Statistical Properties of a DNA Sample Under the Finite-Sites Model

Ziheng Yang

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

Manuscript received June 24, 1996
Accepted for publication August 29, 1996

ABSTRACT

Statistical properties of a DNA sample from a random-mating population of constant size are studied under the finite-sites model. It is assumed that there is no migration and no recombination occurs within the locus. A Markov process model is used for nucleotide substitution, allowing for multiple substitutions at a single site. The evolutionary rates among sites are treated as either constant or variable. The general likelihood calculation using numerical integration involves intensive computation and is feasible for three or four sequences only; it may be used for validating approximate algorithms. Methods are developed to approximate the probability distribution of the number of segregating sites in a random sample of n sequences, with either constant or variable substitution rates across sites. Calculations using parameter estimates obtained for human D-loop mitochondrial DNAs show that among-site rate variation has a major effect on the distribution of the number of segregating sites; the distribution under the finite-sites model with variable rates among sites is quite different from that under the infinite-sites model.

BECAUSE of its high evolutionary rate, the control (D-loop) region of the mitochondrial genome has been widely used in studies of human populations (see, *e.g.*, VIGILANT *et al.* 1991; WARD *et al.* 1991). Most population genetics models for analyzing DNA sequence polymorphisms were developed under the infinite-sites model (KIMURA 1969; WATTERSON 1975), which assumes that every mutation occurs at a different site in the sequence. Although this model may provide reliable approximations for nuclear DNAs or regions of the mitochondrial genome with low mutation rates, the assumption is clearly violated for human D-loop mitochondrial DNAs (mtDNAs). For example, in the 360-bp sequences of the D-loop region of 63 individuals from a North American Indian tribe, the Nuu-Chah-Nulth from Vancouver Island (WARD *et al.* 1991), the minimum number of changes on the most-parsimonious tree, which is an underestimate of the number of mutations in the sample, is 41, while the number of variable sites in the sample is 26, suggesting that many sites must have experienced more than one mutation. Nucleotide sites in the D-loop region appear to be under quite different selectional constraints, and as a result, their evolutionary rates are highly variable (TAMURA and NEI 1993; WAKELEY 1993). The among-site rate variation suggests that most evolutionary changes occur at a few sites in the sequence, while most other sites do not experience any substitutions at all. As the infinite-sites assumption is clearly violated for the mtDNA data, GRIFFITHS and TAVARÉ (1994b) devised algorithms for removing sites and/or individuals to make the data conform with the infinite-sites assumption.

Author e-mail: ziheng@mws4.biol.berkeley.edu

The statistical properties of a DNA sample under the finite-sites model with variable substitution rates among sites may be quite different from those under the infinite-sites model. Indeed, recent simulation studies generating sequences at the level of variation found in human mtDNAs have shown that among-site rate variation has significant effects on distributions of population genetics statistics (BERTORELLE and SLATKIN 1995; ARIS-BROSOU and EXCOFFIER 1996). For example, rate variation causes patterns in the distribution of pairwise differences that were previously attributed to demographic processes such as population expansion (BERTORELLE and SLATKIN 1995; see also LUNDSTROM *et al.* 1992). It shifts the distribution of the number of segregating sites in a DNA sample, invalidating TAJIMA's (1989) D statistic for testing neutrality, which was developed under the infinite-sites model. In one example, TAJIMA's D statistic rejected neutrality with as high a frequency as 23% when evolutionary rates varied among sites, although the data were generated under the neutral model (BERTORELLE and SLATKIN 1995).

This paper studies the distribution of a random sample of DNA sequences under the finite-sites model, with substitution rates assumed to be either constant or variable across sites. Special attention will be paid to the number of segregating sites. A Markov-process model is used for nucleotide substitution, so that multiple hits at one site are allowed. For its simplicity, good fit to real data and widespread use, the gamma distribution will be used to accommodate the among-site rate variation. This distribution involves a shape parameter α, which can be estimated quite reliably using phylogenetic methods (YANG 1993, 1994b; YANG and KUMAR 1996). An estimate for the D-loop region of human

mtDNAs that includes the two hypervariable segments was 0.17 (YANG and KUMAR 1996) and this value will be used in later calculations. When the gamma model is difficult to implement, a discrete rate-class model will be used instead. Results obtained from phylogenetic analyses suggest that the gamma model and the discrete rate-class model produce very similar results (YANG 1995).

TWO SEQUENCES

Constant rate for sites: We consider a random-mating large population of constant size N, without migration. The gene sequence has m completely linked sites. The haploid Wright-Fisher model is considered. With time measured in units of N generations, the coalescence time (t) for two sequences chosen at random from the population is exponentially distributed with mean 1 (see, *e.g.*, HUDSON 1990):

$$f(t) = e^{-t}, \quad t > 0. \tag{1}$$

A Markov process will be used for nucleotide substitution. Let the substitution rate per site per generation be μ, and $\theta = 2N\mu$. Estimates of θ from human mtDNAs were ~0.05 (see, *e.g.*, FU 1994; KUHNER *et al.* 1995), and this value will be used in later calculations. With time measured in units of N generations, the substitution rate per site per N generations will be $\mu = {}^1/_2\theta$. Let $p_{xy}(t)$ be the transition probability from nucleotide x to nucleotide y in time t; this can be calculated for a variety of substitution models (see, *e.g.*, TAVARÉ 1986; YANG 1994a). As an example, suppose that the probability of change between any two nucleotides is the same (JUKES and CANTOR 1969). Under this equal-probability model

$$p_{xy}(t) = \begin{cases} {}^1/_4 + {}^3/_4 \exp(-{}^4/_3\mu t), & \text{if } x = y, \\ {}^1/_4 - {}^1/_4 \exp(-{}^4/_3\mu t), & \text{if } x \neq y. \end{cases} \tag{2}$$

The probability that a site is variable (segregating) in two sequences separated time t ago is

$$q(t) = 1 - p_{xx}(2t) = {}^3/_4 - {}^3/_4 e^{-8/3\mu t}. \tag{3}$$

Conditional on t, the probability of observing S segregating sites in two sequences of length m is given by the binomial probability

$$f(S|\,t) = \binom{m}{S} q(t)^S (1 - q(t))^{m-S}, \; S = 0, 1, \ldots, m. \tag{4}$$

The unconditional distribution is then

$$f(S) = \int_0^\infty \binom{m}{S} q(t)^S (1 - q(t))^{m-S} e^{-t} dt \tag{5}$$

$$= \binom{m}{S} \sum_{i=0}^{m-S} \binom{m-S}{i} (-1)^i ({}^3/_4)^{i+S} \sum_{j=0}^{i+S} \binom{i+S}{j} \frac{(-1)^j}{{}^8/_3\, j\mu + 1}. \tag{6}$$

Equation 6 was obtained by binomial expansions of $(1 - q(t))^{m-S}$ and $q(t)^{i+S}$. In this study, however, Mathematica (WOLFRAM 1991) and a self-written C program were used to evaluate numerically the integral in Equation 5.

Gamma rates for sites: The substitution rate for a site is assumed to be a random variable drawn from a gamma distribution. The density of the distribution is

$$g(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}, \quad r > 0, \alpha > 0, \beta > 0, \tag{7}$$

where α and β are the shape and scale parameters, respectively. Because mutation rate μ and β are confounded, we set $\beta = \alpha$ so that the mean of the distribution is one, with the variance to be $1/\alpha$. The mutation rate for the entire sequence ($m\mu$) is then equal to that under the constant-rate model. The probability that a site is segregating in two sequences separated time t ago is

$$q(t) = \int_0^\infty ({}^3/_4 - {}^3/_4\, e^{-8/3\mu tr}) g(r)\, dr = {}^3/_4 - {}^3/_4 \left(\frac{\alpha}{\alpha + {}^8/_3\, \mu t}\right)^\alpha. \tag{8}$$

Note that this approaches Equation 3 when $\alpha \to \infty$. The distribution of S is

$$f(S) = \int_0^\infty \binom{m}{S} \left[{}^3/_4 - {}^3/_4 \left(\frac{\alpha}{\alpha + {}^8/_3\, \mu t}\right)^\alpha\right]^S \left[{}^1/_4 + {}^3/_4 \left(\frac{\alpha}{\alpha + {}^8/_3\, \mu t}\right)^\alpha\right]^{m-S} e^{-t} dt. \tag{9}$$

Figure 1 shows the distributions of S in two random sequences of $m = 500$ sites under three models. The distribution under the infinite-sites model is calculated using Equation 9.5 of TAVARÉ (1984), which is reproduced below as Equation 30, while those under the finite-sites models are calculated using Equations 5 and 9. As one may expect, relaxing the infinite-sites assumption, especially assuming gamma rates at sites, increases probabilities for small S and decreases those for large S. The constant-rate model gives a similar distribution of S as the infinite-sites model, but the gamma model is much more different. The variance of the distribution under the gamma model is only 35% that under the infinite-sites model. The present result conflicts with ROGERS (1992), who suggested that the infinite-sites model did not introduce much bias. The main reason for this discrepancy seems to be ROGERS' use of a much lower mutation rate, one estimated for the entire mitochondrial genome, although some of his discussions concerned the control region.

MORE THAN TWO SEQUENCES

General case: A sample of n sequences taken from the population are connected to a single common an-

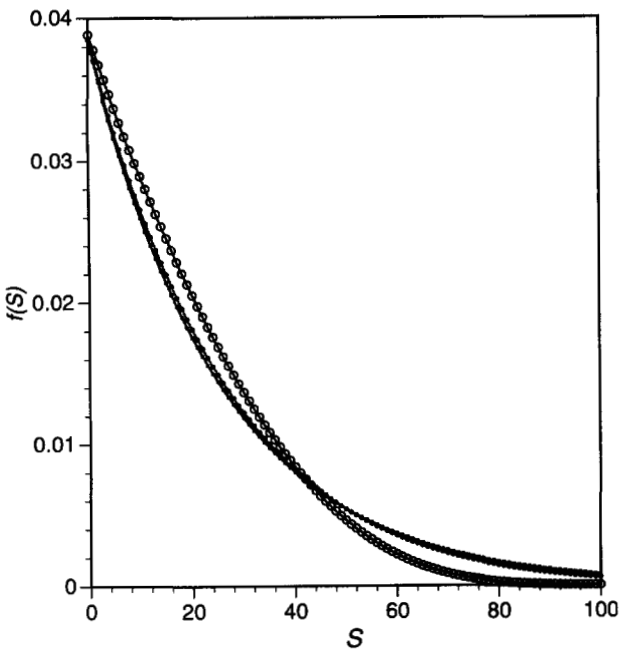

FIGURE 1.—The probability $f(S)$ of observing S segregating sites in a random sample of two sequences under three models: the infinite-sites model (■), the finite-sites model with constant substitution rate among sites (●), and the finite-sites model with the gamma distribution of rates among sites ($\alpha = 0.17$) (○). The sequence has $m = 500$ sites with $m\theta = 500 \times 0.05 = 25$. The means of the three distributions are 25, 23.4, and 18.3, respectively, while their variances are 645, 506.1, and 223.2.

cestor through an ancestral tree (genealogy) (Figure 2). Under the coalescent model, the time T_j during which there are j distinct ancestors approximately follows an exponential distribution with mean $1/\binom{j}{2} = 2/[j(j-1)]$, where time is measured in units of N generations (see *e.g.*, TAJIMA 1983; HUDSON 1990). The joint distribution of the waiting times is then

$$f(T_2, T_3, \ldots, T_n) = \prod_{j=2}^{n} \frac{j(j-1)}{2} \exp\left\{-\frac{j(j-1)}{2} T_j\right\}, \quad T_j > 0. \quad (10)$$

The distribution of node times $t_1, t_2, \ldots, t_{n-1}$ (Figure 2) can be obtained through a linear transform as

$$f(t_1, t_2, \ldots, t_{n-1}) = \frac{(n-1)!n!}{2^{n-1}} \exp\left\{-\sum_{j=1}^{n-1} jt_j\right\}, \quad t_1 > t_2 > \cdots > t_{n-1}. \quad (11)$$

As each of the $(n-1)!n!/2^{n-1}$ genealogies for n individuals has equal probability under the coalescent model, the joint density of genealogy G and its node times is

$$f(G, t_1, t_2, \ldots, t_{n-1}) = \exp\left\{-\sum_{j=1}^{n-1} jt_j\right\}. \quad (12)$$

Substitutions are assumed to occur independently among sites. The sample data can then be summarized

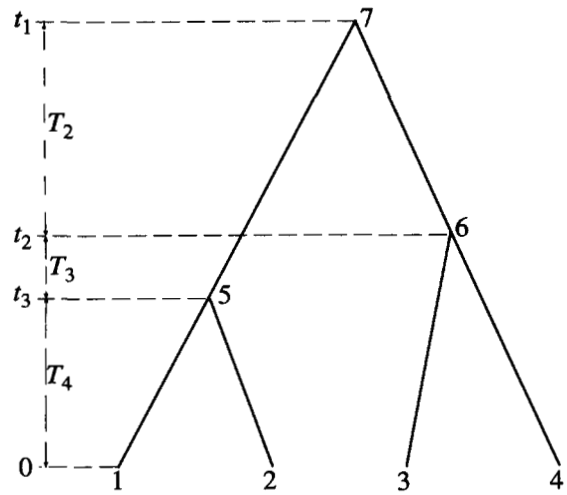

FIGURE 2.—An example genealogy (G_1) for four individuals. T_2, T_3, and T_4 are waiting times between coalescence events, while t_1, t_2, and t_3 are node times.

as counts of different site patterns, represented by the nucleotide compositions in different sequences. Let the nucleotides for the hth pattern be $X^{(h)} = \{x_{1h}, x_{2h}, \ldots, x_{nh}\}'$, where x_{jh} is the nucleotide in the jth sequence. Obviously there are 4^n possible site patterns for n sequences. Let the number of sites in the sample that have the hth pattern be m_h. The conditional probability of observing data $\mathbf{m} = \{m_1, m_2, \ldots, m_{4^n}\}$, given genealogy G and node times $t_1, t_2, \ldots, t_{n-1}$, is given by the multinomial probability

$$f(\mathbf{m} \mid G, t_1, t_2, \ldots, t_{n-1}) = C \prod_{h=1}^{4^n} [f(X^{(h)} \mid G, t_1, t_2, \ldots, t_{n-1})]^{m_h}, \quad (13)$$

where $C = m!/\prod_h m_h!$. The probability of data $\mathbf{m}$ is then an integration over the node times of a specific genealogical tree and a summation over all possible genealogies:

$$f(\mathbf{m}) = \sum_G \int_0^\infty \int_0^{t_1} \cdots \int_0^{t_{n-2}} [f(\mathbf{m} \mid G, t_1, t_2, \ldots, t_{n-1})\, f(G, t_1, t_2, \ldots, t_{n-1})] \times dt_{n-1} \cdots dt_2 dt_1. \quad (14)$$

For the example genealogy G_1 of Figure 2, the conditional probability of observing the hth site pattern in Equation 13 can be calculated as follows under the model of constant rate for sites (*e.g.*, FELSENSTEIN 1981):

$$f(X^{(h)} \mid G_1, t_1, t_2, t_3) = \sum_{x_{7h}} \sum_{x_{6h}} \sum_{x_{5h}} \pi_{x_{7h}} p_{x_{7h}x_{5h}} \times ((t_1 - t_3)\mu) p_{x_{7h}x_{6h}}((t_1 - t_2)\mu) \times p_{x_{5h}x_{1h}}(t_3\mu) p_{x_{5h}x_{2h}}(t_3\mu) p_{x_{6h}x_{3h}}(t_2\mu) p_{x_{6h}x_{4h}}(t_2\mu), \quad (15)$$

where $\pi_{x_{7h}}$ is the probability of observing nucleotide x_{7h} at the root (node 7) of the tree, given by the equilibrium frequency of the nucleotide; for the equal-proba-

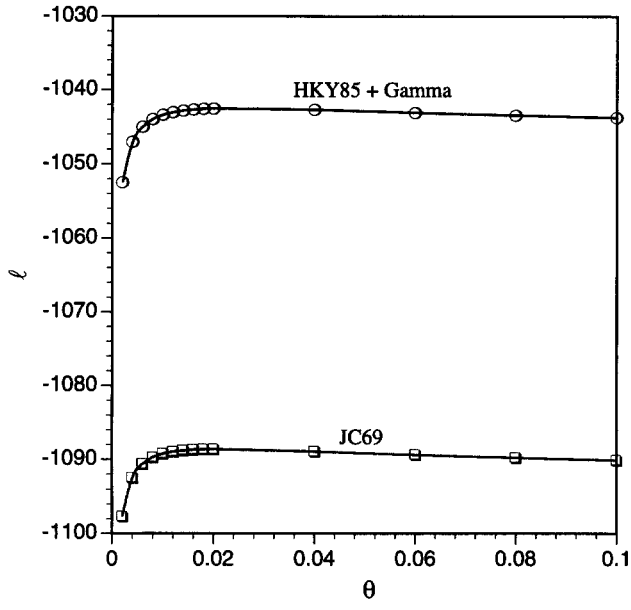

FIGURE 3.—The log likelihood function of θ under two evolutionary models: the equal-probability model of JUKES and CANTOR (1969) (JC69) and the substitution model of HASEGAWA *et al.* (1985) in combination with a discrete-gamma model of rates among sites (YANG 1994b) (HKY85 + Gamma). With the HKY85 model, the transition/transversion rate ratio (α/β in those authors' notation) is fixed at 35, and the gamma parameter is fixed at $\alpha = 0.17$; these are maximum likelihood estimates for this segment of mtDNA. Three West Pygmy sequences taken from the data of VIGILANT *et al.* (1991) (sequence numbers 1, 2, and 38) are used, and there are 684 aligned sites in the sequence. Twelve different site patterns are observed in the data, which are listed below, together with their numbers of occurrences:

West Pygmy (seq. 1):	T	C	A	G	C	A	T	G	G	A	C	A
West Pygmy (seq. 2):	T	C	A	G	C	G	T	G	G	A	C	G
West Pygmy (seq. 38):	T	C	A	G	T	A	C	C	A	G	A	G
	155	205	211	90	10	2	4	2	2	1	1	1

bility substitution model, $\pi_x = {}^1/_4$ for any x. The summations are taken over the unknown nucleotides at the three ancestral nodes. The conditional probabilities under other genealogical trees can be calculated similarly.

When rates are variable across sites, this conditional probability will be calculated as an average over sites, *i.e.*, over the rate distribution. The gamma model of rates for sites involves intensive computation (YANG 1993), and a discrete rate-class model is much simpler to use. Suppose there are K classes of sites, with frequencies $f_1, f_2, \ldots, f_K$ and rates $r_1, r_2, \ldots, r_K$. We have $\Sigma_k f_k = 1$ and $\Sigma_k f_k r_k = 1$. Conditional on rate r_k for the site, we have, for the genealogy of Figure 2,

$$f(X^{(h)} \mid G_1, t_1, t_2, t_3, r_k) = \sum_{x_{7h}} \sum_{x_{6h}} \sum_{x_{5h}} \pi_{x_{7h}} p_{x_{7h}x_{5h}}((t_1 - t_3)\mu r_k) \times p_{x_{7h}x_{6h}}((t_1 - t_2)\mu r_k) p_{x_{5h}x_{1h}}(t_3 \mu r_k) p_{x_{5h}x_{2h}}(t_3 \mu r_k) \times p_{x_{6h}x_{3h}}(t_2 \mu r_k) p_{x_{6h}x_{4h}}(t_2 \mu r_k), \quad (16)$$

so that

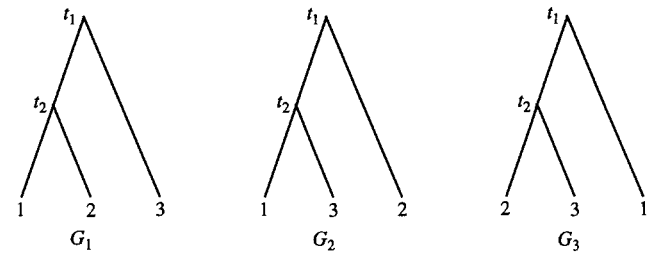

FIGURE 4.—Three genealogical trees and their node times for three individuals.

$$f(X^{(h)} \mid G_1, t_1, t_2, t_3) = \sum_{k=1}^{K} [f_k \times f(X^{(h)} \mid G_1, t_1, t_2, t_3, r_k)]. \quad (17)$$

The discrete-gamma model of YANG (1994b) uses K equal-probability categories to approximate the continuous gamma. In this case, $f_k = 1/K$, and the r_ks are calculated as functions of parameter α of the gamma distribution.

Figure 3 shows the probability of observing the data (the likelihood function) as a function of θ, with the integral of Equation 14 calculated numerically. Three West Pygmy sequences from the data of VIGILANT *et al.* (1991) were analyzed using two models: the equal-probability model of JUKES and CANTOR (1969) (JC69) and the model of HASEGAWA *et al.* (1985) with gamma rates among sites (HKY85+Gamma). The HKY85 model involves four more parameters than the JC69 model and accounts for both nucleotide frequency differences and transition/transversion rate bias. The maximum likelihood estimate of θ under the JC69 model is $\hat{\theta} = 0.021 \pm 0.015$ with $\ell = -1088.64$, while that under the HKY85+Gamma model is $\hat{\theta} = 0.024 \pm 0.018$ with $\ell = -1042.54$. Note that HKY85+Gamma fits the data significantly better than JC69 ($2\Delta\ell = 92.20$ compared with $\chi^2_{1\%} = 9.24$ with 5 d.f.). Obviously there is not much information for estimating θ in only three sequences and the standard errors of the estimates are quite large. It is, nevertheless, a general pattern that use of a simple model causes underestimation of θ.

Calculation of data probabilities by Equation 14 involves evaluation of an $(n - 1)$-dimensional integral for each genealogy G and summation over all possible genealogies. The approach of numerical integration adopted in this paper is not computationally feasible for data of more than four sequences (which involve three-dimensional integrals). It may be useful for validating approximate methods based on Monte Carlo integrations (FELSENSTEIN 1992; GRIFFITHS and TAVARÉ 1994a; KUHNER *et al.* 1995), as such algorithms are typically so complicated that their correctness is not always clear. When the model of JUKES and CANTOR (1969) is applied to the case of three sequences, however, some simplifications are possible, as shown below.

Three sequences: The three possible genealogies for three individuals are shown in Figure 4. The joint den-

TABLE 1

Site configurations in a sample of three DNA sequences and their probabilities under the three genealogies of Figure 4

i	Configuration	No. of sites (m_i)	Probability under G_1	Probability under G_2	Probability under G_3
0	*xxx*	m_0	$p_0(t_1,t_2)$	$p_0(t_1,t_2)$	$p_0(t_1,t_2)$
1	*xxy*	m_1	$p_1(t_1,t_2)$	$p_2(t_1,t_2)$	$p_2(t_1,t_2)$
2	*xyx*	m_2	$p_2(t_1,t_2)$	$p_1(t_1,t_2)$	$p_2(t_1,t_2)$
3	*yxx*	m_3	$p_2(t_1,t_2)$	$p_2(t_1,t_2)$	$p_1(t_1,t_2)$
4	*xyz*	m_4	$p_4(t_1,t_2)$	$p_4(t_1,t_2)$	$p_4(t_1,t_2)$

The p functions are defined in Equation 19.

sity of genealogy G and node times t_1 and t_2 is (See equation 12)

$$f(G, t_1, t_2) = e^{-(t_1+2t_2)}, \quad t_1 > t_2 > 0. \tag{18}$$

Under the substitution model of JUKES and CANTOR (1969), many site patterns have equal probability of occurrence. As a result, the data (nucleotide compositions) at a site can take one of five distinct configurations: *xxx*, *xxy*, *xyx*, *yxx*, and *xyz*, where *x*, *y*, and *z* represent any different nucleotides. These configurations will be labeled 0, 1, 2, 3, and 4, respectively. Conditional on genealogy G_1 of Figure 4 and its node times t_1 and t_2, the probabilities of observing these configurations at a site can be obtained as

$$\begin{aligned}
p_0(t_1, t_2) &= (1 + 3a + 6b + 6c)/16,\\
p_1(t_1, t_2) &= (3 + 9a - 6b - 6c)/16,\\
p_2(t_1, t_2) &= (3 - 3a + 6b - 6c)/16,\\
p_3(t_1, t_2) &= (3 - 3a + 6b - 6c)/16,\\
p_4(t_1, t_2) &= (6 - 6a - 12b + 12c)/16,
\end{aligned} \tag{19}$$

where

$$\begin{aligned}
a &= e^{-8/3t_2\mu},\\
b &= e^{-8/3t_1\mu},\\
c &= e^{-4/3(2t_1+t_2)\mu}.
\end{aligned} \tag{20}$$

Note that $p_2 = p_3$ and $\Sigma_{i=0}^{4} p_i = 1$.

The sample data can be summarized as counts of sites with the five configurations; let them be m_0, m_1, m_2, m_3, and m_4, respectively, with $\Sigma_i m_i = m$ to be the number of sites in the sequence. The conditional probability of observing data $\mathbf{m} = \{m_0, m_1, m_2, m_3, m_4\}$ is

$$f(\mathbf{m}|G_1, t_1, t_2) = C\prod_{i=0}^{4} p_i^{m_i}, \tag{21}$$

where $C = m!/\prod_{i=0}^{4} m_i!$.

The probabilities of observing the five site configurations under genealogies G_2 and G_3 (Figure 4) can be obtained similarly. As summarized in Table 1, these probabilities are p_0, p_2, p_1, p_2, p_4 for G_2, and p_0, p_2, p_2, p_1, p_4 for G_3, from which $f(\mathbf{m}|G_2, t_1, t_2)$ and $f(\mathbf{m}|G_3, t_1, t_2)$ are obtainable.

Assembling these results, we obtain the probability of observing data $\mathbf{m}$ in a random sample of three sequences as

$$\begin{aligned}
f(\mathbf{m}) &= \sum_G \int_0^\infty \int_0^{t_1} f(\mathbf{m}|G, t_1, t_2) f(G, t_1, t_2)\, dt_2 dt_1\\
&= C\int_0^\infty \int_0^{t_1} p_0^{m_0} p_4^{m_4} (p_1^{m_1} p_2^{m_2+m_3} + p_1^{m_2} p_2^{m_1+m_3} + p_1^{m_3} p_2^{m_1+m_2})\\
&\qquad \times e^{-(t_1+2t_2)}\, dt_2 dt_1.
\end{aligned} \tag{22}$$

With gamma rates over sites, the probabilities of Equation 19 are calculated by averaging over the gamma distribution, as $\int_0^\infty p_i(t_1 r, t_2 r) g(r)\, dr$. This means that Equations 19 and 22 still hold, but a, b, c of Equation 20 should be replaced by

$$\begin{aligned}
a &= \left[\frac{\alpha}{\alpha + {}^8/_3\, t_2\mu}\right]^\alpha,\\
b &= \left[\frac{\alpha}{\alpha + {}^8/_3\, t_1\mu}\right]^\alpha,\\
c &= \left[\frac{\alpha}{\alpha + {}^4/_3\, (2t_1 + t_2)\mu}\right]^\alpha.
\end{aligned} \tag{23}$$

The probability $f(S)$ of observing S segregating sites in three sequences of m sites can be calculated as the sum of probabilities of all configurations (m_0, m_1, m_2, m_3, m_4) that satisfy $m_0 = m - S$ or $m_1 + m_2 + m_3 + m_4 = S$. Equivalently, since

$$f(S|G, t_1, t_2) = \binom{m}{S} p_0^{m-S}(1 - p_0)^S, \tag{24}$$

$f(S)$ can also be calculated as

$$f(S) = \binom{m}{S} \int_0^\infty \int_0^{t_1} 3p_0^{m-S}(1 - p_0)^S e^{-(t_1+2t_2)}\, dt_2 dt_1. \tag{25}$$

For the case of constant rate for all sites, the approach used to derive Equation 6 can be used to obtain

$$f(S) = 3\binom{m}{S} \sum_{i=0}^{m-S} \binom{m-S}{i} \frac{(-1)^i}{16^{i+S}} \times \sum_{\substack{j_0,j_1,j_2,j_3 \geq 0\\ j_0+j_1+j_2+j_3=i+S}} \frac{(i+S)!}{j_0!j_1!j_2!j_3!} \frac{(-1)^{j_1+j_2+j_3} 15^{j_0} 3^{j_1} 6^{j_2+j_3}}{[{}^8/_3\,\mu(j_2+j_3) + 1][{}^8/_3\,\mu(j_1+j_2) + {}^{12}/_3\,\mu j_3 + 3]}. \tag{26}$$

Figure 5 shows $f(S)$ under the infinite-sites model and under the finite-sites model with substitution rates among sites assumed to be constant or gamma-distributed. The parameters used are the same as in Figure 1, except that the present sample contains three sequences instead of two. While the distribution under the constant-rate model is similar to that under the

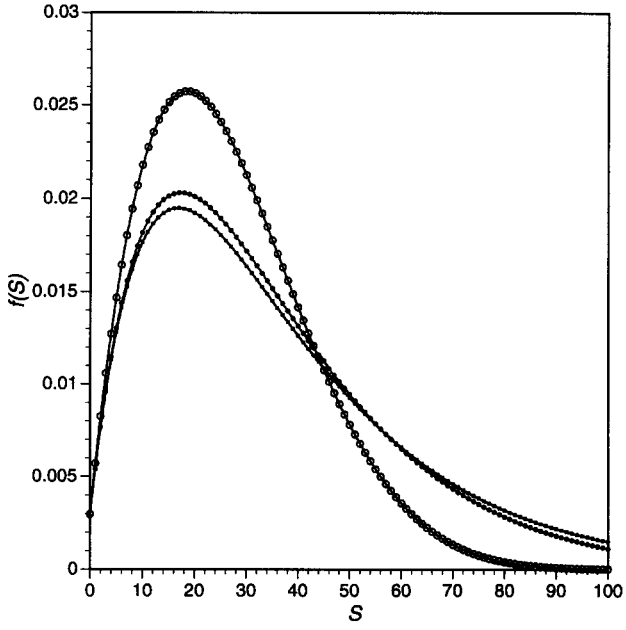

FIGURE 5.—The probability $f(S)$ of observing S segregating sites in a random sample of three sequences under three different models: the infinite-sites model (■), the finite-sites model with constant substitution rate among sites (●), and the finite-sites model with the gamma distribution of rates among sites ($\alpha = 0.17$) (○). The sequence has $m = 500$ nucleotides with $m\theta = 500 \times 0.05 = 25$. The means of the three distributions are 37.5, 34.9, and 26.8, respectively, while their variances are 817.8, 613.1, and 249.1.

infinite-sites model, the distribution with gamma rates is quite different. Relative to the infinite-sites model, the among-site rate variation together with the finite-sites assumption causes a 23% reduction in the mean of the distribution, and a 60% reduction in the variance.

APPROXIMATE DISTRIBUTION OF THE NUMBER OF SEGREGATING SITES

Constant rate among sites: The difference between a finite-sites model and an infinite-sites model lies in the allocation of mutations among sites; the total number of mutations follows the same distribution under the two models. As the distribution of the number of mutations, which is the number of segregating sites under the infinite-sites model, is known (WATTERSON 1975), this observation can be used to derive approximate methods to calculate the distribution of the number of segregating sites under the finite-sites model. We take a *segregating* site as a *mutated* site, a site that has been hit by at least one mutation. This approximation ignores the possibility that a site hit by two or more mutations may become a nonsegregating site. Since sequences for within-species comparison are very similar, such "back" mutations are expected to be rare. The accuracy of the approximation will be examined by computer simulation.

The density function of the total time on a genealogy, $T = \sum_{j=2}^{n} jT_j$, can be found through the convolution of the exponential variables (see Equation 10) as

$$f(T) = \frac{n-1}{2} e^{-(1/2)T}(1 - e^{-(1/2)T})^{n-2}. \quad (27)$$

Conditional on T, the number of substitutions (M) along the tree, which is the number of segregating sites under the infinite sites model, is a Poisson variable with parameter $m\mu T = {}^1/_2 m\theta T$. The probability generating function of M is then

$$G(z) = \mathcal{E}(z^M) = \int_0^\infty \exp\{-{}^1/_2\, m\theta T(1-z)\} f(T)\, dT$$
$$= (n-1)\int_0^1 y^{m\theta(1-z)}(1-y)^{n-2}dy \quad (28)$$
$$= \prod_{j=2}^{n} \frac{j-1}{j-1+m\theta(1-z)} \quad (29)$$

(WATTERSON 1975), from which the distribution of M is obtained as

$$p_M = \frac{1}{M!}\left.\frac{d^M G(z)}{dz^M}\right|_{z=0} = \frac{n-1}{m\theta}\sum_{i=0}^{n-2}(-1)^i\binom{n-2}{i} \times \left(\frac{m\theta}{m\theta+i+1}\right)^{M+1}, \quad M = 0, 1, \cdots \quad (30)$$

(TAVARÉ 1984: page 152). The probability of having no mutations ($M = 0$) is

$$\frac{(n-1)!\Gamma(m\theta+1)}{\Gamma(m\theta+n)}. \quad (31)$$

For large n, Formula 30 causes underflows and overflows and it is easier to evaluate numerically

$$p_M = \int_0^\infty \frac{({}^1/_2\, m\theta T)^M e^{-1/2 m\theta T}}{M!} f(T)\, dT$$
$$= \int_0^1 \frac{(-m\theta \log\{y\})^M}{M!}(n-1)y^{m\theta}(1-y)^{n-2}dy \quad (32)$$

(compare TAVARÉ 1994: page 153). This can also be obtained by differentiating Equation 28 with respect to z.

Let $H(M, S, m)$ be the probability that S sites are hit when M mutations are randomly assigned to m sites in the sequence. This is equivalent to the occupancy problem of putting M balls into m urns and is well studied by statisticians (JOHNSON and KOTZ 1977). Following JOHNSON and KOTZ (1977: page 115), we obtain the following recursive equation:

$$H(M, S, m) = \frac{S}{m} H(M-1, S, m) + \frac{m-S+1}{m} \times H(M-1, S-1, m), \quad (33)$$

with $H(0, 0) = 1$, and $H(M, S) = 0$ for any $M < S$. The first term on the right side is the probability that the first $M-1$ mutations occur at S sites and the Mth mutation occurs at one of these S sites too. The second term is the probability that the first $M-1$ mutations occur at $S-1$ sites and the Mth mutation occurs at

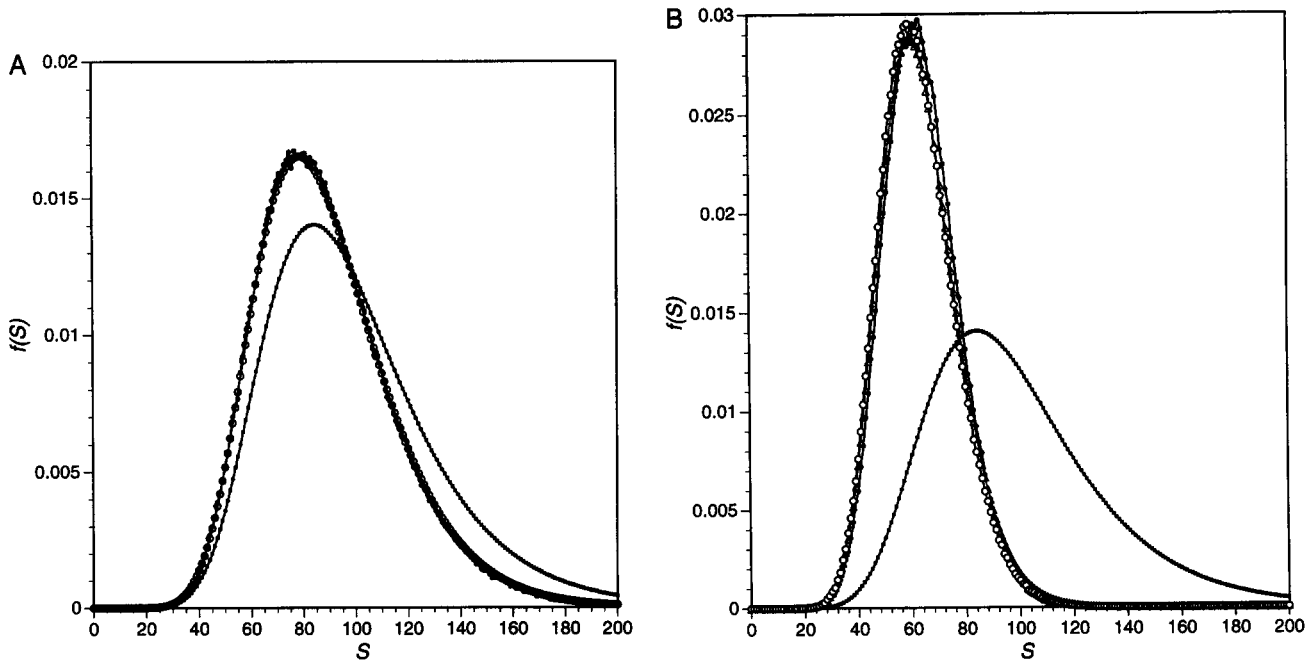

FIGURE 6.—The probability $f(S)$ of observing S segregating sites in a random sample of 30 sequences under the infinite-sites model (■) and under the finite-sites model with either constant substitution rate among sites (A) or variable rates among sites (B). The sequence has $m = 500$ sites with $m\theta = 500 \times 0.05 = 25$. (A) Computer simulations with 10^6 replicates (●) and the approximate method of this paper (Equation 34) (○) are used for the finite-sites model with constant rate among sites. (B) The distribution for the infinite-sites model (■) is the same as in A, shown here for comparison. Two models of variable rates among sites are used. One assumes two classes of sites in the sequence in the proportion 0.91057 and 0.08943 and with rates 0.33546 and 7.7663, respectively. The distribution under this model is obtained by both computer simulation (with 10^6 replicates) (○) and an approximate method (Equation 39) (△). Another model uses the gamma distribution for rates at sites with parameter $\alpha = 0.17$. The distribution under this model is obtained by computer simulation only (with 10^6 replicates) (●), since no approximate method is available. The means and variances of these distributions are listed in Table 2.

one of the $m - S + 1$ sites that have not been hit by any mutation. The recursive equation allows calculation of $H(M, S, m)$ for any reasonably large M and S, with m fixed.

The probability that S sites are segregating (or more precisely, are hit by mutations) in a sample of n sequences of length m can then be calculated as

$$f(S) \doteq \sum_{M=S}^{\infty} p_M H(M, S, m). \quad (34)$$

Alternatively, the probability that S sites are hit by mutations, conditional on the total time T on the genealogical tree, is

$$f(S|T) \doteq \binom{m}{S}(1 - e^{-(1/2)\theta T})^S e^{-(1/2)\theta T(m-S)}. \quad (35)$$

The unconditional probability can be obtained using $f(T)$ of Equation 27 as

$$f(S) \doteq \int_0^\infty f(S|T)f(T)\,dT = (n-1)!\binom{m}{S}\sum_{k=0}^{S}(-1)^k\binom{S}{k} \times \frac{\Gamma((m-S+k)\theta+1)}{\Gamma((m-S+k)\theta+n)}. \quad (36)$$

Variable rates among sites: A model of discrete rate classes will be considered. Since sites in the same rate class have the same mutation rate, the probability of the number of segregating sites in any rate class given the number of mutations in the class is known from results of the previous section. The probabilities of different allocations of mutations into different classes of sites are also easily obtainable. The probability of the number of segregating sites can then be calculated by considering how many mutations occur in each class of sites and how many sites in the class are hit by mutations. The case of two rate classes is considered here, as using three or more rate classes would require much more intensive computation.

Suppose that the first rate class has m_1 sites with rate r_1 and the second rate class has m_2 sites with rate r_2. We have $m_1 + m_2 = m$ to be the total number of sites in the sequence, and $m_1 r_1 + m_2 r_2 = m$, so that $m\mu$ is the total mutation rate for the sequence. We regard m_1 and m_2 as constants rather than random variables to simplify the analysis; this treatment should have only minor effect as long as the sequence is not very short. Conditional on the occurrence of a mutation, the probability that it hits a site in class one is

$$p_1 = \frac{m_1 r_1}{m_1 r_1 + m_2 r_2}, \quad (37)$$

and the probability that it hits a site in class two is $p_2 = 1 - p_1$.

TABLE 2

Mean and variance of the number of segregating sites in a random sample of 30 sequences

Model	Method	Curve in Figure 6	Mean	Variance
Infinite sites	Equation 30	■ in 6A and 6B	99.0	1105.9
Finite sites, constant rate	Simulation	● in 6A	88.5	675.7
	Approximation (Equation 34)	○ in 6A	89.0	706.3
Finite sites, two rate classes	Simulation	○ in 6B	62.5	195.8
	Approximation (Equation 39)	△ in 6B	63.4	209.5
Finite sites, gamma rates	Simulation	● in 6B	63.7	183.1

The full distributions and details of the models are given in the legend to Figure 6.

The probability that M mutations occur on the genealogical tree is given by p_M of Equation 30. Conditional on M, the probability that M_1 mutations occur at sites of the first rate class and $M_2 = M - M_1$ mutations occur at sites of the second class is given by the binomial probability

$$\binom{M}{M_1} p_1^{M_1}(1 - p_1)^{M-M_1}. \quad (38)$$

Within site class i ($i = 1, 2$), the probability that S_i sites are hit by the M_i mutations is $H(M_i, S_i, m_i)$. Summing over all combinations of S_1 and S_2 that satisfy $S_1 + S_2 = S$, generated by all combinations of M_1 and M_2, we obtain the probability of observing S segregating sites in a sample of size n as

$$f(S) \doteq \sum_{M=S}^{\infty} \sum_{M_1=0}^{M} \left\{ p_M \binom{M}{M_1} p_1^{M_1}(1 - p_1)^{M-M_1} \times \sum_{S_1=0}^{S} [H(M_1, S_1, m_1)H(M_2, S_2, m_2)] \right\}. \quad (39)$$

In previous calculations, a gamma parameter $\alpha = 0.17$ has been used. This estimate was obtained from maximum likelihood analysis of a data set of 25 D-loop human mtDNAs (YANG and KUMAR 1996). The log likelihood from this analysis is $\ell = -1624.1$. Fitting a two-rate-class model (YANG 1995) to the same data leads to the following maximum likelihood estimates: $\hat{f}_1 = 0.91057$ and $\hat{f}_2 = 0.08943$ for the proportions of the two rate classes, and $\hat{r}_1 = 0.33546$ and $\hat{r}_2 = 7.7663$ for the rates, with $\ell = -1622.4$. The two-rate-class model involves one more parameter than the gamma model and provides a slightly better fit to these data. (The constant-rate model has $\ell = -1676.0$ and fits the data much more poorly than the variable-rates models.) Parameter estimates for the two-rate-class model are used to calculate the distribution of the number of segregating sites in a random sample of 30 sequences. The distributions under the infinite-sites model and under the finite-sites constant-rate model are also calculated. Computer simulations are performed under the finite-sites models to check the accuracy of the approximation. The gamma model is also used in the simulation for comparison. The results are shown in Figure 6, A and B, and Table 2.

Since the approximate method counts some nonsegregating sites as segregating, we expect it to shift the distribution to the right and overestimate both the mean and variance of the distribution. This is indeed the case, as the approximate distributions have longer tails than the correct distribution obtained by simulation (Figure 6, A and B). The approximation is worse under the two-rate-class model than under the constant-rate model, probably because considerable back substitutions leading to nonsegregating sites have occurred at the fast-changing sites. Overall, the approximation appears to be quite accurate. Furthermore, we note that the distributions under the variable-rates models are quite different from that under the infinite-sites model. The variances of the former are less than $^1/_5$ that of the latter (Table 2). From JOHNSON and KOTZ (1977: page 109), the expected number of mutated sites, given the number of sites and the number of mutations, reaches the maximum when rates among sites are identical. Therefore the expected number of segregating sites in a sample under a variable-rates model should always be smaller than that under the constant-rate model. It is also noteworthy that the distribution under the gamma model is very close to that under the two-rate-class model.

As shown in Figure 7, A and B, the accuracy of the approximation seems to be high and does not seem to depend on the sample size (n). Figure 7 also shows that the variance under the infinite-sites model increases with n without bound, while that under the finite-sites model seems to approach a finite limit.

DISCUSSION

Analysis of the finite-sites model appears to be much more difficult than that of the infinite-sites model. Unfortunately, it is also quite clear from the results of this paper and previous simulation studies (BERTORELLE and SLATKIN 1995; ARIS-BROSOU and EXCOFFIER 1996) that the among-site rate variation, in combination with the finite-sites assumption, has substantial effects on analyses of D-loop mitochondrial data. We note that adding more complexity to the finite-sites model, such

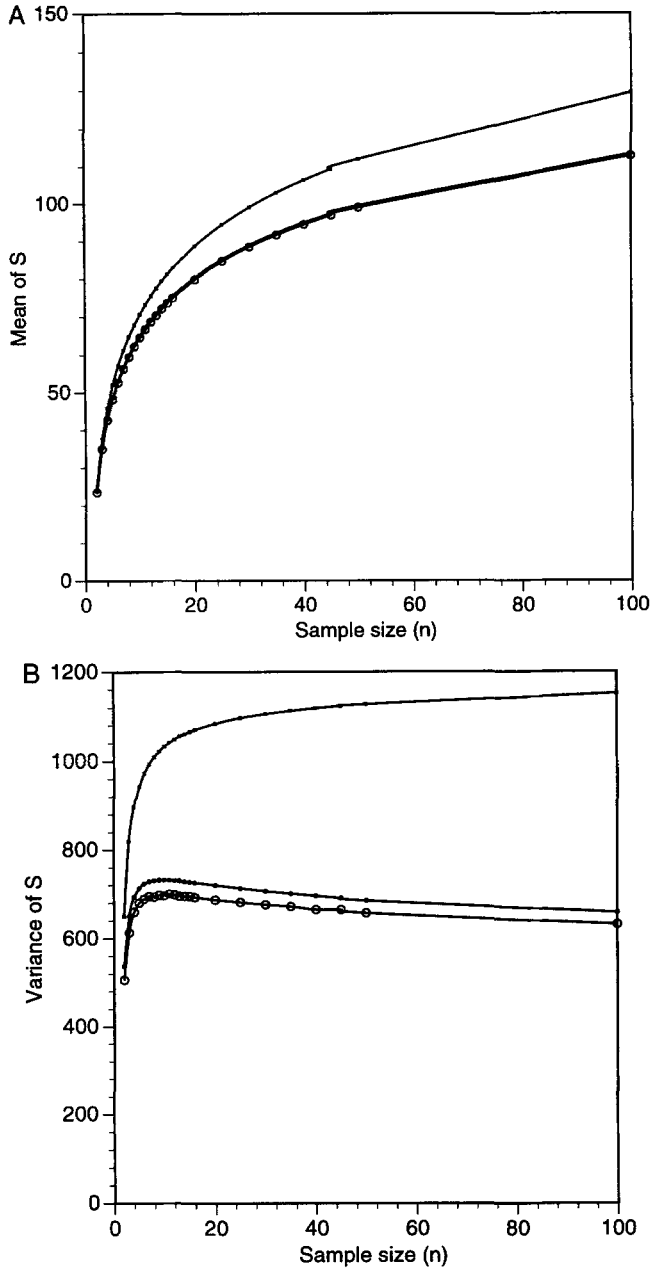

FIGURE 7.—The mean (A) and variance (B) of the number of segregating sites as functions of the sample size (n). The approximate method given by Equation 34 (●) is compared with computer simulation with 500,000 replicates (○). Results under the infinite-sites model (■) are also shown.

as using more realistic substitution models that account for base frequency differences and transition/transversion rate bias does not pose much additional difficulty in the analysis. These factors are, however, not as important as the among-site rate variation, from phylogenetic analyses of among-species data.

Previously, GRIFFITHS (1980) and O'BRIEN (1982) obtained the generating function of the number of segregating sites between two finite sequences, and GOLDING and STROBECK (1982) derived recursive equations for calculating the probability distribution of nucleotide site differences between two sequences. A constant rate was assumed for all sites. Results of this paper may be considered an extension to those results, both to variable mutation rates among sites and to more sequences. LUNDSTROM *et al.* (1992) fitted a model of two rate classes to some summary statistics for a sample of DNA sequences.

The effect of selection is not considered in this paper, and by adopting a coalescent approach, neutral evolution is assumed. However, the major reason for the existence of among-site rate variation is the different selectional constraints at different sites (see, *e.g.*, LI *et al.* 1985). The D-loop region of the mitochondrial genome, like other genes, is clearly under such selectional constraints (WAKELEY 1993). When natural selection is operating, the genealogy of the individuals may not be independent of the mutations they carry, and so the distributions of genealogies and node times used in this paper may not be accurate. Further studies considering more realistic models and incorporating factors such as migration and population size fluctuation are needed.

Addendum: After submission of this paper, TAJIMA (1996) published a study of the finite-sites model. While the present study aims to approximate the full distribution of the number of segregating sites in a random sample of n DNA sequences, TAJIMA derived analytical expressions for the expectation (but not the variance) under several substitution models, assuming either constant rate or gamma rates among sites. His general conclusion that the among-site rate variation has a substantial effect on analysis of D-loop mitochondrial data conforms with that of this paper.

I thank BRUCE RANNALA, MONTY SLATKIN, and THOMAS WIEHE for comments. I thank NAOYUKI TAKAHATA and two anonymous referees for a number of suggestions for improving the manuscript. One referee suggested derivation of Equations 6 and 26, while another suggested Equation 36 using a different approach. This study was supported by a grant from National Institute of Health (GM-40282) to M. SLATKIN.

LITERATURE CITED

ARIS-BROSOU, S., and L. EXCOFFIER, 1996 The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. **13:** 494–504.

BERTORELLE, G., and M. SLATKIN, 1995 The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. Mol. Biol. Evol. **12:** 887–892.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. Genet. Res. **60:** 209–220.

FU, Y.-X., 1994 A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

GOLDING, G. B., and C. STROBECK, 1982 The distribution of nucleotide site differences between two finite sequences. Theor. Pop. Biol. **22:** 96–107.

GRIFFITHS, R. C., 1980 Genetic identity between populations when mutation rates vary within and across loci. J. Math. Biol. **10:** 195–204.

GRIFFITHS, R. C., and S. TAVARÉ, 1994a Simulating probability distributions in the coalescent. Theor. Pop. Biol. **46:** 131–159.

GRIFFITHS, R. C., and S. TAVARÉ, 1994b Ancestral inference in population genetics. Statist. Sci. **9:** 307–319.

HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating the human-

ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology,* Vol. 7, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.

JOHNSON, N. J., and S. KOTZ, 1977 *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory.* Wiley, New York.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism,* edited by H. N. MUNRO. Academic Press, New York.

KIMURA, M., 1969 The number of heterogeneous nucleotide sites maintained in a finite population due to steady flux of mutation. Genetics **61:** 893–903.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

LI, W.-H., C.-I. WU and C.-C. LUO, 1985 Evolution of DNA sequences, pp. 1–94 in *Molecular Evolutionary Genetics,* edited by J. MACINTYRE. Plenum Press, New York.

LUNDSTROM, R., S. TAVARÉ and R. H. WARD, 1992 Modeling the evolution of the human mitochondrial genome. Math. Biosci. **112:** 319–335.

O'BRIEN, P., 1982 Allele frequencies in a multidimensional Wright-Fisher model with general mutation. J. Math. Biol. **15:** 227–237.

ROGERS, A., 1992 Error introduced by the infinite-sites model. Mol. Biol. Evol. **9:** 1181–1184.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

TAJIMA, F., 1996 The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics **143:** 1457–1465.

TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Pop. Biol. **26:** 119–164.

TAVARÉ, S., 1986 Some probabilistic and statistical problems on the analysis of DNA sequences, pp. 57–86 in *Lectures on Mathematics in the Life Sciences,* Vol. 17, American Mathematical Society, Providence, RI.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES and A. C. WILSON, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503–1507.

WAKELEY, J., 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

WARD, R. H., B. L. FRAZIER, K. DEW-JAGER and S. PÄÄBO, 1991 Extensive mitochondrial diversity within a single Amerindian tribe. Proc. Natl. Acad. Sci. USA **88:** 8720–8724.

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

WOLFRAM, S., 1991 *Mathematica: A System for Doing Mathematics by Computer,* 2nd ed. Addison-Wesley, Reading, MA.

YANG, Z., 1993 Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10:** 1396–1401.

YANG, Z., 1994a Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39:** 105–111.

YANG, Z., 1994b Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.

YANG, Z., 1995 A space-time process model for the evolution of DNA sequences. Genetics **139:** 993–1005.

YANG, Z., and S. KUMAR, 1996 Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13:** 650–659.

Communicating editor: N. TAKAHATA