# Modeling the Evolution of the Human Mitochondrial Genome

RON LUNDSTROM
*Collaborative Research, Inc., Waltham, Massachusetts 012154*

SIMON TAVARÉ
*Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089-1113*

AND

R. H. WARD
*Department of Human Genetics, School of Medicine, University of Utah, Salt Lake City, Utah 84112*

*Received 4 March 1992; revised 25 August 1992*

ABSTRACT

Mitochondrial DNA data have been used extensively to study evolution and early human origins. These applications require estimates of the rate at which nucleotide substitutions occur in the DNA sequence. We consider the problem of estimating substitution rates in the presence of site-to-site rate variation. A coalescent model is presented that allows for different substitution rates for purines and pyrimidines, as well as more detailed models that allow fast and slow rates within each of the purine and pyrimidine classes. A method for estimating such rates is presented. Even for these simple models of site heterogeneity, there are, typically, insufficient data to obtain reliable estimates of site-specific substitution rates. However, estimates of the average rate across all sites appear to be relatively stable even in the presence of site heterogeneity. Simulations of models with site-to-site variation in mutation rate show that hypervariable sites can produce peaks in the pairwise difference curves that have previously been attributed to population dynamics.

## 1. INTRODUCTION

The human mitochondrial genome, a circular molecule consisting of approximately 16,500 nucleotides, is one of the best understood molecular systems at the population level. The complete nucleotide sequence of human mitochondrial DNA (mtDNA) has been determined [1], and the most variable portions of the molecule have been sequenced in



 0025-5564/92/$5.00

several population studies [6,12,27,28]. In addition, the variability of the entire molecule has been assayed in populations worldwide, via restriction fragment length polymorphisms. This intense interest is due to the fact that mtDNA is ideally suited for the study of early human origins and microevolution. As mitochondria are maternally inherited, their transmission is consistent with haploid genetics, and hence the genealogy of mtDNA lineages is much simpler than that of nuclear DNA. Also, maternal inheritance implies that mtDNA is immune to sexual recombination, and changes that occur in mtDNA sequences must be due to mutation rather than a shuffling of the nucleotides during meiosis.

Another useful property of mtDNA for evolutionary studies is its increased rate of nucleotide substitution due to the lack of a mismatch repair apparatus in the replication mechanism. The mutation rate for mtDNA has been estimated to be five to ten times faster than the rate for nuclear genes [11], which means that mutations occur frequently enough for polymorphism to exist in modern populations. In general, mutations that have been used for evolutionary studies also appear to be selectively neutral; so far there is no conclusive evidence that these mtDNA mutations cannot be considered as neutral markers in the lineages in which they occur [2]. The rapid rate of evolution of mtDNA has been exploited to answer many questions about human evolution. Assuming a constant evolutionary rate, the number of mutations between diverse human groups has been used to obtain estimates of the time when the most recent common ancestor of all human mitochondria lived [4]. Other studies have used mtDNA to study the origin of ethnic groups, such as the timing and number of colonizations of the New World by Native Americans [23, 28].

Each of these applications requires a knowledge of the rate at which mutations occur in an mtDNA sequence. Estimates of this rate have been obtained by comparing a single DNA sequence from each of several species whose times of divergence are presumed known. Divergence is calculated from the number of nucleotide differences between species using one of several methods that correct for multiple mutations at a site, and rate estimates are obtained by dividing the sequence divergence by the divergence time [13,14]. This strategy suffers from several shortcomings. The divergence time between taxa is frequently unknown or is subject to significant error. Using interspecies differences may cause error if the substitution rate varies significantly between species. Also, regions of the mitochondrial genome appear to mutate often enough that multiple mutations at some sites make it impossible to estimate the actual number of substitutions that have occurred since the species diverged.

A more fundamental problem is posed by the possibility that different portions of the mitochondrial genome may evolve at different rates, and, in fact, each nucleotide site may have a distinct mutation rate. One solution would be to estimate the nucleotide substitution rate for each site individually. However, since information about mutation rates is obtained by replication over groups of similar sites, site-specific estimates are unreliable. An alternative approach postulates the existence of discrete classes of sites, each with a characteristic mutation rate. This reduces the number of parameters to be estimated, after which a statistical method to estimate these parameters from existing data sets can be developed. We have proposed a general method of rate estimation when sites are homogeneous [19]. In the present paper we show how this method can be extended to assess site heterogeneity. Section 2 introduces the coalescent model with mutation, and Section 3 gives a method of inference for this model. In Section 4, we illustrate the methods by analyzing a sample of 63 sequences from an Amerindian population.

## 2. THE COALESCENT PROCESS WITH MUTATION

When a sample of $n$ individuals is taken from a population, the ancestry of each individual can be traced backwards in time. The $n$ ancestral lineages in the sample will first coalesce to $n-1$ lineages and will continue coalescing until all $n$ lineages join together at a common ancestor. This ancestry can be represented by an inverted binary tree whose nodes represent the times at which the various lineages coalesce. Under the coalescent model [16], the amount of time that the sample has exactly $j$ ancestral lineages has an exponential distribution with parameter $\binom{j}{2}$, and each of the $\binom{j}{2}$ possible pairs of lineages is equally likely to coalesce. Time is measured in units of $N$ generations, where $N$ is the effective population size. For mitochondrial DNA the effective population size is equivalent to the number of females that contribute to the next generation, normally assumed constant over generations. This defines the distribution of a random ancestral tree that results from the sampling process.

Conditional on the ancestral tree determined by the sample, the distribution of a DNA sequence with $s$ sites is defined for each individual as follows. First, the ancestor at the top of the tree is assigned a sequence according to an initial distribution $\boldsymbol{\pi}$. Mutations occur at the $i$th site along each branch in the ancestral tree at the points of a Poisson process with rate $\theta_i/2$, the processes for distinct branches and distinct sites being independent.

Nucleotide changes at a given site occur according to a Markov

chain. At the $i$th site, the matrix $\mathbf{P}^{(i)} = \{p_{jk}^{(i)}\}$ gives the probability of changing to nucleotide $k$ when the site presently contains nucleotide $j$. Because the molecular character of each nucleotide differs, the probability of mutation at a particular site depends on the nucleotide occupying that site. Therefore the matrix $\mathbf{P}^{(i)}$ need not have $p_{jj}^{(i)} = 0$ for all $j$. We will assume that the process is stationary, as would be the case if each $\mathbf{P}^{(i)}$ were irreducible with stationary distribution $\boldsymbol{\pi}^{(i)}$, and $\boldsymbol{\pi} = \boldsymbol{\pi}^{(1)} \times \boldsymbol{\pi}^{(2)} \times \cdots \times \boldsymbol{\pi}^{(s)}$, corresponding to independent alllocation of nucleotides to the ancestral sequence. Mutations that occur in the tree generate a DNA sequence for each individual in the sample.

Since the distribution of the ancestral tree has no parameters, the process is completely determined by the mutation parameters $\theta_i$, $\boldsymbol{\pi}^{(i)}$, and $\mathbf{P}^{(i)}$, $i = 1, \ldots, s$. These parameters cannot all be estimated because of confounding between them. A unique set of parameters is obtained by using the rate matrices $\mathbf{Q}^{(i)}$ determined by

$$q_{jk}^{(i)} = \begin{cases} \theta_i p_{jk}^{(i)} & \text{if } j \neq k, \\ \theta_i\left(p_{jj}^{(i)} - 1\right) & \text{if } j = k. \end{cases}$$

In matrix notation this becomes

$$\mathbf{Q}^{(i)} = \theta_i(\mathbf{P}^{(i)} - \mathbf{I}).$$

The quantity $q_{jk}^{(i)}/2$ is interpreted as the rate at which nucleotide $k$ is substituted for nucleotide $j$, $j \neq k$, at site $i$.

The rate matrices $\mathbf{Q}^{(i)}$ are uniquely determined by the process and hence can be estimated. However, there are many values of $\theta_i$ and $\mathbf{P}^{(i)}$ corresponding to a given matrix $\mathbf{Q}^{(i)}$. For example, if $\tilde{\mathbf{P}}^{(i)} = [1/(1+\gamma)](\gamma\mathbf{I} + \mathbf{P}^{(i)})$ and $\tilde{\theta}^{(i)} = (1+\gamma)\theta^{(i)}$ for some $\gamma > 0$, then $\tilde{\mathbf{Q}}^{(i)} = \tilde{\theta}^{(i)}(\tilde{\mathbf{P}}^{(i)} - \mathbf{I}) = \mathbf{Q}^{(i)}$, and therefore $\tilde{\theta}^{(i)}$ and $\tilde{\mathbf{P}}^{(i)}$ define the same process (in distribution) as $\theta_i$ and $\mathbf{P}^{(i)}$. When values of $\theta_i$ and $\mathbf{P}^{(i)}$ are needed, we define

$$\theta_i = -\min_j q_{jj}^{(i)} \tag{1}$$

and

$$\mathbf{P}^{(i)} = \mathbf{I} + \mathbf{Q}^{(i)}/\theta_i. \tag{2}$$

The choice in (1) and (2) is particularly appropriate in the numerical analysis routines we need later.

In this model, $\mathbf{Q}^{(i)}$ can be any matrix with nonnegative off-diagonal entries and zero row sums, provided there is a unique stationary vector satisfying $\boldsymbol{\pi}^{(i)}\mathbf{Q}^{(i)} = \mathbf{0}$. If $u_{jk}^{(i)}$ are the substitution rates expressed as the probability of mutating from nucleotide $j$ to nucleotide $k$ in a single generation at the $i$th site, then $u_{jk}^{(i)} \approx q_{jk}^{(i)}/2N$ for $j \neq k$. Hence absolute rates can be obtained from relative rates if the effective population size $N$ is known, and vice versa.

As formulated, the model has too many parameters to be reasonably estimated even if many individuals are sampled. It is therefore necessary to consider special cases that reduce the number of parameters. The simplest model assumes that all sites have identical mutation rates, that is, $\mathbf{Q}^{(i)} \equiv \mathbf{Q}$. In this case, $\mathbf{Q}$ can be specified by from one to 12 parameters as in [13], [14], or [25]. Related models in the population genetics setting may be found in [8]–[10] and [20].

Since transversions are rarely observed in samples of mitochondrial DNA from local populations, an alternative model can be defined that assumes $\mathbf{Q}^{(i)} \equiv \mathbf{Q}^{(R)}$ for purine sites and $\mathbf{Q}^{(i)} \equiv \mathbf{Q}^{(Y)}$ for pyrimidine sites, where

$$\mathbf{Q}^{(Y)} = \begin{pmatrix} -\alpha_1 & \alpha_1 \\ \alpha_2 & -\alpha_2 \end{pmatrix}, \qquad \mathbf{Q}^{(R)} = \begin{pmatrix} -\beta_1 & \beta_1 \\ \beta_2 & -\beta_2 \end{pmatrix}, \tag{3}$$

with

$$\boldsymbol{\pi}^{(Y)} = \frac{1}{\alpha_1 + \alpha_2}(\alpha_2, \alpha_1), \qquad \boldsymbol{\pi}^{(R)} = \frac{1}{\beta_1 + \beta_2}(\beta_2, \beta_1). \tag{4}$$

In this case, $s_1$ of the sites contain purines and $s_2$ contain pyrimidines. $s_1$ and $s_2$ are known, as they are the observed numbers of purine and pyrimidine sites, respectively. This leaves the four parameters $\alpha_1$, $\alpha_2$, $\beta_1$, and $\beta_2$ to estimate.

Because sites may not be homogeneous, a more detailed model postulates the existence of fast and slow sites within each of the purine and pyrimidine classes. For the pyrimidine class, let $\mathbf{Q}^{(F)}$ denote the $2\times 2$ matrix for the fast sites and $\mathbf{Q}^{(S)}$ be the $2\times 2$ matrix for slow sites. Suppose that a site is fast with probability $f$ and slow with probability $1-f$, and, once classified, sites remain fast or slow thereafter. This is accomplished by taking

$$\mathbf{Q}^{(i)} = \begin{pmatrix} \mathbf{Q}^{(F)} & 0 \\ 0 & \mathbf{Q}^{(S)} \end{pmatrix}. \tag{5}$$

In this case, $\mathbf{Q}^{(i)}$ is not irreducible, but the extra parameter $f$ uniquely specifies the stationary distribution as

$$\boldsymbol{\pi}^{(i)} = \left(f\boldsymbol{\pi}^{(F)}, (1-f)\boldsymbol{\pi}^{(S)}\right). \tag{6}$$

A similar model can be used for the purine sites.

## 3. PARAMETER ESTIMATION

In this section we describe a general method for estimating substitution rates from sequence data and discuss how it might be applied to the models described earlier.

### 3.1. *GENERAL PRINCIPLES*

Ideally, we would base our inferential procedure on the probability $P_{n,\mathbf{x}}$ that the random sample of size $n$ has $x_j$ individuals of allele $j$, for $j = 1,2,\ldots,r$, where $r$ is the number of possible alleles in the model. For example, if there are $s$ sites in the sequence, each of which can be classified as a fast or slow nucleotide, then $r = 8^s$. While there is, in principle, a way to compute $P_{n,\mathbf{x}}$ [see Eq. (12), below], there is at present no computationally feasible method to calculate maximum likelihood estimates of the parameters using this likelihood. We explore the use of some alternative statistics that are computationally tractable and also provide a good summary of the information about substitution rates available in the data.

To this end, let $\mathbf{x} = (x_1, x_2, \ldots, x_r) \in \mathbf{Z}_+^r$ be a vector with $\sum x_j = n$, and let $V_{n,\mathbf{x}}$ be the fraction of the $s$ sites in the sample where $x_j$ individuals have the $j$th of the $r$ possible "alleles" at that *site*. For example, the alleles might correspond to the $r = 4$ possible nucleotides at a site or the $r = 8$ possible fast or slow nucleotides at a site. In any event, we use the set of statistics $\{V_{n,\mathbf{x}}\}$ to estimate the parameters of the model.

When summarizing the sequence data to obtain $\{V_{n,\mathbf{x}}\}$, the labels of each individual at each site are not recorded, preserving the information in individual sites but destroying joint information between sites. Sites collectively contain synergistic information about the phylogeny, so there is some loss of information associated with the $\{V_{n,\mathbf{x}}\}$ that is counterbalanced by their mathematical and computational tractability.

Parameter estimates can then be obtained by minimizing the squared error function

$$\sum_{\mathbf{x}} \left(V_{n,\mathbf{x}} - E[V_{n,\mathbf{x}}]\right)^2. \tag{7}$$

In the models described here, sites are classified into classes within which different sites behave in identical probabilistic fashion. It follows from the form of (7) that inferences about rates within different classes may be made separately.

As an alternative to least squares estimation, we might assume that sites are independent, in which case rates may once more be estimated separately within the classes. The likelihood for a given class is of the form

$$\sum_{\mathbf{x}} V_{n,\mathbf{x}} \log E[V_{n,\mathbf{x}}]. \tag{8}$$

The independent-sites model is likely to be a good approximation when the mutation rates are large, for then the role of phylogeny is washed out.

In either case, the expected value $E[V_{n,\mathbf{x}}]$ needs to be computed. Let $P_{n,\mathbf{x}}$ be the probability that a typical site has configuration $\mathbf{x}$, so that within each class,

$$E[V_{n,\mathbf{x}}] = P_{n,\mathbf{x}}. \tag{9}$$

In order to compute $P_{n,\mathbf{x}}$ we use the coalescent structure to derive a system of equations satisfied by the $P_{n,\mathbf{x}}$. Related methods are described in [15], [17], [19], and [22].

We look back at the ancestry of the sample to derive a recursion for the probabilities $P_{n,\mathbf{x}}$ that is determined by whether the most recent event in the sample's history is a mutation or a coalescence. Let $\theta$ be the mutation parameter at a single site, and let $\mathbf{P}$ be the mutation matrix. Define $(n,\mathbf{x}_{j,k})$ to be a sample having one more gene with allele $j$ and one fewer gene with allele $k$ at that locus than the sample with state $(n,\mathbf{x})$. If the last event before sampling was a mutation, the process could have been in state $(n,\mathbf{x}_{j,k})$ and could have experienced a mutation that changed some gene from $j$ to $k$. This happens with probabilities

$$\left(\frac{\theta}{\theta+n-1}\right)\left(\frac{x_j+1}{n}\right)p_{jk}P_{n,\mathbf{x}_{j,k}} \qquad \text{if } j \neq k \tag{10a}$$

and

$$\left(\frac{\theta}{\theta+n-1}\right)\left(\frac{x_k}{n}\right)p_{kk}P_{n,\mathbf{x}} \qquad \text{if } j = k. \tag{10b}$$

It is convenient to allow negative entries in $\mathbf{x}$, with the convention that $P_{n,\mathbf{x}}$ is zero in this case. Define $(n-1,\mathbf{x}_k)$ to be the state having one

fewer genes with allele $k$. If the last event before sampling was a coalescence, the process could have been in state $(n-1,\mathbf{x}_k)$ and a gene with allele $k$ was chosen to split. This happens with probability

$$\left(\frac{n-1}{\theta+n-1}\right)\left(\frac{x_k-1}{n-1}\right)P_{n-1,\mathbf{x}_k}. \tag{11}$$

Summing over all the possible states of the process at the last event before sampling gives

$$P_{n,\mathbf{x}} = \frac{\theta}{\theta+n-1}\sum_{j,k}\frac{x_j+1-\delta_{jk}}{n}p_{jk}P_{n,\mathbf{x}_{j,k}} + \frac{n-1}{\theta+n-1}\sum_{k}\frac{x_k-1}{n-1}P_{n-1,\mathbf{x}_k}. \tag{12}$$

The allele frequencies $P_{n,\mathbf{x}}$ are calculated from Eq. (12) together with the initial conditions $P_{1,\mathbf{e}_j}=\pi_j$ for $\mathbf{e}_j=(\delta_{1j},\ldots,\delta_{Kj})$, using an algorithm given in the Appendix. We emphasize that the recursion in (12) may be used to study *general* mutation structure for any number of alleles at a given locus. In this paper, however, we are primarily concerned with the case in which a locus is a particular site in the DNA sequence.

### 3.2. *APPLICATION TO FAST–SLOW MODELS*

We study first the purine–pyrimidine model of the previous section. Since the purine and pyrimidine sites can be analyzed separately, we illustrate the methods using the pyrimidine sites. Entirely analogous arguments apply to the purine sites. Now $V_{n,(x_1,x_2)}$ is the fraction of the *pyrimidine* sites that have $x_1$ C nucleotides and $x_2=n-x_1$ T nucleotides. Further,

$$E\left[V_{n,(x_1,x_2)}\right]=P^{(Y)}_{n,(x_1,x_2)},$$

where $P^{(Y)}_{n,(x_1,x_2)}$ is computed from Eq. (12) using the pyrimidine parameters $\mathbf{Q}^{(Y)}$.

Next we consider the model that allows fast and slow sites within the pyrimidine class. There are four types of pyrimidines: fast C, fast T, slow C, and slow T. Label these types 1, 2, 3, and 4, and let $P_{n,(x_1,x_2,x_3,x_4)}$ be the probability that $x_j$ individuals have type $j$, $1\leqslant j\leqslant 4$, at a particular site. Because a site cannot contain both fast and slow types, $P_{n,\mathbf{x}}=0$ unless $x_1=x_2=0$ or $x_3=x_4=0$. Let $P^{(F)}_{n,(x_1,x_2)}$ be the probabilities computed from Eq. (12) using the fast parameters $\mathbf{Q}^{(F)}$, and let $P^{(S)}_{n,(x_3,x_4)}$ be the probabilities computed from Eq. (12) using the slow parameters

$\mathbf{Q}^{(S)}$. Direct substitution into Eq. (12) shows that

$$P_{n,(x_1,x_2,0,0)} = fP^{(F)}_{n,(x_1,x_2)}$$

and

$$P_{n,(0,0,x_3,x_4)} = (1-f)P^{(S)}_{n,(x_3,x_4)}$$

satisfy the recursion when $\mathbf{Q}$ and $\boldsymbol{\pi}$ are given by Eqs. (5) and (6). Since the locations of fast and slow sites are unknown, the observed pyrimidine sites are a mixture of fast and slow. Hence,

$$E\left[V_{n,(x_1,x_2)}\right] = fP^{(F)}_{n,(x_1,x_2)} + (1-f)P^{(S)}_{n,(x_1,x_2)}. \tag{13}$$

To simplify the analysis, we will assume here that $\mathbf{Q}^{(S)} = h\mathbf{Q}^{(F)}$, that is, that the rate matrices for the fast and slow sites differ only by the scalar $h$. In particular, this implies that $\boldsymbol{\pi}^{(S)} = \boldsymbol{\pi}^{(F)}$. With this assumption, the model has four parameters: $f$, $h$, and two parameters that specify $\mathbf{Q}^{(F)}$. When $h = 0$, the slow sites have $\mathbf{Q}^{(S)} = 0$, with the interpretation that some fraction $1-f$ of the sites is fixed due to molecular constraints [5,7]. We have continued to take $\boldsymbol{\pi}^{(S)} = \boldsymbol{\pi}^{(F)}$ in this boundary case too, realizing that this assumption may be rather poor. Further investigation of this assumption seems worthwhile.

## 4. APPLICATIONS TO DATA

We illustrate our methods by analyzing a representative set of 63 mtDNA sequences from the Nuu-Chah-Nulth of Vancouver Island, British Columbia [28]. The sequence data, which are detailed in Figure 1 of [28], represent the first 360 nucleotides of the mitochondrial control region. In total there are 28 unique lineages defined by the occurrence of 26 variable sites. Of the 201 pyrimidine sites, 21 were variable, whereas only 5 of the 159 purine sites were variable. The contemporary traditional Nuu-Chah-Nulth population numbers some 2400 individuals, of whom 600 are females of child-bearing age. Apart from a decline in population that occurred immediately following European contact in the late 18th century, the archeological and ethnographic data suggest that the ancestral population leading to the contemporary Nuu-Chah-Nulth was relatively stable during most of the past 8000 years. Hence, despite the short-term demographic fluctuations that undoubtedly occurred, we have taken the long-term effective population size of this tribe to be 600 females, with upper and lower bounds of 900 and 300, respectively [19]. Also, like many tribal groups, the Nuu-Chah-Nulth are subdivided into distinct bands, ranging in size

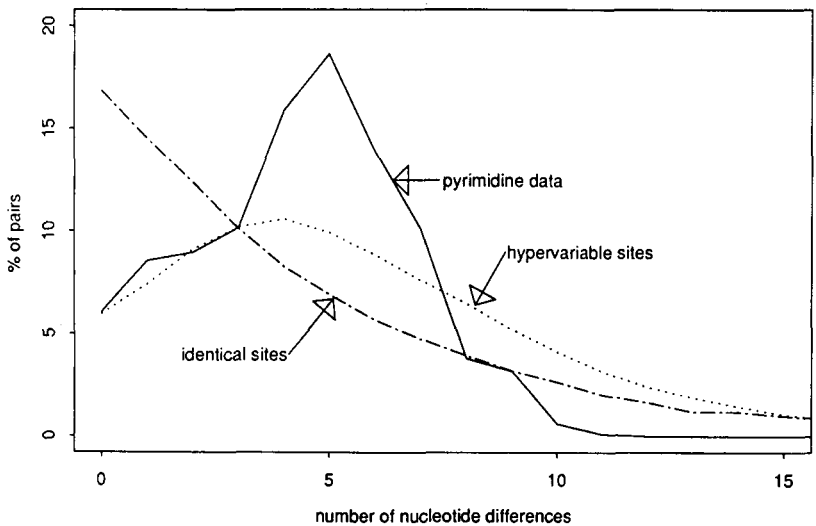


FIG. 1. Distribution of pairwise differences in actual and simulated data.

from 40 to 600 individuals. However, as expected for haploid molecules, the aggregation of the Nuu-Chah-Nulth into local bands exerts little influence on the distribution of mitochondrial lineages because the distribution of pairwise differences was random with respect to band affiliation [26]. In order to analyze sequences that would represent the full spectrum of diversity within this tribe, the 63 individuals in the sample were chosen from 13 of the 14 extant bands in such a way that they were maternally unrelated for four generations. Hence, the 63 sequences can be considered to be a random sample of the Nuu-Chah-Nulth population as it existed in the mid-19th century [28].

As a baseline for comparison, we used the purine–pyrimidine model with identical sites within each class and obtained estimates of the four parameters [19]. These estimates are given in Table 1. Confidence intervals were obtained empirically from simulations. For the identical-sites model, the squared error and likelihood functions given by Eqs. (7) and (8) have unique global minima and maxima, respectively.

Using the estimates given by the identical-sites models, samples were simulated via the coalescent model, and their properties were compared with those of the actual data. These values predict 4.97 variable purine sites and 21.21 variable pyrimidine sites, in agreement with both the data and the simulations. The total number of mutations observed in the simulated data set had a 5th percentile of 19 and a 95th percentile of 62, while the number of state charges had a 5th percentile of 13 and

TABLE 1

Estimates of Substitution Rates in the Mitochondrial Control Region Assuming Identical Sites Within Purines and Pyrimidines[a]

| Rate parameter | Rate estimate | Lower bound | Upper bound |
|---|---|---|---|
| $\alpha_1$ (C → T) | 0.02 | 0.01 | 0.04 |
| $\alpha_2$ (T → C) | 0.03 | 0.02 | 0.06 |
| $\beta_1$ (A → G) | 0.005 | 0.002 | 0.015 |
| $\beta_2$ (G → A) | 0.014 | 0.004 | 0.05 |

[a]See Lundstrom et al. [19].

a 95th percentile of 46. Note that the latter pair bracket the minimum number of 41 mutations in the observed data as inferred by parsimony. However, the simulated data deviate from the actual data in two important respects. First, the distribution of pairwise differences in the actual data had a deficiency of identical or closely related sequences compared to the simulations. This creates a peak in the distribution of pairwise differences as shown in Figure 1, rather than the decaying distribution predicted by the identical-site model [24]. Second, the actual data had an excess of distinct sequences compared to the simulated data. This was most evident in the pyrimidines, where the data defined 24 distinct sequences and the simulations defined only 9–17.

In an attempt to rectify this lack of fit, we used the fast–slow model. For the purine sites, the estimates are the same as in the identical-site model. That is, the model estimates that all sites are fast and the fast rates are (0.005,0.014), as in [19]. Since there were only five variable purine sites, we lack sufficient data for reliable parameter estimation.

The situation is quite different for the pyrimidine data. In the fast–slow model, the objective function (7) has many local minima and a relatively flat contour, making it numerically difficult to calculate the parameter estimates. Plotting $\alpha_1$ (the C to T rate) against $\alpha_2$ (the T to C rate) reveals that the local minima all lie near the line $\alpha_2 = 1.5\alpha_1$, the condition that gives the correct nucleotide frequencies. Table 2 gives the value of the objective function along the line $\alpha_2 = 1.5\alpha_1$. The smallest estimates appear when all sites are classified as fast with rates $(\alpha_1, \alpha_2) = (0.02, 0.03)$, the estimates given by the identical-sites model. As the estimates of the fast rate increase by a factor of 7 to (0.14,0.22), the percentage of slow sites increases to 80%, and the slow sites are invariant. As the fast rates increase above this point, the slow sites have rates 0.04 to 0.01 times that of the fast sites. The overall best fit occurs

TABLE 2

Estimates of Substitution Rates in Pyrimidine Sites using the Fast–Slow Model

| Fast sites | | | Slow sites | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha_1$ | $\alpha_2$ | No. of sites | $\alpha_1$ | $\alpha_2$ | No. of sites | Mean rate | Squared error | Alleles | No of var pos* |
| 0.02 | 0.03 | 197 | 0.00 | 0.00 | 4 | 0.0238 | 4.980e-04 | 13.3 | 21.4 |
| 0.04 | 0.06 | 106 | 0.00 | 0.00 | 95 | 0.0255 | 4.900e-04 | 13.7 | 21.5 |
| 0.06 | 0.08 | 80 | 0.00 | 0.00 | 121 | 0.0269 | 4.851e-04 | 14.0 | 20.9 |
| 0.08 | 0.12 | 60 | 0.00 | 0.00 | 141 | 0.0291 | 4.801e-04 | 14.6 | 21.0 |
| 0.10 | 0.15 | 52 | 0.00 | 0.00 | 149 | 0.0309 | 4.779e-04 | 15.3 | 21.0 |
| 0.12 | 0.18 | 46 | 0.00 | 0.00 | 155 | 0.0327 | 4.772e-04 | 15.6 | 21.5 |
| 0.14 | 0.21 | 42 | 0.00 | 0.00 | 159 | 0.0344 | 4.776e-04 | 16.2 | 21.4 |
| 0.18 | 0.27 | 25 | 0.01 | 0.01 | 176 | 0.0337 | 4.819e-04 | 15.9 | 21.1 |
| 0.24 | 0.37 | 15 | 0.01 | 0.02 | 186 | 0.0333 | 4.859e-04 | 16.1 | 21.4 |
| 0.34 | 0.52 | 7 | 0.01 | 0.02 | 194 | 0.0311 | 4.929e-04 | 15.1 | 20.8 |
| 0.41 | 0.62 | 5 | 0.02 | 0.02 | 196 | 0.0306 | 4.943e-04 | 15.1 | 21.2 |
| 0.62 | 0.94 | 2 | 0.02 | 0.03 | 199 | 0.0284 | 4.971e-04 | 14.4 | 21.1 |
| 0.70 | 1.05 | 1 | 0.02 | 0.03 | 200 | 0.0276 | 4.976e-04 | 13.9 | 20.9 |
| 2.00 | 3.04 | 6 | 0.02 | 0.03 | 195 | 0.0890 | 5.765e-04 | 25.3 | 23.1 |

*var pos, variable positions

when 77% of the sites are invariant and the fast rates are (0.12,0.18). In spite of many local minima, the average rate across all sites given in column 7 varies from only 0.023 to 0.034.

To determine how well the values in Table 2 modeled the data, we simulated 1000 samples using each set of parameters and compared the simulations with the data. The mean number of distinct sequences and variable sites in the simulations appear in columns 9 and 10 of Table 2. The fast–slow model increased the number of distinct sequences from 14 to 16, still short of the 24 observed in the data. Also, the pairwise difference curve still shows the exponential shape that is characteristic of the identical-sites model, rather than the peaked distribution of the data.

More extreme estimates in the fast–slow model do have the ability to explain both the enhanced number of distinct sequences and the peaked pairwise difference curve. The last row in Table 2 shows the effect of six hypervariable sites with rate 100 times that of the remaining slow sites. In this case, the value of the least squares objective function is somewhat higher, but the simulations show the correct number of alleles and the pairwise difference distributions show a peak. This is shown in Figure 1.

The poor fit of the curve with six hypervariable sites to actual data is not surprising, as our methods attempt to fit the distribution of the statistics $\{V_{n,\mathbf{x}}\}$ rather than the pairwise difference curve. A peak in the pairwise difference distribution has been explained by population dynamics not included in the model, such as migration or a population expansion [21,24]. However, this analysis shows that a peak can also result from the effects of site heterogeneity.

To investigate the possibility of hypervariable sites in the data, we omitted the seven most variable pyrimidine sites in the data, which reduced the number of alleles to 14, the number predicted by the model simulations, but only slightly changed the pairwise difference distribution. As changes in population size and hypervariable sites are both plausible, the observed data are very likely to have resulted from a combination of both effects.

The estimate of the mean rate across all sites given in column 7 of Table 2 is fairly stable and is equal to the mean rate estimated by the identical-sites model. Even with hypervariable sites, where the mean rate increases dramatically, the increase is due to the extremely large value attributed to the six hypervariable sites, while the majority of sites have rate comparable to the mean rate for the other parameter values. Hence the simplest identical-site model can reasonably estimate the mean rate across all sites in spite of site-to-site variability.

## 5. CONCLUSIONS

We have introduced a model for analyzing DNA sequence data from a population. This model allows for site heterogeneity; that is, it allows sites in the sequence to have distinct substitution rates. We have also proposed a method of rate estimation for this model based on the statistics $\{V_{n,\mathbf{x}}\}$, which record the distribution of base frequencies at each site. Although these are not sufficient statistics, they provide an adequate summary of the information about substitution rates available in the data. The distribution of the statistics $\{V_{n,\mathbf{x}}\}$ lends itself well to numerical calculation and therefore provides a means for estimation of substitution rates. Ideally, we would like to use maximum likelihood estimation based on the distribution of the counts $\{V_{n,\mathbf{x}}\}$. Although this is theoretically possible, it is not yet computationally feasible.

Allowing distinct rates for each site creates too many parameters to be estimated unless sites are grouped into classes. We have grouped sites into purines and pyrimidines and have further allowed fast and slow sites within each of these groups. Even for this simple model, very large sample sizes are needed for accurate estimation of rates. Although larger sample sizes may increase the precision of the estimates, longer

regions of DNA are also needed to provide larger groups of similar sites. However, when longer sequences are available, there is no guarantee that the additional sites will not introduce additional site heterogeneity. The 360 base pairs of the control region in the Amerindian data are flanked by less variable regions, so that extending the sequence will not necessarily provide more information. Therefore, there may never be sufficient data to estimate site-specific substitution rates in this portion of the molecule without a more detailed understanding of the molecular function of individual sites.

In spite of the difficulties in estimating site-specific substitution rates, the simplest model, which assumes that all sites are identical, can reasonably be used to estimate the average rate across all sites. In the applications of the fast–slow model, the estimates of average rate remained relatively constant even when estimates of the fast and slow rates were variable. Hence the estimates obtained from the identical-sites model are useful for modeling most aspects of the data, though they may not adequately describe subtleties such as the number of distinct sequences.

It is important to emphasize that the effects of site heterogeneity were found, on the basis of simulations, to be qualitatively similar to effects that have been attributed to population dynamics. Peaks in the pairwise difference curve can be generated by an expanding population or by hypervariable sites. In the latter case, the hypervariable sites must have rates greater than 1 and work in concert with a group of sites with rates less than 1. Under these conditions, the number of distinct sequences in the sample is greatly enhanced. Even a small number of hypervariable sites could move the peak in the pairwise difference curve and cause significant error in estimates of population dynamics based on pairwise differences. Similarly, we have found that population fluctuations in conjunction with the finite-sites model can have a marked influence on the position of the peak in the pairwise difference curve; see [18]. A model that simultaneously accounts for variable population size and site-specific rate variability will be necessary to discern the effects of each.

## APPENDIX

We will show that Eq. (12) has a unique solution and therefore characterizes the probabilities $P_{n,\mathbf{x}}$. Let

$$l_n = \binom{n+r-1}{r-1}$$

be the number of ordered $r$-tuples of nonnegative integers that add to

$n$. Also, let $\mathbf{y}_n$ be an $l_n$-vector that contains the probabilities $P_{n,\mathbf{x}}$ for a fixed value of $n$ and the $l_n$ possibilities of $\mathbf{x}$. Since Eq. (12) is linear in $P_{n,\mathbf{x}}$, it can be written as

$$\left[I - \frac{\theta}{\theta + n - 1}\mathbf{A}_n\right]\mathbf{y}_n = \mathbf{B}_{n-1}\mathbf{y}_{n-1}, \tag{A1}$$

where $\mathbf{A}_n$ is a matrix of size $l_n \times l_n$ and $\mathbf{B}_{n-1}$ has size $l_n \times l_{n-1}$. Moreover, a column sum of the matrix $\mathbf{A}_n$ is given by

$$\sum_{j,k} \frac{x_k}{n} p_{kj} = \sum_k \frac{x_k}{n} = 1,$$

which means that $\mathbf{A}_n^T$ is a stochastic matrix. Hence $[\mathbf{I} - [\theta/(\theta + n - 1)]\mathbf{A}_n]$ is diagonally dominant and nonsingular. This proves the uniqueness of the solution to (12).

Equation (A1) also suggests a numerical method for computing $\mathbf{y}_n$. Since $\mathbf{A}_n$ and $\mathbf{B}_{n-1}$ are sparse matrices, an iterative method can be used to solve (A1) for $\mathbf{y}_n$. If $\mathbf{L}$, $\mathbf{D}$, and $\mathbf{U}$ are the lower triangular, diagonal, and upper triangular pieces of $\mathbf{A}_n$, one form of the Gauss–Seidel iterates for solving (A1) is given by

$$\mathbf{z}_{i+1} = \left(\mathbf{I} - \frac{\theta}{\theta + n - 1}\mathbf{L}\right)^{-1}\left[\frac{\theta}{\theta + n - 1}(\mathbf{U} + \mathbf{D})\mathbf{z}_i + \mathbf{B}_{n-1}\mathbf{y}_{n-1}\right].$$

The iteration matrix

$$\frac{\theta}{\theta + n - 1}\left(\mathbf{I} - \frac{\theta}{\theta + n - 1}\mathbf{L}\right)^{-1}(\mathbf{U} + \mathbf{D})$$

has spectral radius less than $\theta/(\theta + n - 1)$; hence the Gauss–Seidel iterates $\mathbf{z}_i$ converge with geometric rate $\theta/(\theta + n - 1)$ to the solution of (A1) [3]. Note that our choice of $\theta$ and $\mathbf{P}$ in (1) and (2) optimizes the convergence rate. Computer storage is not needed for the matrices $\mathbf{A}_n$ or $\mathbf{B}_{n-1}$ because the necessary non-zero entries are easily calculated when needed. Two arrays of length $l_n$ are necessary to store the solution and the right-hand side of (A1).

*This research was supported in part by National Science Foundation grant DMS90-05833, by National Institutes of Health grant GM41746, and by a Fellowship from the Program in Mathematics and Molecular Biology at the University of California at Berkeley, which is supported by the National Science Foundation under grant DMS87-20208. The Government has*

*certain rights in this material. We wish to thank Dick Hudson and two reviewers for helpful comments.*

REFERENCES

1 S. Anderson, A. Bankier, B. Barrell, M. deBruijn, A. Coulson, J. Drouin, I. Eperon, D. Nierlich, B. Roe, F. Sanger, P. Schreier, A. Smith, R. Staden, and I. Young, Sequence and organization of the human mitochondrial genome, *Nature* 290:457–465 (1981).

2 J. C. Avise, Mitochondrial DNA and the evolutionary genetics of higher animals, *Phil. Trans. Roy. Soc. Lond. B* 312:325–342 (1986).

3 R. Burden, J. Faires, and A. Reynolds, *Numerical Analysis*, Prindle, Weber, and Schmidt, Boston, 1978.

4 R. Cann, M. Stoneking, and A. Wilson, Mitochondrial DNA and human evolution, *Nature* 325:31–36 (1987).

5 G. A. Churchill, A. von Haeseler, and W. C. Navidi, Sample size for a phylogenetic inference, *Mol. Biol. Evol.*, 9:753–769 (1992).

6 A. Di Rienzo and A. C. Wilson, Branching pattern in the evolutionary tree for human mitochondrial DNA, *Proc. Natl. Acad. Sci. USA* 88:1597–1601 (1991).

7 W. M. Fitch and E. Margoliash, A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case, *Biochem. Genet.* 1:65–71 (1967).

8 G. Golding and C. Strobeck, The distribution of nucleotide site differences between two finite sequences, *Theor. Popul. Biol.* 22:96–107 (1981).

9 R. C. Griffiths, Allele frequencies in multidimensional Wright–Fisher models with a general symmetric mutation structure, *Theor. Popul. Biol.* 17:51–70 (1980).

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

11 D. Hartl and A. Clark, *Principles of Population Genetics*, Sinauer Associates, Sunderland, Mass., 1989.

12 S. Horai and K. Hayasaka, Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA, *Am. J. Hum. Gen.* 46:828–842 (1990).

13 T. Jukes and C. Cantor, Evolution of protein molecules, in *Mammalian Protein Metabolism*, H. Munro, Ed., Academic, New York, 1969, pp. 21–123.

14 M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *J. Mol. Evol.* 2:87–90 (1980).

15 J. F. C. Kingman, *Mathematics of Genetic Diversity* (Regional Conf. Ser. Appl. Math., Vol. 34), SIAM, Philadelphia, 1980.

16 J. F. C. Kingman, On the genealogy of large populations, *J. Appl. Probab.* 19A:27–43 (1982).

17 R. Lundstrom, Stochastic models and statistical methods for DNA sequence data, Ph.D. Thesis, Univ. Utah, 1990.

18 R. Lundstrom, The coalescent process when population size varies, with application to DNA sequences, submitted.

19 R. Lundstrom, S. Tavaré, and R. H. Ward, Estimating mutation rates from molecular data using the coalescent, *Proc. Natl. Acad. Sci. USA* 89:5961–5965 (1992).

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

21 A. Rogers and H. Harpending, Population growth makes waves in the distribution of pairwise genetic differences, *Mol. Biol. Evol.*, 9:552–569 (1992).

22 S. Sawyer, D. Dykhuizen, and D. Hartl, Confidence interval for the number of selectively neutral amino acid polymorphisms, *Proc. Natl. Acad. Sci. USA* 84:6225–6228 (1987).

23 T. Schurr, S. Ballinger, Y. Gan, J. Hodge, D. A. Merriwether, D. Lawrence, W. Knowler, K. Weiss, and D. Wallace, Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages, *Am. J. Hum. Genet.* 47:613–623 (1990).

24 M. Slatkin and R. Hudson, Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations, *Genetics* 129:555–562 (1991).

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

26 D. Valencia, Mitochondrial DNA evolution in the Nuu-Chah-Nulth, M.S. Thesis, Univ. Utah, 1992.

27 L. Vigilant, R. Pennington, H. Harpending, T. Kocher, and A. C. Wilson, Mitochondrial DNA sequences in single hairs from a South African population, *Proc. Natl. Acad. Sci. USA* 86:9350–9354 (1989).

28 R. H. Ward, B. L. Frazier, K. Dew, and S. Pääbo, Extensive mitochondrial diversity within a single Amerindian tribe, *Proc. Natl. Acad. Sci. USA* 88:8720–8724 (1991).