Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 5961–5965, July 1992
Evolution

Estimating substitution rates from molecular data using the coalescent

(molecular evolution/population genetics/mutation rates/mitochondrial DNA)

RON LUNDSTROM*, SIMON TAVARÉ†, AND R. H. WARD*

*Department of Human Genetics, Eccles Institute of Human Genetics, 10 North 2030 East Street, Room 2100, Building 533, Salt Lake City, UT 84112; and †Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113

Communicated by John C. Avise, March 6, 1992 (received for review September 19, 1991)

ABSTRACT **A coalescent model is used to estimate the rate at which neutral substitutions occur in a DNA sequence, without the necessity for an independent estimate of divergence times. Given a random sample of molecular sequences from a finite population, the distribution of the time to a common ancestor can be obtained from the coalescent model. With this principle, summary statistics are developed that use the distribution of molecular diversity within the sample to estimate the relative magnitude of nucleotide substitution rates. If, in addition, the effective population size is known, absolute substitution rates can also be estimated. These techniques are illustrated by estimating the transition rates that underlie the evolution of the first 360 nucleotides of the mitochondrial control region in an Amerindian tribal population.**

The ability to obtain valid estimates of nucleotide substitution rates is a fundamental problem for evolutionary biology. At the level of population genetics, estimates of substitution rates are essential for evaluating the effects of random drift in a population or for calculating the probability of fixation of a mutant allele. At the level of "macro evolution", substitution rates are required to estimate the evolutionary time span since taxa diverged. In addition, the distribution of rates among different evolutionary lineages or among different genomic regions is fundamental to understanding how the tempo of evolution varies. The rate of substitution that characterizes a particular genomic region also defines its relevance for phylogenetic analysis. However, in spite of the central role that substitution rates play in the neutral theory of evolution, numerical estimates are often compromised by a lack of appropriate statistical methods or by a lack of sufficient data.

In the area of molecular evolution, estimates of substitution rates are traditionally 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 substitutions at a site, and rate estimates are obtained by dividing the sequence divergence by the divergence time (1–3). This strategy suffers from several shortcomings. The divergence time between taxa is frequently unknown or subject to significant error. Using interspecific differences may cause error when the substitution rate varies significantly between species. Also, the method applies only to a single (consensus) sequence from each species and does not readily extend to samples of several sequences from each species.

Although these problems may introduce only a slight bias when estimating global substitution rates for distantly related taxa, the situation is quite different when estimating the rate of neutral substitutions in the context of population genetics. When dealing with intraspecific molecular data, estimates of the divergence time between individuals within a population are not generally available. In those cases when individuals are connected by a pedigree, the individuals are too closely related for significant evolutionary changes to have occurred. Even when divergence times can be estimated, as for geographic isolates, the relative error is substantial. Lacking information about divergence between individuals, population geneticists have estimated rates using the sampling structure of a finite population (4, 5).

The relationship between divergence times and population structure is explicitly given by the coalescent model (6, 7), which gives the statistical distribution of the time to a common ancestor of a random sample. This distribution, which replaces the estimates of divergence time in the interspecies analysis, can be used in conjunction with a mutation model, assuming either infinite sites or finite sites (8, 9). The infinite-sites model requires that no site experience more than one substitution and that all types of substitutions occur at the same rate at every site—conditions violated by most molecular data sets. Hence, it is more appropriate to use the finite-sites model in conjunction with the coalescent to account for the fact that multiple substitutions can occur, sites experience different substitution rates, and transitions occur more frequently than tranversions.

With the advent of molecular techniques, such as the PCR, it is now possible to obtain a comprehensive set of molecular sequences from population-based samples (10–12). The large sample sizes provided by such surveys increase the resolving power of the data, allowing the substitution process to be studied in more detail. As a consequence, individual rates for each type of nucleotide substitution can be estimated, as well as a single "overall" substitution rate. We describe methods that use the finite-sites model to estimate substitution rates based on sequence data from a random sample and compare their performance against an extension of the traditional pairwise difference (PD) estimator. These techniques are illustrated by estimating the substitution rates in the 5′ end of the mitochondrial control region by using a sample of 63 sequences from a single North American Indian tribe (12).

Materials and Methods

The ancestry of a random sample of n DNA sequences from a finite population of effective size N can be modeled by the coalescent (6), which has two components. The first component generates the genealogy of the n sequences by giving the distribution of time between nodes in the ancestral tree. Under the coalescent model, the amount of time that the sample has j distinct ancestors is an exponential random variable with parameter $\binom{j}{2} = j(j-1)/2$, where time is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: LS, least-squares; IS, independent sites; MSE, mean square error; PD, pairwise difference.

measured in units of N generations. The second component of the coalescent superimposes nucleotide substitutions on the branches of a given ancestral tree. Substitutions occur at rate $\theta/2$ along each branch, where θ is the scaled substitution parameter. For haploids, $\theta = 2Nu$, where u is the probability of a neutral substitution per site per generation. We restrict attention to selectively neutral single nucleotide substitutions and do not consider insertion or deletion events.

When a nucleotide substitution occurs at a site that presently contains nucleotide i, it is changed to nucleotide j with probability p_{ij}. The product $\theta p_{ij} = q_{ij}$ $(i \neq j)$ gives the rate at which sites currently containing nucleotide i mutate to nucleotide j. The rate matrix $\mathbf{Q} = \theta(\mathbf{P} - \mathbf{I})$ uniquely defines the substitution process. For DNA sequences, $\mathbf{Q}$ is a 4×4 matrix representing substitutions among the four nucleotides. In our model, $\mathbf{Q}$ can be any irreducible matrix with nonnegative off-diagonal entries and zero row sums, which encompasses the one- and six-parameter models given in refs. 1 and 3. The goal of this paper is to estimate the substitution rates q_{ij}. If u_{ij} are the substitution rates expressed as the probability of mutating from nucleotide i to nucleotide j in a single generation at a given site, then $u_{ij} \approx q_{ij}/2N$ for $i \neq j$. Hence absolute rates can be obtained from relative rates when the effective population size N is known. Associated with the matrix $\mathbf{Q}$ is the stationary distribution $\boldsymbol{\pi}$, which satisfies $\boldsymbol{\pi}\mathbf{Q} = 0$. The ith component of $\boldsymbol{\pi}$ gives the frequency of the ith nucleotide in the ancestral sequence at the top of the tree, so that the expected number of changes from nucleotide i to nucleotide j along a single branch of length t is $q_{ij}\pi_i t/2$.

Estimating Rates from the Coalescent

The estimation methods described below assume that a random sample is available from a finite population. A sequence of sites, each site having identical substitution matrix, is taken from each individual in the sample. The proportion of sites with a given distribution of nucleotides can be used as a statistic to summarize the sequence data. Let $\mathbf{x} = (x_1, x_2, x_3, x_4)$ be a vector with $\Sigma x_j = n$ and let $V_{n,\mathbf{x}}$ be the fraction of the s sites in the sample, where x_j individuals have nucleotide j at that site. Parameter estimation is based on the set of statistics $\{V_{n,\mathbf{x}}\}$. The expected value $E[V_{n,\mathbf{x}}]$ is equal to the probability $P_{n,\mathbf{x}}$ that a single site has configuration $\mathbf{x}$. Parameter estimates are obtained by minimizing the squared error function

$$\sum_{\mathbf{x}} (V_{n,\mathbf{x}} - P_{n,\mathbf{x}})^2, \qquad [1]$$

which can be classified as a method-of-moments procedure because it involves equating the fraction of sites with a given configuration to its expected value and solving [in the least-squares (LS) sense] for the parameters of interest. Hence, we call the estimators obtained by this procedure LS estimators. Although the LS method does not assume sites are independent, it also does not explicitly use any information about dependencies between sites.

The second method explicitly assumes that sites are independent and maximizes the resulting log-likelihood, which is proportional to

$$\sum_{\mathbf{x}} V_{n,\mathbf{x}} \log(P_{n,\mathbf{x}}). \qquad [2]$$

Eq. 2 would be the likelihood for the data if sites were phylogenetically independent—that is, if each site had its own ancestral tree and each tree were statistically independent of all other trees. However, because all sites share a common (random) ancestry generated by the coalescent, they are *not* independent. Because the estimators obtained by this method are not true likelihood estimators, we use the name independent-sites (IS) estimators. Since these estimators need not have the desirable properties usually associated with maximum likelihood—such as consistency, asymptotic normality, or asymptotic minimum variance—we rely on simulation techniques to assess the behavior of both procedures.

For either method, the probabilities $P_{n,\mathbf{x}}$ need to be calculated in terms of the parameters. For a general rate matrix $\mathbf{Q}$, an explicit expression for $P_{n,\mathbf{x}}$ is not known. However, the probabilities $P_{n,\mathbf{x}}$ can be computed recursively (13, 14). Let $P_{n,\mathbf{x}_{j,k}}$ be the probability that a site has one more individual with nucleotide j and one less individual with nucleotide k than a site in configuration $\mathbf{x}$, and let $P_{n-1,\mathbf{x}_k}$ be the probability that a site in a sample of size $n-1$ has one less individual with nucleotide k. The probabilities $P_{n,\mathbf{x}}$ can be computed by solving the linear recursion

$$P_{n,\mathbf{x}} = \frac{\theta}{\theta + n - 1} \sum_{j,k} \frac{x_j + 1 - \delta_{jk}}{n} p_{jk} P_{n,\mathbf{x}_{j,k}} + \frac{n-1}{\theta + n - 1} \sum_k \frac{x_k - 1}{n-1} P_{n-1,\mathbf{x}_k}, \qquad [3]$$

where δ_{kj} is the Kronecker δ function, and the recursion starts with initial conditions $P_{1,i} = \pi_i$. It is convenient to allow negative entries in $\mathbf{x}$, with the convention that in this case, $P_{n,\mathbf{x}}$ is zero.

When analyzing data in which transversions are absent, purine and pyrimidine transitions can be considered separately, and each nucleotide site can be considered as binary. Two rate matrices, each of dimension 2, are used for purine and pyrimidine transitions. Each has the form

$$\mathbf{Q} = \begin{pmatrix} -q_{01} & q_{01} \\ q_{10} & -q_{10} \end{pmatrix}, \qquad [4]$$

where "0" and "1" represent the bases adenine and guanine in the case of purines and represent the bases cytosine and thymine in the case of pyrimidines. In this special case, the probabilities $P_{n,\mathbf{x}}$ are given by ref. 15:

$$P_{n,(i,n-i)} = \binom{q_{10} + i - 1}{i} \binom{q_{01} + n - i - 1}{n - i} \Big/ \binom{q_{01} + q_{10} + n - 1}{n}. \qquad [5]$$

Although Eq. 5 is sufficient to analyze data sets that lack transversions, it is not sufficient in general.

Lastly, for the special case of binary sites we developed a method based on d, the mean number of PDs in the sample. Although d estimates the single parameter θ in the infinite-sites model, in the finite-sites model d has expected value (13)

$$E[d] = \frac{s(q_{01} + q_{10})\,(1 - \pi_0^2 - \pi_1^2)}{(q_{01} + q_{10}) + 1} = \frac{s\bar{\theta}}{(q_{01} + q_{10}) + 1}, \qquad [6]$$

where $\bar{\theta} = \pi_0 q_{01} + \pi_1 q_{10}$ is the overall or average rate and s is the length of the sequence. When rates are small, the factor $(q_{01} + q_{10}) + 1$ is nearly unity, and d/s should estimate $\bar{\theta}$. To estimate q_{01} and q_{10} from d, let f_0 and f_1 be the fraction of zeros and ones in the n sequences collectively, with expected values π_0 and π_1. Equating f_0 and d to their expected values and solving gives

$$q_{01} = \frac{df_0}{s(1 - f_0^2 - f_1^2) - d}, \qquad q_{10} = \frac{df_1}{s(1 - f_0^2 - f_1^2) - d}. \qquad [7]$$

Eq. 7 gives PD estimators for the parameters, which are analogous to the estimators given in refs. 16 and 17. In the

simulations below, we compare the PD estimates for $\bar{\theta}$, q_{01}, and q_{10} with those obtained from the IS and LS methods.

Results and Discussion

Although statistical properties of each estimation procedure are difficult to obtain analytically, samples from the coalescent model are readily simulated (18). Our simulation method recursively generates a sample of m sequences from a sample of size $m - 1$, as follows. The process starts with two identical sequences of length s, generated by selecting a nucleotide at each site at random according to the stationary vector π. In the recursive step, one of the m sequences is chosen at random and either (*i*) with probability $(m - 1)/(s\theta + m - 1)$, that sequence is duplicated to get $m + 1$ sequences or (*ii*) with probability $s\theta/(s\theta + m - 1)$, it experiences a mutation at a randomly chosen site. The process continues until $n + 1$ sequences are obtained, at which time the last sequence is discarded. The resulting n sequences are a random sample from the coalescent (13, 19).

By generating a series of simulated samples of binary sites, we empirically calculated means, variances, and percentiles for all three estimators over a range of parameter values chosen to reflect the Amerindian data used as an example. Accordingly, we chose to simulate 63 sequences of 201 sites, selecting 15 values of substitution rates ranging from 0.005 to 0.1 to bracket the estimated values by approximately one order of magnitude. Substitution matrices were constructed by combining the 15 values into pairs, omitting pairs in which the two rates differed by more than an order of magnitude. For each matrix, at least 500 samples were generated, with up to 3000 samples simulated for selected matrices.

Performance of Estimators

Estimators for the binary model were evaluated by computing the observed bias, SD, and mean square error (MSE) from the simulated sets of data. Table 1 gives a representative set of the results when $q_{01} = q_{10}$ (first six rows) as well as when q_{01} is held constant and q_{10} increases (second six rows).

Overall, the performance of the LS and IS estimation methods was essentially identical, except that when substitution rates exceeded 0.01, the LS method tended to exhibit a 5–10% lower SD than the IS method. For all three methods, the largest observed relative error was 0.004/0.06 = 6.8%, which occurred when $q_{01} = 0.06$. Overall, the relative error averaged 2.1%, 2.2%, and 3.4% for the IS, LS, and PD methods, respectively. This level of bias is negligible when compared with the stochastic variation or the uncertainty in effective population sizes.

The middle section of Table 1 shows that the SDs for the IS and LS methods are generally ≈40% of the value being estimated and ≈70% for the PD method. We noted that for a specific value of q_{01}, the SD of the estimate decreased when the q_{10} increased. For example, the estimate for 0.005 had SD 0.0027 when paired with 0.008, and the SD decreased to 0.0023 when paired with the value 0.05.

The square root of MSE gives the average error in estimation. Overall, for the IS and LS methods the average error in estimation was 62% of the error for the PD method. As indicated by Table 2, the estimates $\bar{\theta}$ show similar properties. Table 2 also indicates that, as expected, the mean PD underestimates the average rate in the finite-sites model. In addition, the PD method cannot yield rate estimates for a single site, whereas the LS and IS methods could be applied to sites individually. Overall, these comparisons indicate that the LS and IS methods are more efficient and versatile than methods based on the mean PD.

Application to the Nuu-Chah-Nulth Data

These estimation techniques are suitable for mitochondrial DNA because it is effectively haploid and lacks recombination. Ward *et al.* (12) sequenced the first 360 base pairs of the mitochondrial control region for a sample of 63 Nuu-Chah-Nulth (Nootka). The sample comprised individuals who were maternally unrelated for four generations, chosen from 13 of the 14 tribal bands and, thus, deviates from a truly random sample. For mitochondrial DNA, the effective population size is approximated by the number of reproducing females. Because there were ≈600 females of childbearing age in the population, this value was used as an estimate of the long-term effective population size (12).

Among the 63 sequences, there were 26 variable sites that defined 28 distinct alleles, hereafter called lineages. There were no apparent tranversions in the data (12). Of the 360 sites sequenced, 201 were pyrimidines and 159 were purines. The LS and IS methods were applied to the pyrimidine and purine sites separately to estimate the transition rates for each class of nucleotides.

Twenty-one of the 201 pyrimidine sites are variable. For this data set, the squared-error function defined by Eq. 1 has a unique global minimum when $q_{01} = 0.02$ (cytosine → thymine rate) and $q_{10} = 0.03$ (thymine → cytosine rate). The IS procedure gave identical results. Assuming these to be the true values, simulations with $n = 63$ and $s = 201$ give the SD of these estimates as 0.007 and 0.01. The expected number of substitutions per unit time is $q_{01}\pi_0 + q_{10}\pi_1$, which is estimated as 0.024. Scaling these quantities by twice the effective population size gives the probabilities of a cytosine →

Table 1. Empirical properties of estimation methods based on simulations of 201 sites and 63 individuals

	Bias of q_{01} estimate*			SD of q_{01} estimate*			MSE of q_{01} estimate†		
q_{01}, q_{10}*	LS	IS	PD	LS	IS	PD	LS	IS	PD
5, 5	−0.28	−0.27	−0.22	2.51	2.54	3.69	6.40	6.53	13.68
8, 8	0.30	0.30	0.35	3.58	3.64	5.48	12.88	13.32	30.15
10, 10	0.10	0.11	0.24	4.41	4.46	6.55	19.47	19.88	43.01
20, 20	0.27	0.31	0.27	7.85	8.14	12.03	61.62	66.40	144.86
50, 50	0.97	0.95	0.23	15.09	15.95	26.71	228.50	255.18	713.62
60, 60	1.42	1.79	2.86	18.41	19.86	33.77	341.07	397.46	1148.40
5, 8	0.00	0.00	−0.00	2.65	2.69	3.99	7.03	7.24	15.89
5, 20	0.05	0.06	0.02	2.55	2.59	3.69	6.51	6.69	13.62
5, 50	0.06	0.05	−0.01	2.28	2.29	3.16	5.22	5.25	9.97
60, 8	4.11	3.92	4.04	28.94	29.42	43.85	854.55	881.06	1938.93
60, 20	0.10	0.24	0.86	21.66	22.56	34.62	469.34	509.13	1199.01
60, 50	1.73	1.83	1.82	17.77	18.56	29.52	318.79	348.00	874.96

*Parameters, bias, and SD × 10^3.
†MSE × 10^6.

Table 2. Properties of $\bar{\theta}$ estimators based on simulations of 201 sites and 63 individuals

q_{01}, q_{10}*	$\bar{\theta}$*	Bias of $\bar{\theta}$*		SD of $\bar{\theta}$*		MSE of $\bar{\theta}$†	
		LS	PD	LS	PD	LS	PD
5, 5	5	−0.30	−0.32	2.47	3.55	6.20	12.7
5, 20	8	0.06	−0.28	4.01	5.43	16.1	29.5
5, 50	9	0.08	−0.68	4.09	4.97	16.8	25.1
20, 50	29	0.55	−2.6	9.60	12.7	92.4	167
50, 50	50	0.76	−5.5	14.7	20.5	216	452

*Parameters, bias, and SD × 10^3.
†MSE × 10^6.

thymine transition in a single generation as 17 × 10^{-6} per site and a thymine → cytosine transition as 25 × 10^{-6} (Table 3). The overall probability of a pyrimidine transition is 20 × 10^{-6}.

We calculated a 90% confidence region for the pair (q_{01}, q_{10}) using the simulated empirical distributions, as follows. To determine whether a point (r_{01}, r_{10}) is in the confidence region, 500 samples were generated using the values (r_{01}, r_{10}), and the resulting empirical distribution was used to construct a 90% probability region in the plane. If this region contains the original estimate (0.02, 0.03), then (r_{01}, r_{10}) is contained in the 90% confidence region for (q_{01}, q_{10}). The one-dimensional projections of this confidence region give confidence intervals of (0.01, 0.04) and (0.02, 0.06) for q_{01} and q_{10}, respectively. Scaling by twice the effective population size gives the lower and upper bounds for the mutation probabilities (Table 3).

The parameter estimates for purines are less precise than those for pyrimidines because there were only 5 variable sites among the 159 purine sites. As before, the LS and IS procedures give identical results. The estimate for q_{01} is 0.005 (adenine → guanine rate), and the estimate for q_{10} is 0.014 (guanine → adenine rate), with 90% confidence intervals of (0.002, 0.015) and (0.004, 0.05), respectively. The expected number of purine substitutions per unit time is 0.007. Scaling these estimates by twice the effective population size gives the probabilities of an adenine → guanine transition in a single generation as 4 × 10^{-6} per site and the probability of a guanine → adenine transition as 12 × 10^{-6} (Table 3). The overall probability of a purine substitution is 6.0 × 10^{-6}. These results suggest that in this region of the molecule, pyrimidine transitions occur about three times faster than purine transitions.

In the absence of transversions, the transition/transversion ratio cannot be estimated directly. However, parsimony analysis of the molecular phylogeny for these sequences indicates a minimum of 41 transitions are necessary to explain these data. With 41 transitions and no tranversions, the hypothesis that the transition/transversion ratio is $< k$ is rejected with 75% confidence when $k < 30$. A reasonable lower bound for the transition/transversion ratio of this region is, therefore, 30:1, which is consistent with results obtained by direct inspection of similar population data (11).

Table 3. Estimates of substitution rates and probabilities in the mitochondrial control region based on the distribution of variable sites in 63 Amerindians

Type of transition	q_{ij} estimate	P of transition*	Lower bound*	Upper bound*
C → T	0.02	17	8	33
T → C	0.03	25	17	50
A → G	0.005	2	2	13
G → A	0.014	12	3	42
Pyrimidine		20	11	40
Purine		6	2	18
Total		14	7	30

*Probabilities, upper bounds, and lower bounds × 10^6.

By using the fact that this region has 45% purines and 55% pyrimidines, the overall probability of a transition per generation per site in this region was computed to be 14 × 10^{-6}, considerably higher than published estimates for the entire mitochondrial molecule (20). In evolutionary terms, the estimates in Table 3 represent 46% divergence per million yr for purines, 16% divergence per 100,000 yr for pyrimidines, and 11% total divergence per 100,000 yr. These rates also imply that ≈14 children per million differ from their mother at a particular base pair.

Although there are no benchmarks to validate these estimates against, rates can be obtained by comparing human and chimpanzee data. Assuming a transition/transversion ratio of 30:1, Ward *et al.* (12) estimated the divergence of human and chimpanzee sequences for this portion of the control region to be 33% per million yr, or 4.1 × 10^{-6} for the probability of a substitution per generation per site, compared with the statistical lower bound of 7 × 10^{-6} for the total sequence rate (Table 3).

Because the validity of these estimates depends on the appropriateness of the model for the data set being analyzed, we investigated the fit of the model to the data. Simulations using the finite-sites model in conjunction with the coalescent mirrored some, but not all, aspects of the Amerindian data. Nucleotide frequencies in the simulated data sets agreed with actual frequencies, and the predicted number of variable sites (4.97 purine and 21.21 pyrimidine) agreed with the observed data (5 purine and 21 pyrimidine). Also, the total number of mutations in the simulated data set (35–56 per sample) bracketed the minimum of number of mutations in the observed data (41) as inferred by parsimony. However, the distribution of PDs in the observed data had a deficiency of identical or closely related sequences, compared with the model expectations, and a detailed analysis of the pyrimidine sites indicated a greater number of observed lineages (24), compared with the simulated data (9 to 17 lineages per sample).

Three different factors might contribute to these differences between the simulated and observed data and, thus, influence our estimates of mutation rates. (*i*) Site-specific variability in mutation rates, which could lead to an increased number of lineages, was investigated in two ways. First, inference using a more general model that postulates the existence of fast and slow sites estimated that 1–20% of the sites had mutation rates 7–100 times faster than the remaining 80–99% of the sites (21), and simulations using this model exhibited a better fit to the data in terms of the number of lineages and the distribution of PDs. Second, by removing the seven most variable sites, the model became more consistent with the observed data. The net result of incorporating site-specific heterogeneity of mutation rates is to raise the average estimate of mutation rate for the entire region by at least 50% due to a small number of fast sites, the rates of which may differ by more than an order of magnitude. (*ii*) Admixture between genetically distinct tribes or a polyphyletic origin of the initial founding population, which might account for the excess number of lineages in the observed data, was also investigated. Because the Nuu-Chah-Nulth have four clades that may predate the colonization of the Americas (12), we repeated the pyrimidine analysis on the largest clade and on the two largest clades to examine the possible effect of multiple founders. We also analyzed an extended data set including 81 individuals from the linguistically distinct Haida and Bella Coola tribes (22) to examine the possible influence of admixture. After accounting for the changes in effective population size, the resulting estimates ranged from 35% lower to 25% higher, indicating that even appreciable amounts of admixture have only a minimal

influence on the estimates. (*iii*) Lastly, our presumption that 600 represents the long-term effective population size may be unrealistic because of demographic fluctuations that have undoubtedly occurred in the past. However, consideration of the archeological record (23) and early historical accounts (24) suggests that the effective population size for this group was unlikely to have been <300 or >900. Adjusting for these extreme values would lower our estimates by 33% or raise them by 100%.

Overall, these comparisons suggest that, although biological populations differ in a number of ways from the model assumptions, each factor considered alone exerts only a relatively minor effect, compared with the difference between our estimates and previous estimates. However, because a more profound effect may occur when these factors act in concert, the values in Table 3 should be regarded as provisional estimates only.

Conclusions

Assuming a finite-sites model of mutation, we have presented three methods to estimate the rate at which nucleotide substitutions occur in a DNA sequence. Two of these methods, one based on a LS approach and one based on a likelihood approach assuming independent sites, gave comparable results, with acceptable MSE. Because confidence intervals for these two estimators are readily obtained by simulation, either provides a reasonable approach to obtain estimates of the average mutation rate for a given segment of DNA, assuming neutral substitution. Because the underlying model provides estimates of the product of the effective population size and the substitution rate, these methods could also be used to estimate effective population size in the event the substitution rates are known. The LS and IS methods have the additional advantage that they can be extended to DNA sequences in which transversions are common, to protein sequences, as well as to other kinds of data, such as repetitive elements, where mutation from one neutral allele to another is common (13).

Because the method based on the mean PD had substantially higher MSEs, it is less reliable. In addition, because PDs are computed from the entire sequence for two individuals at a time, the method never considers the phylogenetic information in the entire sample simultaneously. By contrast, the statistics $\{V_{n,x}\}$ consider all individuals at the same time, albeit at a single site, and can, therefore, take advantage of the phylogenetic information available for individual sites. In fact, the probabilities $P_{n,x}$ can be calculated by listing all possible phylogenies and summing the conditional probability of a single site given the phylogeny times the prior probability of that phylogeny, as defined by the coalescent (13). However, our method can only use phylogenetic information of sites individually, one at a time, rather than the phylogeny of all sites collectively. Were the true phylogeny known, more powerful methods may exist that could take advantage of this additional information. However, if the phylogeny of the data set must be inferred, the reliability of the rate estimates will depend on the reliability of the estimated phylogeny. Because phylogeny estimation for large numbers of taxa is a computationally intensive problem, it is a strength of these methods that there is no need to explicitly estimate the sample phylogeny. Hence, despite the loss of information that would be available by considering all sites jointly, our methods provide a practicable alternative to the theoretically more precise, but largely unattainable, strategy of incorporating a true phylogeny into the estimation procedure.

Lastly, inspection of the Amerindian mitochondrial data (12) analyzed by these methods indicates the importance of using the computationally more complex finite-sites model, rather than the simpler infinite-sites model. When these binary sites are considered in pairs, the occurrence of pairs where all four possible combinations of nucleotides occur indicates that at least one site in the pair must have experienced multiple substitutions (25). In fact, parsimony analysis shows at least 41 transitions are distributed among 26 variable sites. Hence, the infinite-site model cannot be used to explain the evolution of these Amerindian mitochondrial sequences. Our simulations of the 201 pyrimidine sites using the estimated rates $q_{01} = 0.02$ and $q_{10} = 0.03$ confirm that, even with these seemingly low rates, a large number of sites experienced multiple mutations. On average, 25% of the mutations occurred at sites that had already experienced a mutation, and the total number of mutations exceeded the number of observed variable positions by an average of seven. As indicated above, it is highly probable that some sites in this region of the mitochondrial molecule have a higher rate of mutation than others, which would increase the probability of multiple substitutions even further. Hence, future extensions of these methods will need to take account of site-specific rate heterogeneity, as well as the influence of changing population size.

We thank Bill Navidi, Bob Griffiths, Joe Felsenstein, and Geoff Watterson for their suggestions. Two reviewers suggested helpful changes, and the Mathematics Undergraduate Computing Lab provided countless hours of computer time for the simulations. This research was supported, in part, by National Institutes of Health Grant GM41746, by National Science Foundation Grants DMS88-03284 and DMS90-05833, 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.

1. Jukes, T. & Cantor, C. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. (Academic, New York), pp. 21–123.
2. Kaplan, N. & Risko, K. (1982) *Theor. Popul. Biol.* **21**, 318–328.
3. Kimura, M. (1980) *J. Mol. Evol.* **2**, 87–90.
4. Neel, J. & Thompson, E. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1904–1908.
5. Strobeck, C. (1983) *Theor. Popul. Biol.* **24**, 160–172.
6. Kingman, J. F. C. (1982) *J. Appl. Probab.* **19**, 27–43.
7. Griffiths, R. C. (1989) *J. Math. Biol.* **27**, 667–680.
8. Griffiths, R. C. (1980) *Theor. Popul. Biol.* **17**, 51–70.
9. Golding, G. & Strobeck, C. (1981) *Theor. Popul. Biol.* **22**, 96–107.
10. Vigilant, L., Pennington, R., Harpending, H., Kocher, T. & Wilson, A. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9350–9354.
11. Horai, S. & Hayasaka, K. (1990) *Am. J. Hum. Genet.* **46**, 828–842.
12. Ward, R. H., Frazier, B. L., Dew, K. & Pääbo, S. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8720–8724.
13. Lundstrom, R. (1990) Ph.D. thesis (Univ. of Utah, Salt Lake City).
14. Sawyer, S., Dykhuizen, D. & Hartl, D. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 6225–6228.
15. Wright, S. (1968) *Evolution and the Genetics of Populations* (Univ. of Chicago Press, Chicago), Vol. 2.
16. Ball, M., Neigel, J. & Avise, J. C. (1990) *Evolution* **44**, 360–370.
17. Avise, J. C., Ball, M. & Arnold, J. (1988) *Mol. Biol. Evol.* **5**, 331–344.
18. Hudson, R. R. (1983) *Evolution* **37**, 203–217.
19. Ethier, S. N. & Griffiths, R. C. (1987) *Ann. Probab.* **15**, 515–545.
20. Brown, W. M., George, M., Jr. & Wilson, A. C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1967–1971.
21. Lundstrom, R., Ward, R. H. & Tavaré, S. (1992) *Math. Biosci.*, in press.
22. Ward, R. H., Redd, A., Valencia, D., Frazier, B. & Pääbo, S. (1992) *Proc. Natl. Acad. Sci. USA*, in press.
23. Dewhirst, J. (1978) *Sound Heritage* (Prov. Arch. Brit. Colum., Victoria, BC), Vol. 7, pp. 1–30.
24. Duff, W. (1964) *The Indian History of British Columbia* (Prov. Mus. of British Columbia, Victoria), Vol. 1, pp. 1–117.
25. Felsenstein, J. (1988) *Annu. Rev. Genet.* **22**, 521–565.