Theoretical Population Biology **53**, 143–151 (1998)
Article No. TP971348

TPB

Maximum Likelihood Estimation of Population Divergence Times and Population Phylogenies under the Infinite Sites Model

Rasmus Nielsen*

Department of Integrative Biology, University of California, Berkeley, California 94720-3140

E-mail: rasmus@mws4.biol.berkeley.edu

Received January 24, 1997

In this paper, a maximum likelihood estimator of population divergence time based on the infinite sites model is developed. It is demonstrated how this estimator may be applied to obtain maximum likelihood estimates of the topology of population phylogenies. This approach addresses several classical problems occurring in the inference of the phylogenetic relationship of populations, most notably the problem of shared ancestral polymorphisms. The method is applied to previously published data sets of human African populations and of Caribbean hawksbill turtles. © 1998 Academic Press

INTRODUCTION

In recent years, the genetic analysis of population subdivision has undergone dramatic development. Phylogenetic approaches have been championed by numerous authors (Vigilant *et al.*, 1989; Slatkin and Maddison, 1990; Mountain and Cavalli-Sforza, 1994; Patton *et al.*, 1994; Templeton *et al.*, 1995). Perhaps the most notable application of phylogenetic approaches is the analysis of the divergence of human ethnic groups out of Africa (Vigilant *et al.*, 1989; Watson *et al.*, 1996). In this type of analysis a gene genealogy is estimated and conclusions regarding migration and divergence between populations are inferred from the estimated gene genealogy. One of the most serious challenges in this type of analysis is the potential lack of concordance between population phylogenies and gene genealogies even in the absence of migration between the populations. The branching pattern in a gene genealogy consisting of genes sampled from several populations may be caused both by divergence between the genes after the time of population separation and by divergence before the time of population separation. This problem of shared ancestral polymorphisms has, in certain parts of the literature, been termed lineage sorting (see, for example, Avise, 1994). The problem of shared ancestral polymorphisms may be relevant not only within species but also between species. For example, the difficulty in determining the branching pattern of humans, chimpanzees, and gorillas (see for example Horai *et al.*, 1992) may very well reflect that a large proportion of the divergence between individuals occurred in the time before speciation. Several authors have attempted to account for the effect of shared ancestral polymorphisms. Simple probabilistic arguments based on coalescent theory have been used to assess the probability of disconcordance between population phylogeny and gene-genealogy (Takahata, 1989; Wu, 1991; Hudson, 1992). However, a more appropriate approach would be to estimate the parameters of the underlying demographic process directly. For example, if one is interested in the divergence time between populations, this time should be estimated directly instead of

[*] Current address: Museum of Comparative Zoology, Harvard University, 26 Oxford St., Cambridge, Massachusetts 02138.

0040-5809/98 $25.00

estimating the coalescence time of the gene genealogy and then subsequently relating these estimates to the divergence times of the populations. Obtaining a direct estimate of population divergence times requires integration over all possible coalescence times and topologies of the gene genealogies. One of the reasons why such an analysis has not been performed is that no analytical tools have been previously available for performing such integration. However, recent computational advances in population genetics have made this type of analysis tractable (Kuhner *et al.* 1995, Griffiths and Tavaré, 1994a, b). In this paper, a maximum likelihood method for directly estimating the divergence time of populations under the infinite sites model is presented. It is shown how this approach leads to a method for estimating population phylogenies. In the applications section, the population divergence times of two human populations from Africa are estimated using a previously published data set of human mitochondrial DNA. In addition, the population phylogeny of three populations of hawksbill turtles is estimated using previously published data.

THE METHOD OF GRIFFITHS AND TAVARÉ IN THE ONE-POPULATION CASE

Before embarking on the analysis of multiple populations it would be useful to first review some of the results and terminology applied in the method of estimating of $\theta = 4N\mu$ (where μ is the mutation rate in the entire haplotype and N is the population size) described by Griffiths and Tavaré (1994a, 1995). Subsequently, it will be shown how this methodology can be applied in the estimation of parameters of the demographic process for multiple populations.

In the following, we will assume that the divergence within a population follows Kingman's coalescent process (Kingman 1982, see Tavaré 1984 or Hudson 1991 for a review). This implies, among other assumptions, that we assume random mating, selective neutrality between genes and a constant population size. However, as demonstrated by Griffiths and Tavaré (1994), changes in population size can easily be incorporated into the model. We will further assume that the mutational process follows an infinite sites model (Kimura, 1969). This implies that multiple mutations in the same nucleotide site are not allowed. The sample of haplotypes, under the infinite sites model, can be represented by a matrix of binary characters since only one mutation can occur in each site. For example, had we obtained a sample from a single population consisting of 5 haplotypes containing 4 variable sites:

atgc
accc
acgc
accc
gcct

this data set could be coded in the binary matrix (**S**) with arbitrary labeling of zeros and ones as

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{n} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 1 \end{bmatrix},$$

where **n** is a vector containing the counts or multiplicities of each haplotype (type of haplotype). The quantity we are interested in obtaining is the likelihood of θ, that is, the probability of observing our particular sample of ordered haplotypes given θ. Denote this probability $p(\mathbf{S}, \mathbf{n})$. To obtain an expression for this probability we will consider the genealogical history of the haplotypes. We will sum over all other possible ancestral samples at a previous time which by one mutation or one coalescence event could be transformed into the present sample. In order to do this some notation must be introduced.

A sample in which a coalescence event just occurred between haplotypes of the kth type is given by $(\mathbf{S}, \mathbf{n} - \mathbf{e}_k)$ where $\mathbf{e}_k$ is a unit vector that subtracts 1 from entry k of **n**. If the last event was a mutation, there are two possibilities. If the mutation originally happened in a haplotype with only one copy in the sample, then the sample before the mutation is given by $(\mathbf{S}^l, \mathbf{n})$, where $\mathbf{S}^l$ denotes a matrix identical to **S** but with the lth column removed corresponding to the elimination of one segregating site. If the mutation occurred from a haplotype (j) with multiple copies in the sample, then the sample before mutation is given by $(\mathbf{S}^{kl}, \mathbf{n}^k + \mathbf{e}_j)$, where $\mathbf{S}^{kl}$ denotes a matrix identical to **S** but with column l and row k removed and $\mathbf{n}^k$ denotes a vector identical to **n** but with entry k removed. This corresponds to eliminating the one segregating site that distinguished haplotype k from haplotype j.

Now, note that conditional on either a mutation or coalescence event occurring, the probability that it was a coalescence event is

$$\frac{n-1}{n-1+\theta}$$

and the probability that it was mutation is

$$\frac{\theta}{n-1+\theta}$$

where n is the sample size.

Under the infinite sites model, mutations to previously existing haplotypes (back mutations) are not allowed. Thus, mutations can have occurred as the most recent event only in haplotypes of multiplicity 1. The probability of a mutation in haplotype k ($n_k = 1$) is $1/n$, where n_k is the multiplicity of haplotype k. Likewise, the probability of a coalescence event in haplotype k ($n_k > 1$) given a coalescence occurred is $n_k(n_k - 1)/n(n-1)$.

Now, by summing over all possible previous states in the genealogy we obtain the following recursion

$$p(\mathbf{S}, \mathbf{n}) = \frac{1}{n(n-1+\theta)}\left[\sum_{k \in Z_1} n_k(n_k - 1)\, p(\mathbf{S}, \mathbf{n} - \mathbf{e}_k) + \theta \sum_{k \in Z_2} p(\mathbf{S}^l, \mathbf{n}) + \theta \sum_{k \in Z_3} p(\mathbf{S}^{kl}, \mathbf{n}^k + \mathbf{e}_j)\right] \quad (1)$$

(Griffiths and Tavaré, 1995). The first sum (the coalescence case) is over all haplotypes with multiplicities larger than one, i.e.,

$$Z_1 = \{k : n_k \geqslant 2\}.$$

The second sum (mutation from a single copy haplotype) is over all haplotypes with multiplicities 1 that differ from all other haplotypes by at least two mutations and that contain a site with a unique mutation, i.e.,

$$Z_2 = \{k : n_k = 1, \mathbf{s}_{.l} = \mathbf{e}_k \; or \; \mathbf{s}_{.l} = \mathbf{e}_k^c, \mathbf{s}_{k.}^l \neq \mathbf{s}_j^l, k \neq j\}.$$

where $\mathbf{e}_k^c$ is the complement of $\mathbf{e}_k$, $\mathbf{s}_i$ is the ith row of $\mathbf{S}$ and $\mathbf{s}_{.l}$ is the lth column of $\mathbf{S}$. The third sum (mutation from a multiple copy haplotype) is over haplotypes with multiplicities 1 that contain a site that differs from all other haplotypes and which is indistinguishable from another haplotype except for this site, i.e.,

$$Z_3 = \{k : n_k = 1, \mathbf{s}_{.l} = \mathbf{e}_k \; or \; \mathbf{s}_{.l} = \mathbf{e}_k^c, \mathbf{s}_{k.}^l = \mathbf{s}_j^l, k \neq j\}.$$

As an example, assume that the following sample was observed:

$$\mathbf{S} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{n} = \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

then $Z_1 = \{1\}$, $Z_2 = \{2\}$ and $Z_3 = \{\ \}$. Therefore

$$p(\mathbf{S}, \mathbf{n}) = \frac{1}{3(2+\theta)}\left(2p\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) + \theta p\left(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}\right)\right).$$

For a more rigorous derivation of Eq. (1) consult Ethier and Griffiths (1987) which provide a treatment of sample probabilities under the infinite sites model using measure theory. Hudson and Kaplan (1986) derived a similar recursive equation for the infinite alleles model. Eq. (1) first appeared in the exact form presented here in Griffiths and Tavaré (1995).

Note that in theory, $p(\mathbf{S}, \mathbf{n})$ can be calculated directly from this recursion by iteration and by specifying the boundary condition

$$p(2, m) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^m \quad (2)$$

(Watterson, 1975) where $p(2, m)$ is the probability of obtaining a particular sample of two haplotypes with m segregating sites. However, for samples of sizes larger than 15–20, the number of possible states in the recursion is so large that a direct evaluation is not computationally possible. Instead, the Markov chain Monte Carlo approach by Griffiths and Tavaré (1994a, 1994b, 1995) can be applied to provide estimates of $p(\mathbf{S}, \mathbf{n})$. Griffiths and Tavaré (1995) developed a method for evaluating recursion similar to Eq. (1) and applied the method to evaluate the probabilities of unrooted trees. The derivation of a Markov chain Monte Carlo approach for Eq. (1) follows trivially from the derivation for the tree probabilities given by Griffiths and Tavaré (1995) since it only involves a slight change in state space. Sequences under the infinite sites model can simply be interpreted as unrooted genealogical trees, with possible multifurcations in cases where the data do not allow distinction between different alleles (lineages).

By defining a Markov chain with the same state space as the recursion (Eq. (1)), one may evaluate $p(\mathbf{S}, \mathbf{n})$ by evaluating only a subset of all possible paths in the recursion. Paths are chosen by simulating along the Markov chain. In particular, by defining a function

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

where v is given by

$$v = |\{k : n_k = 1, s_{.l} = e_k \; or \; s_{.l} = e_k^c \text{ for some } l\}|,$$

and $|\{...\}|$ indicates the size of a set, and a Markov chain with the following transition probabilities

$$(\mathbf{S}, \mathbf{n}) \to (\mathbf{S}, \mathbf{n} - \mathbf{e}_k) \text{ with probability } \frac{n_k(n_k - 1)}{f(\mathbf{S}, \mathbf{n})\, n(n + \theta + 1)}$$

$$(\mathbf{S}, \mathbf{n}) \to (\mathbf{S}^l, \mathbf{n}) \text{ with probability } \frac{\theta}{f(\mathbf{S}, \mathbf{n})\, n(n + \theta + 1)}$$

$$(\mathbf{S}, \mathbf{n}) \to (\mathbf{S}^{kl}, \mathbf{n} + \mathbf{e}_j) \text{ with probability } \frac{\theta}{f(\mathbf{S}, \mathbf{n})\, n(n + \theta + 1)} \tag{3}$$

where the conditions under which transitions of the three types are allowed, is provided by $k \in Z_1$, $k \in Z_2$, and $k \in Z_3$, respectively. Repeated simulation of the Markov chain until hitting the absorptive state $(n = 2)$ provides an estimate of $p(\mathbf{S}, \mathbf{n})$

$$\hat{p}(\mathbf{S}, \mathbf{n}) = \frac{\sum_{j=1}^{v} p(2, m(\eta)) \prod_{i=0}^{\eta - 1} f(\mathbf{S}(i), \mathbf{n}(i))}{v}, \tag{4}$$

where v is the number of simulations performed, η is the random number of states the chain visits until the absorbing state $(n = 2)$ is hit, $p(2, m(\eta))$ is the number of segregating sites between the two remaining haplotypes at time η and $f(\mathbf{S}(i), \mathbf{n}(i)$ is the value of $f(\mathbf{S}, \mathbf{n})$ at the ith state the Markov chain visits. A proof of this result will not be provided here. Readers interested in the derivation should consult Griffiths and Tavaré (1994a, b).

TWO POPULATIONS

The aim of this section is to derive a method for evaluating the likelihood function for two populations that diverged some time (T) in the past. T refers here to the scaled time, that is, the divergence time measured in generations divided by the effective populations size. In the two population case we also denote the matrix of haplotypes by $\mathbf{S}$ but there are now two vectors $\mathbf{n}_1$ and $\mathbf{n}_2$, containing the multiplicities of the haplotypes in population 1 and in population 2, respectively. Also let $\mathbf{S}(T-)$, $\mathbf{n}_1(T-)$, and $\mathbf{n}_2(T-)$ denote the particular values of $\mathbf{S}$, $\mathbf{n}_1$ and $\mathbf{n}_2$, respectively, in the ancestry at time right before T and $\mathbf{S}(T+)$, $\mathbf{n}_1(T+)$ and $\mathbf{n}_2(T+)$ right after T, looking forward in time (Fig. 1). Then, the probability of obtaining a particular sample of haplotypes from two populations is given by

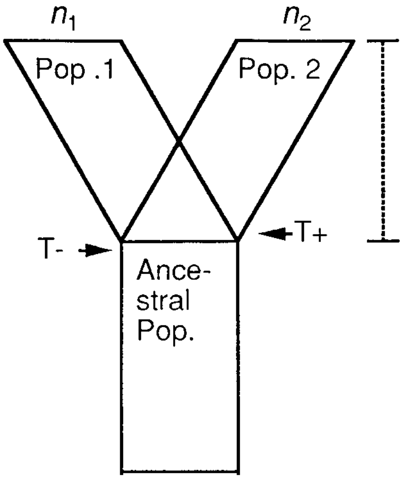

FIG. 1. The divergence of two populations from a common ancestral population.

$$\begin{aligned} p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid T) &= \sum_{\mathbf{S}(T+),\, \mathbf{n}_1(T+),\, \mathbf{n}_2(T+)} p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid \mathbf{S}(T+), \mathbf{n}_1(T+), \\ &\quad \mathbf{n}_2(T+))\, p(\mathbf{S}(T+), \mathbf{n}_1(T+), \mathbf{n}_2(T+)) \\ &= \sum_{\mathbf{S}(T+),\, \mathbf{n}_1(T+),\, \mathbf{n}_2(T+)} p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid \mathbf{S}(T+), \mathbf{n}_1(T+), \\ &\quad \mathbf{n}_2(T+))\; p(\mathbf{S}(T-), \mathbf{n}_1(T-) + \mathbf{n}_2(T-)) \\ &\quad \cdot p(\mathbf{S}(T+), \mathbf{n}_1(T+), \\ &\quad \mathbf{n}_2(T+) \mid \mathbf{S}(T-), \mathbf{n}_1(T-) + \mathbf{n}_2(T-)) \end{aligned}$$

where p_2 denotes the probability of observing a two population sample. Notice that the sum is over all possible values of $(\mathbf{S}(T+), \mathbf{n}_1(T+), \mathbf{n}_2(T+))$ that could lead to a sample of $(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2)$. If both the haplotypes and the two populations are ordered, no combinatorial factor is involved when considering the events at the time of population divergence. Therefore

$$\begin{aligned} p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid T) &= \sum_{\mathbf{S}(T+),\, \mathbf{n}_1(T+),\, \mathbf{n}_2(T+)} p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid \mathbf{S}(T+), \\ &\quad \mathbf{n}_1(T+), \mathbf{n}_2(T+))\; p(\mathbf{S}(T-), \mathbf{n}_1(T-) + \mathbf{n}_2(T-)) \end{aligned} \tag{5}$$

This probability provides the likelihood function of the population divergence time (T).

It was previously shown how to estimate $p(\mathbf{S}, \mathbf{n})$. It remains to be demonstrated how to calculate $p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid \mathbf{S}(T+), \mathbf{n}_1(T+), \mathbf{n}_2(T+))$ and perform the summation over all possible values of $(\mathbf{S}(T+)$, $\mathbf{n}_1(T+)$,

$\mathbf{n}_2(T+)$). In order to develop such a method we need to provide a recursive equation for the probability of obtaining two particular samples from two different populations given that the two populations diverged from a common population some time T ago and given that the last mutation or coalescence in the genealogy of the genes happened at a time τ in the past, where $\tau < T$. Notice that since both the time to coalescence and the time to a mutation are exponentially distributed, the relative probability of observing a mutation or coalescence is independent of T. In fact, the relative rate argument still normally applied in coalescence models still holds when conditioning on the time of the next event. The reason for this is that $P(x < y \mid \min(x, y) = T) = (a/(a+b))$, where x and y are exponential random variables with rates a and b. Now for $t < T$, coalescences occur with rates $(n_1(n_1-1)/2$ and $n_2(n_2-1)/2$ and mutations occur with rate $(n_1+n_2)\,\theta/2$, assuming the population sizes are constant. Let $f(\tau)$ be the density function of an exponential random variable with parameter $[n_1(n_1-1)+n_2(n_2-1)+(n_1+n_2)\,\theta]/2$ and let $F(\tau)$ be the corresponding CDF. For $\tau < T$, τ is the time to the next coalescence event or mutation looking back in time. If $\tau \geqslant T$, no coalescence events or mutations happened before T. This event will occur with probability $1-F(T)$. In that case,

$$p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid \tau, T)_{\tau \geqslant T} = p(\mathbf{S}, \mathbf{n}_1 + \mathbf{n}_2). \tag{6}$$

Conditional on $\tau < T$, we arrive at the following equation

$$p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid \tau, T)_{\tau < T}$$
$$= \frac{\left(\begin{aligned} &\sum_{k \in Z_{11}} n_{1k}(n_{1k}-1)\; p_2(\mathbf{S}, \mathbf{n}_1 - \mathbf{e}_k, \mathbf{n}_1 \mid T-\tau) \\ &\quad + \theta \sum_{k \in Z_{21}} p_2(\mathbf{S}^l, \mathbf{n}_1, \mathbf{n}_2 \mid T-\tau) \\ &\quad + \theta \sum_{k \in Z_{31}} p_2(\mathbf{S}^{kl}, \mathbf{n}_1^k + \mathbf{e}_j, \mathbf{n}_2 \mid T-\tau) \\ &\quad + n_{2k}(n_{2k}-1) \sum_{k \in Z_{12}} p_2(\mathbf{S}, \mathbf{n}_1 - \mathbf{e}_k, \mathbf{n}_2 \mid T-\tau) \\ &\quad + \theta \sum_{k \in Z_{22}} p_2(\mathbf{S}^l, \mathbf{n}_1, \mathbf{n}_2 \mid T-\tau) \\ &\quad + \theta \sum_{k \in Z_{32}} p_2(\mathbf{S}^{kl}, \mathbf{n}_1, \mathbf{n}_2^k + \mathbf{e}_j \mid T-\tau) \end{aligned} \right)}{n_1(n_1-1+\theta) + n_2(n_2-1+\theta)} \tag{7}$$

where n_{ik} is the multiplicity of haplotype k in population i and n_i is the size of the sample from population i. This equation follows from precisely the same arguments provided when establishing the one population recursion. However, now four events are possible; a coalescence in population 1, a coalescence in population 2, a mutation in population 1, or a mutation in population 1. Again, haplotypes can only be newly mutated if they have multiplicity 1 in the entire sample (population 1 and population 2). Therefore the set Z_{1i} is given by

$$Z_{1i} = \{k : n_{ki} \geqslant 2\}.$$

Likewise, Z_{2i} and Z_{3i} is given by

$$Z_{2i} = \{k : n_{ki} = 1,\, n_{k(3-i)} = 0,\, \mathbf{s}_{.l} = \mathbf{e}_k$$
$$or\ \mathbf{s}_{.l} = \mathbf{e}_k^c,\, \mathbf{s}_{k.}^l \neq \mathbf{s}_j^l,\, k \neq j\}$$

and

$$Z_{3i} = \{k_1 : n_{ki} = 1,\, n_{k(3-i)} = 0,\, \mathbf{s}_{.l} = \mathbf{e}_k$$
$$or\ \mathbf{s}_{.l} = \mathbf{e}_k^c,\, \mathbf{s}_{k.}^l = \mathbf{s}_j^l,\, k \neq j\}.$$

Notice that we can use Eq. (6) and Eq. (7) to write

$$p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid T) = \int_0^T p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid \tau, T)_{\tau < T}\, f(\tau)\, d\tau$$
$$+ (1 - F(T))\; p(\mathbf{S}, \mathbf{n}_1 + \mathbf{n}_2). \tag{8}$$

This expression suggests that $p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid T)$ can be evaluated by Monte Carlo integration in a simulation scheme similar to the one devised in the one-population case by Griffiths and Tavaré (1995). Analogous to the one-population case, Markov chain Monte Carlo simulations can be performed along Eq. (7) by specifying the following transition probabilities

$(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2) \to (\mathbf{S}, \mathbf{n}_1 - \mathbf{e}_k, \mathbf{n}_{3-i})$ with probability

$$\frac{n_{ki}(n_{ki}-1)}{f(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2)(n_1(n_1+\theta+1) + n_2(n_2+\theta+1))}$$

$(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2) \to (\mathbf{S}^l, \mathbf{n}_1, \mathbf{n}_2)$ with probability

$$\frac{\theta}{f(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2)(n_1(n_1+\theta+1) + n_2(n_2+\theta+1))}$$

$(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2) \to (\mathbf{S}^{kl}, \mathbf{n}_i + \mathbf{e}_j, \mathbf{n}_{3-i})$ with probability

$$\frac{\theta}{f(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2)(n_1(n_1+\theta+1) + n_2(n_2+\theta+1))} \tag{9}$$

where the conditions under which transitions of type 1, 2 and 3 are allowed are provided by Z_{1i}, Z_{2i} and Z_{3i} respectively and $f_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2)$ is defined as

$$f_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2) = \sum_{k=1}^{d} \frac{(n_{k1}-1)\, n_{k1} + (n_{k2}-1)\, n_{k2} + \theta(v_1+v_2)}{(n_1+\theta-1)\, n_1 + (n_2+\theta-1)\, n_2}, \tag{10}$$

$$v_i = |\{k : n_{ki}=1, n_{k(3-i)}=0, \mathbf{s}_{.l}=e_k \; or \; s_{.l}=e_k^c \text{ for some } l\}|.$$

While simulating along this Markov chain, time is kept track of by summing up deviates from the exponential distribution, i.e., the time to kth mutation or coalescence event is given by $t=\sum_{i=1}^{k} t_i$ where t_i is obtained by simulating an exponential distribution with parameter $(n_1(n_1-1)+n_2(n_2-1)+\theta(n_1+n_2))/2$ if there were n_i copies of the gene in population i at time $t_{(i-1)}$. Now, simulations can be performed along this Markov chain until $t \geqslant T$. Denote the configuration of the sample at time t by $(\mathbf{S}(t), \mathbf{n}_1(t), \mathbf{n}_2(t))$. Then

$$p_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid T) = E\left[p(S(T+), n_1(T+)+n_2(T+)) \times \prod_{j=1}^{N_T} f_2(\mathbf{S}(j), \mathbf{n}_1(j), \mathbf{n}_2(j)) \right]$$

which suggest the following estimator of p_2:

$$\hat{p}_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2 \mid T) = \frac{\sum_{i=1}^{v} \prod_{j=1}^{N_T} f_2(\mathbf{S}(j), \mathbf{n}_1(j), \mathbf{n}_2(j)) \prod_{j=1}^{\eta} f(\mathbf{S}(j), \mathbf{n}(j))}{v}. \tag{11}$$

In other words, the likelihood function of T can be evaluated by performing simulations of the Markov chain given by Eq. (9) while $t \leqslant T$ and along the Markov chain given by Eq. (3) when $t > T$ while evaluating $f_2(\mathbf{S}, \mathbf{n}_1, \mathbf{n}_2)$ and $f(\mathbf{S}, \mathbf{n})$ before and after T respectively.

Boundary conditions are given by $p(2, m)$ if $t > T$. While $t \leqslant T$, boundary conditions are obtained by taking the convolution of the distribution of the number of segregating sites which arose while the populations are isolated and the number of segregating sites which arose in the time before isolation. In particular

$$p_2(2, m \mid T) = \sum_{i=0}^{m} \frac{(\theta(T-t))^{m-i}\, e^{-\theta(T-t)}}{(m-i)!} \left(\frac{1}{1+\theta}\right)\left(\frac{\theta}{1+\theta}\right)^{i} = \frac{\begin{pmatrix} (\theta(T-t))^{m}\, e^{-\theta(T-t)} \\ \times F(1, -m; 1; -[(T-t)+\theta(T-t)]^{-1}) \end{pmatrix}}{(1+\theta)\, m!}, \tag{12}$$

(Mathematica, 1988), where F is the generalized hypergeometric function.

MULTIPLE POPULATIONS

Analogous to the two population case, a method for estimating the likelihood of an entire population tree with specified divergence times can be established. For r populations, a vector of population divergence times $\mathbf{T}=(T_1, T_2, ..., T_{r-1})$ is constructed such that the first entry specifies the time from the present where population 1 and population 2 merges, the second entry specifies the divergence time for population (1, 2) and 3, etc. The quantity of interest is the conditional probability of obtaining a set of samples from multiple populations given a specified topology of the population tree and the population divergence times ($\mathbf{T}$). Equations equivalent to Eqs. (3) and (7) can easily be established. For example, the conditional probability of observing a set of samples from r populations, given that the last mutation or coalescence happened at time τ before the last population divergence time (looking backward in time), is given by

$$p_r(\mathbf{S}, \mathbf{n}_1, ..., \mathbf{n}_r \mid \tau, T) = \frac{1}{\sum_{j=1}^{r} n_j(n_j-1+\theta)} \times \sum_{i=1}^{r} \begin{pmatrix} \sum_{k \in Z_{1i}} n_{ik}(n_{ik}-1)\, p_r(\mathbf{S}, \mathbf{n}_1, ..., \mathbf{n}_i-\mathbf{e}_k, ..., \mathbf{n}_r \mid T-\tau) \\ + \sum_{k \in Z_{2i}} \theta p_r(\mathbf{S}^l, \mathbf{n}_1, ..., \mathbf{n}_r \mid T-\tau) \\ + \sum_{k \in Z_{3i}} \theta p_r(\mathbf{S}^{kl}, n_1, ..., \mathbf{n}_i^k+\mathbf{e}_j, ..., \mathbf{n}_r \mid T-\tau) \end{pmatrix}. \tag{13}$$

The Markov chain Monte Carlo method for estimating the likelihood follows trivially from the derivation provided for two populations and will not be derived here. Note however, that the likelihood is evaluated by simulations through a Markov chain backwards in time until all populations have merged into one ancestral population. By evaluating the likelihood for different population trees, the maximum likelihood tree can be found just as in the case of single sequence phylogenetic inference.

APPLICATIONS

Two applications of the methods developed here will be presented. First, a maximum likelihood estimate of the divergence time between two African human populations will be obtained. Second, the population phylogeny of three Caribbean turtle populations will be estimated.

There has been considerable interest in the divergence of human populations in Africa (Vigilant *et al.*, 1989; Watson *et al.*, 1996; Tishkoff *et al.*, 1996). One of the reasons for this interest is that the ancestral human population can probably be traced to Africa. As an illustration of the methods for estimating population divergence times, the divergence time between the Mbuti population (formerly known as the eastern pygmies) and the Biaka population (formerly known as the western pygmies) will here be estimated. For this purpose previously published DNA sequences of the mitochondrial D-loop region from the two populations were obtained from GenBank (Mbuti: HUMMTDL073, HUMMTDL072, HUMMTDL071, HUMMTDL070, HUMMTDL069, HUMMTDL068, HUMMTDL067, HUMMTDL066, HUMMTDL065, HUMMTDL032, HUMMTDL031, HUMMTDL030, HUMMTDL006, HUMMTDL005, HUMMTDL004 and Biaka: HUMMTDL047, HUMMTDL046, HUMMTDL045, HUMMTDL044, HUMMTDL043, HUMMTDL042, HUMMTDL041, HUMMTDL040, HUMMTDL039, HUMMTDL038, HUMMTDL037, HUMMTDL002, HUMMTDL001). The sequences were aligned by ClustalV. Sequences with missing data were subsequently deleted. Likewise, sites with ambiguous alignment were also removed. The methods discussed above assume an infinite sites model. This implies that only one substitution is allowed to occur in each nucleotide site. However, the set of haplotypes discussed above is not consistent with the infinite sites model. In fact, certain parts of the D-loop region are rather saturated with substitutions. A binary data set, consistent with the infinite sites model, was therefore obtained from the aligned sequences by only considering transversional changes. The resulting data set is fully compatible with the infinite sites model (Table I).

The population size of the Mbutis and the Biakas appear to be almost the same. Both populations today consist of approximately 30,000 individuals. Furthermore, there is no evidence of changes in the effective population size from the demographic data or from the mismatch distribution (Watson *et al.*, 1996). Likewise, comparable levels of heterozygosity in the two populations also suggest that the effective population size is approximately the same (see, for example, Vigilant *et al.*, 1989). Therefore, in the following it will be assumed that θ (four times the effective population size times the mutation rate) for the Mbuti population is the same as θ for the Biaka population. There are therefore only two unknown parameters to estimate: T and θ (where T is the divergence of the two populations divided by the effective population size). The likelihood for different values of T and θ was evaluated by performing 1,000,000 runs through the Markov chain (see above). The maximum likelihood values for θ and T were estimated to 0.9 and 1.8 ($l = -43.3$) respectively. Assuming an effective diploid population size of 1,000 would then imply that the total rate of transversions in the D-loop region is approximately 0.00045. More importantly, the divergence time of the two populations would be approximately 1.800 generations. Assuming a generation time of 20, this translates into 36,000 years.

TABLE I

The Sequences of Sites Containing Tranversions in the Mbuti and Biaka mtDNA Control Region. The Last Two Columns Provide the Counts of Each Haplotype in Each of the Populations

Haplotype	Mbuti	Biaka
1 0 1 1 0 1 1 0 0	3	0
0 1 1 1 0 1 1 0 0	1	0
1 0 1 1 0 1 1 0 1	1	0
1 0 0 1 0 0 1 0 0	1	0
1 0 0 1 0 1 1 0 0	2	0
1 0 1 1 1 1 0 1 0	0	2
1 0 1 0 0 1 0 1 0	0	1
1 0 1 1 0 1 0 1 0	0	8
1 0 0 1 0 1 1 0 0	0	2

Another application of the method is the estimation of population phylogenies. This type of application will be demonstrated on a previously published data set on the Caribbean hawksbill turtle (Table II). Bass *et al.* (1996) collected data of the mitochondrial DNA control region of the Caribbean hawksbill turtle in order to test hypotheses on female nest site choice. They observed a high degree of isolation between different reproductive populations. Data from 3 populations (Table II) is applied here to provide an estimate of the population phylogeny of the three populations. As in the two-population case, it is assumed that all three populations, as well as the ancestral populations, have the same value of θ. The method described above can now be applied to

TABLE II

The Sequences of Variable Sites for Three Populations of the Caribbean Hawksbill Turtle. The Last Three Columns Provide the Counts of Each Haplotype in Each of the Three Populations.

Haplotype	USVI	Barbados	Mexico
0 0 0 0 0 0 0 0 0	1	11	0
1 0 0 0 0 1 0 1 1	0	1	0
0 1 0 0 0 0 0 0 0	0	3	0
1 0 1 0 1 1 1 1 1	14	0	0
1 0 1 1 0 1 1 1 1	0	0	2
1 0 1 0 0 1 1 1 1	0	0	13

estimate the likelihood for different values of θ and the two population divergence times (T_1 and T_2) for each of the three rooted population phylogenies. As in normal phylogenetic inference, the phylogeny with the highest likelihood is said to be most supported by the data. Notice that it is not obvious which population phylogeny is most supported by the data (Fig. 2a). Only the USVI (U.S. Virgin Islands) and the Barbados populations share a haplotype. Based on haplotype sharing measures these two populations should therefore be grouped together. However, the average number of nucleotide differences between the Mexican and the USVI population is much less than in any other comparison (Table II). Measures based on this type of information would therefore group the USVI and the Mexican population together. These two approaches differ in that the first approach assumes that the effect of mutation is negligible whereas the second type of approach assumes that genetic drift is of little importance. Which approach is the most reasonable depends on assumptions regarding the relative effect of mutation and drift. In contrast, the likelihood method introduced here applies all of the information in the data and accounts for both drift and mutation.

In order to estimate the population phylogenies, 1,000,000 runs through the Markov chain were performed for each value of T_1, T_2, and θ. A full search was completed for the three parameters for each of the population phylogenies. For the three population phylogenies (Mexico, (USVI, Barbados)), (USVI, (Mexico, Barbados), and (Barbados, (USVI, Mexico), maximum likelihood values of -44.5, -44.8, and -43.9, respectively, were estimated. The maximum likelihood values of T_1 and T_2 were approximately 0.75 and 1.0 (Fig. 2b). In other words, this method will group the USVI and the Mexico populations together. In this case, the results obtained by methods based on haplotype sharing would differ from the maximum likelihood solution. Notice that the maximum likelihood method is this case provides a unique estimate of the population phylogeny whereas methods based on haplotype trees provide no resolution and methods based on allele sharing would reach the wrong conclusion. However, also notice that the likelihood values differ only slightly. Obviously, there is only little information in this data set regarding the population phylogeny.

DISCUSSION

There are several advantages to the presented methodology. It allows maximum likelihood estimates of

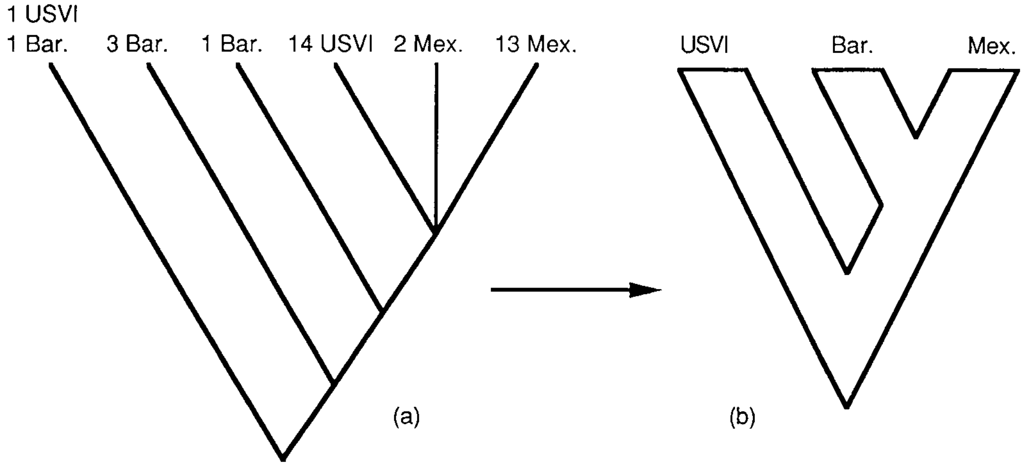

FIG. 2. (a) The maximum parsimony tree of the haplotypes shown in Table 2 and (b) the estimated population phylogeny using the coalescent likelihood approach. Notice that the parsimony tree provides no resolution of the relationship between populations. The number in front of the population name indicates how many of the particular haplotype are found in the population.

population divergence times with corresponding confidence intervals to be obtained. This in return solves the problem of ancestral polymorphisms while applying all of the information in the sample regarding the population phylogeny. The maximum likelihood method provides a unique estimate of the population phylogeny taking account of both mutation and drift. Other methods rely on assumptions regarding the relative effect of mutation and drift.

However, there are also several problems with this methodology. At the present stage the method is only implemented for the infinite sites model. Most data sets do not immediately conform to this model. This problem is richly illustrated by the analysis of the human mitochondrial DNA discussed above in which only transversional differences were considered. The method could, in principle, be implemented for finite sites models of DNA evolution (see, for example, Griffiths and Tavaré, 1994a, and the corresponding program SEQUENCE). Unfortunately, such models would most likely increase the computational time drastically. Since the method, even at the present stage, is computationally intensive it does not seem to be immediately feasible to perform this type of analysis for a finite sites model. However, for closely related populations where the data does conform to the infinite sites model, the presented method provides a strong alternative to classical approaches in the analysis of population divergence in models without migration.

ACKNOWLEDGMENTS

This study is supported by NIH grant GM40282 to Montgomery Slatkin and a personal grant to R.N. from the Danish Research Council. I thank M. Slatkin and two anonymous referees for comments and R. Hudson for important suggestions in the initial stages of this project.

REFERENCES

Avise, J. C. 1994. "Molecular Markers, Natural History and Evolution," Chapman & Hall, London/New York.

Bass, A. L., Good, D. A., Bjorndal, K. A., Richardson, J. I., Hillis, Z.-M., Horrocks, J. A., and Bowen, B. W. 1996. Testing models of female reproductive migratory behavior and population structure in the Caribbean Hawksbill turtle, *Eretmochelys imbricata*, with mtDNA Sequences, *Mol. Ecol.* **5**, 321–328.

Ethier, S. N., and Griffiths, R. C. 1987. The infinitely-many-sites model as a measure-valued diffusion, *Ann. Prob.* **15**, 515–545.

Griffiths, R. C. 1989. Genealogical-tree probabilities in the infinitely-many-sites model, *J. Math. Biol.* **27**, 667–680.

Griffiths, R. C., and Tavaré, S. 1994a. Simulating probability distributions in the coalescent, *Theor. Popul. Biol.* **46**, 131–159.

Griffiths, R. C., and Tavaré, S. 1994b. Ancestral inference in population genetics, *Stat. Sci.* **9**, 307–319.

Griffiths, R. C., and Tavaré, S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model, *Math. Biosci.* **127**, 77–98.

Horai, S., Satta, Y., Haysaka, K., Kondo, R., and others. 1992. Mans place in Hominoidea revealed by mitochondrial DNA genealogy, *J. Mol. Evol.* **35**, 32–43.

Hudson, R. R. 1991. Gene genealogies and the coalescent process, *in* "Oxford Surveys in Evolutionary Biology" (D. Futuyma and J. Antonovivs, Eds.), Vol. 7, pp. 1–44. Oxford Univ. Press, London.

Hudson, R. R. 1992. Gene trees, species trees and the segregation of ancestral alleles, *Genetics* **131**, 509–512.

Hudson, R. R., and Kaplan, N. L. 1986. On the divergence of alleles in nested subsamples from finite populations, *Genetics* **113**, 1057–1076.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, *Genetics* **61**, 893–903.

Kingman, J. F. C. 1982. The coalescent, *Stochast. Proc. Appl.* **13**, 235–248.

Kuhner, M. K., Yamato, J., and Felsenstein, J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling, *Genetics* **40**, 1421–1430.

Mathematica. 1988. Wolfram Research, Inc. Illinois.

Mountain, J. L., and Cavalli-Sforza, L. L. 1994. Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms, *Proc. Natl. Acad. Sci.* **91**, 6515–6519.

Patton, J. L., Dasilva, M. N. F., and Malcolm, J. R. 1994. Gene genealogy and differentiation among Arboreal Spiny Rats (Rodentia, Echimyidae) of the amazon basin——A test of the riverine barrier hypothesis, *Evolution* **48**, 1314–1323.

Slatkin, M., and Maddison, W. P. 1990. Detecting isolation by distance using phylogenies of genes, *Genetics* **126**, 249–260.

Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees, *Genetics* **122**, 957–966.

Tavaré, S. 1984. Lines-of-descent and genealogical processes, and their application in population genetics models, *Theor. Popul. Biol.* **26**, 119–164.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T., and Kidd, K. K. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins, *Science* **271**, 1380–1387.

Templeton, A. R., Routman, E., and Phillips, C. A. 1995. Separating population structure from population history—A cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the Tiger salamander, *Ambystoma tigrinum*, *Genetics* **140**, 767–782.

Vigilant, L., Pennington, R., Harpending, H., and Kocher, T. D. 1989. Mitochondrial DNA sequences in single hairs from a southern African population, *Proc. Natl. Acad. Sci.* **86**, 9350–9354.

Watson, E., Bauer, K., Aman, R., Weiss, G., von Haeseler, A., and Pääbo, S. 1996. mtDNA sequence diversity in Africa, *Am. J. Hum. Genet.* **59**, 437–444.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination, *Theor. Pop. Biol.* **7**, 256–276.

Wu, C. I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphism, *Genetics* **127**, 429–435.