaries for estimation and inference. While this represents an important first step in a new field, it fails to make full use of the information in the data. The trade-off is that the computational complexity of the analysis can increase substantially. Given the time taken to collect extensive collections of sequence data, it seems reasonable to explore statistical techniques that make fuller use of these hard-won data.

Computer-intensive likelihood methods have been used to calculate probabilities on complex pedigrees arising in human genetics. Cannings, Thompson and Skolnick (1978) describe the structure of the models, and Thompson and Guo (1991) and Thompson (1994) show how Markov chain Monte Carlo (MCMC) methods can be applied. The techniques we use are rather different, in that we generate independent runs of a process of random length as opposed to the dependent observations on a single ergodic process typical of MCMC. The techniques have in common the fact that they are in principle quite old, MCMC dating back to Metropolis et al. (1953) and Hastings (1970), and the present method at least to Forsythe and Leibler (1950) in the context of solving linear equations. See Halton (1970) for further examples.

As sequence data become more prevalent, population geneticists will be able to refine their analyses by studying the joint evolution of several different genomic regions at once. While this should give a more comprehensive picture of the evolutionary history of our species, the analysis will be much more complicated. This derives in part from the fact that recombination between different regions of nuclear genes is very common, and recombination has the effect of scrambling evolutionary history. A second difficulty concerns another statistical issue, that of sampling strategy. In our look at the Nuu-Chah-Nulth, we supposed that the sample was indeed "random." In practice this is of course very difficult to arrange. The sensitivity of estimates and inferences to nonrandom sampling certainly needs to be quantified. Finally, much of the population genetics theory currently used to analyze molecular variability is based on rather simplified, selectively neutral, models of reproduction. These assumptions, embodied in this paper in the coalescent, warrant further study with a view to assessing how well conclusions based on them might apply to human populations. The development of statistical approaches to such issues will be an important part of population genetics for many years to come.

ACKNOWLEDGMENTS

Professor Tavaré is supported in part by NSF Grant DMS-90-05833, and both authors were supported in part by the IMA during the preparation of this paper. We thank Ryk Ward for many helpful discussions about the mitochondrial data discussed here, and Sue Wilson for comments on an earlier draft.

REFERENCES

ANDERSON, S., BANKIER, A., BARRELL, B., DE BRUIJN, M., COULSON, A., DROUIN, J., EPERON, I., NIERLICH, D., ROE, B., SANGER, F., SCHREIER, P., SMITH, A., STADEN, R. and YOUNG, I. (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290** 457–465.

CANN, R. M., STONEKING, M. and WILSON, A. (1987). Mitochondrial DNA and human evolution. *Nature* **325** 31–36.

CANNINGS, C., THOMPSON, E. A. and SKOLNICK, M. H. (1978). Probability functions on complex pedigrees. *Adv. in Appl. Probab.* **10** 26–61.

DEWHIRST, J. (1978). Nootka Sound: a 4000 year perspective. *Sound Heritage* **7** 1–29.

FELSENSTEIN, J. (1982). Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology* **57** 379–404.

FLADMARK, K. R. (1975). *A Paleoecological Model for Northwest Coast Prehistory*. National Museums of Canada, Ottawa.

FORSYTHE, G. E. and LEIBLER, R. A. (1950). Matrix inversion by the Monte Carlo method. *Math. Comp.* **26** 127–129.

GRIFFITHS, R. C. (1987). An algorithm for constructing genealogical trees. Statistics Research Report 163, Dept. Mathematics, Monash Univ.

GRIFFITHS, R. C. and TAVARÉ, S. (1994a). Simulating probability distributions in the coalescent. *Theoret. Population Biol.* **46** 131–159.

GRIFFITHS, R. C. and TAVARÉ, S. (1994b). Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosciences.* To appear.

GRIFFITHS, R. C. and TAVARÉ, S. (1994c). Sampling theory for neutral alleles in a varying environment. *Philos. Trans. Roy. Soc. London Ser. B* **344** 403–410.

GUSFIELD, D. (1991). Efficient algorithms for inferring evolutionary trees. *Networks* **21** 19–28.

HALTON, J. H. (1970). A retrospective and prospective study of the Monte Carlo method. *SIAM Rev.* **12** 1–63.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

HUDSON, R. R. (1991). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* **7** (D. Futuyma and J. Antonovics, eds.) 1–44. Oxford Univ. Press.

KINGMAN, J. F. C. (1982a). On the genealogy of large populations. In *Essays in Statistical Science* (J. Gani and E. J. Hannan, eds.) 27–43. Applied Probability Trust, Sheffield, UK.

KINGMAN, J. F. C. (1982b). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics* (G. Koch and F. Spizzichino, eds.) 97–112. North-Holland, Amsterdam.

LUNDSTROM, R., TAVARÉ, S. and WARD, R. H. (1992). Estimating mutation rates from molecular data using the coalescent. *Proc. Nat. Acad. Sci. U.S.A.* **89** 5961–5965.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

SCHURR, T., BALLINGER, S., GAN, Y., HODGE, J., MERRIWETHER, D. A., LAWRENCE, D., KNOWLER, W., WEISS, K. and WALLACE, D. (1990). Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. *American Journal of Human Genetics* **47** 613–623.

SHIELDS, G. F, SCHMIECHEN, A. M., FRAZIER, B. L., REDD, A., VOEVODA, M. I., REED, J. K. and WARD, R. H. (1993). mtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. *American Journal of Human Genetics* **53** 549–562.

STONEKING, M. (1993). DNA and recent human evolution. *Evolutionary Anthropology* **2** 60–73.

TAVARÉ, S. (1994). Calibrating the clock: using stochastic processes to measure the rate of evolution. In *Calculating the Secrets of Life* (E. S. Lander, ed.). National Academy Press, Washington, DC. To appear.

THOMPSON, E. A. (1994). Monte Carlo likelihood in genetic mapping. *Statist. Sci.* **9** 355–366.

THOMPSON, E. A. and GUO, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. Appl. Med. Biol.* **8** 149–169.

TORRONI, A., SCHURR, T. G., YANG, C., SZATHMARY, E. J. E., WILLIAMS, R. C., SCHANFIELD, M. S., TROUP, G. A., KNOWLER, W. C., LAWRENCE, D. N., WEISS, K. M. and WALLACE, D. C. (1992). Native American mitochondrial DNA analysis indicates the Amerind and Na-Dene populations were founded by two independent migrations. *Genetics* **130** 153–162.

WARD, R. H., FRAZIER, B. L., DEW, K. and PÄÄBO, S. (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Nat. Acad. Sci. U.S.A.* **88** 8720–8724.

WARD, R. H., REDD, A., VALENCIA, D., FRAZIER, B. and PÄÄBO, S. (1993). Genetic and linguistic differentiation in the Americas. *Proc. Nat. Acad. Sci. U.S.A.* **90** 10,663–10,667.

WATTERSON, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biol.* **7** 256–276.