Substitution Rate Variation Among Sites in Mitochondrial Hypervariable Region I of Humans and Chimpanzees

Laurent Excoffier and Ziheng Yang†*

*Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland; and †Department of Biology, Galton Laboratory, University College London, England

Mitochondrial D-loop hypervariable region I (HVI) sequences are widely used in human molecular evolutionary studies, and therefore accurate assessment of rate heterogeneity among sites is essential. We used the maximum-likelihood method to estimate the gamma shape parameter α for variable substitution rates among sites for HVI from humans and chimpanzees to provide estimates for future studies. The complete data of 839 humans and 224 chimpanzees, as well as many subsets of these data, were analyzed to examine the effect of sequence sampling. The effects of the genealogical tree and the nucleotide substitution model were also examined. The transition/transversion rate ratio (κ) is estimated to be about 25, although much larger and biased estimates were also obtained from small data sets at low divergences. Estimates of α were 0.28–0.39 for human data sets of different sizes and 0.20–0.39 for data sets including different chimpanzee subspecies. The combined data set of both species gave estimates of 0.42–0.45. While all those estimates suggest highly variable substitution rates among sites, smaller samples tend to give smaller estimates of α. Possible causes for this pattern were examined, such as biases in the estimation procedure and shifts in the rate distribution along certain lineages. Computer simulations suggest that the estimation procedure is quite reliable for large trees but can be biased for small samples at low divergences. Thus, an α of 0.4 appears suitable for both humans and chimpanzees. Estimates of α can be affected by the nucleotide sites included in the data, the overall tree length (the amount of sequence divergence), the number of rate classes used for the estimation, and to a lesser extent, the included sequences. The genealogical tree, the substitution model, and demographic processes such as population expansion do not have much effect.

Introduction

It is now well known that mutation (substitution) rates in the human mitochondrial D-loop are highly heterogeneous. Initially, two segments of the control region were noted to be particularly variable and were thus called hypervariable regions I and II (HVI and HVII), respectively (Vigilant 1986; Vigilant et al. 1991). Later, mutation rates were found to be highly variable even within each of the two fast-evolving segments, with a few mutational hot spots evolving rapidly while most sites have very low rates of change or remain unchanged (Hasegawa et al. 1993; Tamura and Nei 1993; Wakeley 1993). Rate variation among sites is usually modeled by a gamma distribution (Uzzell and Corbin 1971). The shape parameter α is the inverse of the coefficient of variation of mutation rates, with $\alpha \gg 1$ meaning relatively little rate variation and $\alpha < 1$ meaning extreme rate variation. Previous estimates of the α parameter for the human D-loop DNA range from 0.11 to 0.47 (Tamura and Nei 1993; Wakeley 1993; Yang and Kumar 1996; Meyer, Weiss, and von Haeseler 1999), indicating that substitution rates are highly variable among sites in the D-loop (see also Kocher and Wilson 1991).

Variable mutation rates can have major impacts on various aspects of molecular evolutionary studies. For instance, ignoring variable rates among sites may lead to underestimation of sequence distances (Hasegawa et al. 1993; Tamura and Nei 1993) and mislead phylogeny reconstruction (Yang 1996*a*). It also leads to biased estimates of the time to the most recent common ancestor from a population sample (Lundstrom, Tavaré, and Ward 1992*a*; Hasegawa et al. 1993; Tamura and Nei 1993; Yang 1996*b*). Tests of selective neutrality can be misleading when the assumed infinite-sites model is violated and rates vary considerably among sites (Bertorelle and Slatkin 1995; Aris-Brosou and Excoffier 1996; Tajima 1996). Furthermore, accurate reconstruction of population demographic history based on pairwise mismatch analysis requires the among-site rate variation to be taken into account (Lundstrom, Tavaré, and Ward 1992*b*; Rogers et al. 1996; Weiss and von Haeseler 1998; Schneider and Excoffier 1999). It thus appears important not only to estimate the average mutation rate of the D-loop, but also to characterize the substitution rate pattern among sites in the DNA sequence.

In this paper, we characterize substitution rates in HVI of humans and chimpanzees. Using the maximum-likelihood (ML) method (Yang 1994), we obtain estimates of the transition-transversion rate ratio ($\kappa = \alpha/\beta$ in the notation of Kimura [1980] and Hasegawa, Kishino, and Yano [1985]) and the gamma shape parameter (α) for mutation rate heterogeneity among sites. Our initial objective was to provide reliable parameter estimates for future studies, which may not include enough data for independent estimation. However, we found that the estimates were somewhat variable among different samples of sequences, and we thus explored possible reasons for the variation. Monte Carlo simulations were performed to study possible biases in the estimation procedure. The effects of the assumed genealogical tree, demographic processes such as population expansion, inclusion of sites in the data, and possible shifts in the mutation rate distribution among sites in certain lineages were explored.

Key words: D-loop, mitochondrial DNA, humans, chimpanzees, mutation rate variation, gamma distribution.

Address for correspondence and reprints: Ziheng Yang, Department of Biology, 4 Stephenson Way, London NW1 2HE, England. E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 16(10):1357–1368. 1999

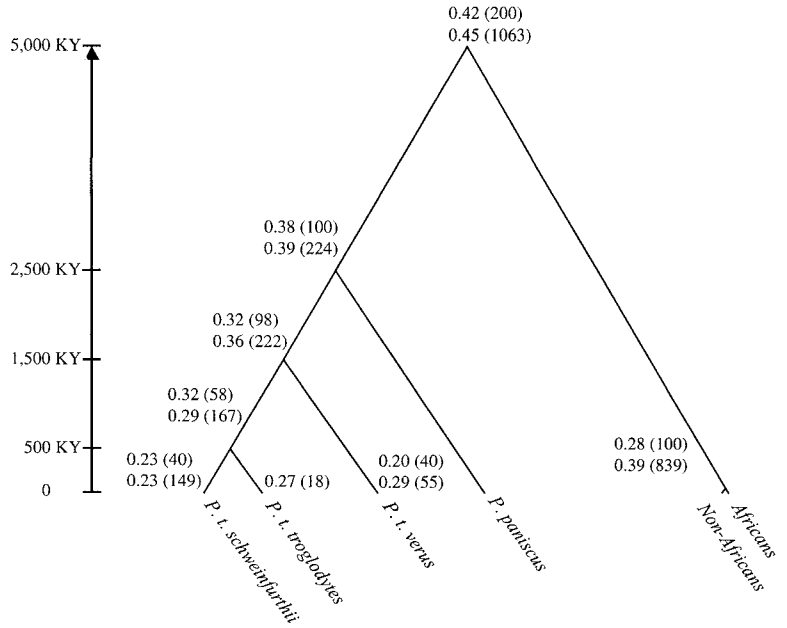

FIG. 1.—Phylogenetic tree of the human and chimpanzee populations and estimates of the gamma parameter α for HVI from different samples of sequences. The number of sequences involved in α estimation is given in parentheses for each sample. The time of divergence is in thousands of years (KY) before present.

Materials and Methods

Compilation of the Human and Chimpanzee Sequences

At the time of this study, over 4,000 mitochondrial D-loop HVI sequences from humans (Handt, Meyer, and von Haeseler 1998) and 224 sequences from chimpanzees had been published in the literature (Morin et al. 1994; Horai et al. 1995; Goldberg and Ruvolo 1997; Wise et al. 1997). For ML estimation of substitution rate parameters, we first compiled a combined data set of 100 human sequences and 100 chimpanzee sequences. Sequences were selected to be representative of overall human diversity. Out of the 100 human sequences, 30 were randomly drawn from sub-Saharan Africans, including Eastern, Western, and Southern Africans, and 70 were randomly drawn from non-Africans, covering all other continental regions. Only two sequences are available from *Pan paniscus* (*P.p.*), and both were used. The other chimpanzee sequences were 18 sequences from *Pan troglodytes troglodytes* (*P.t.t.*), 40 from *Pan troglodytes verus* (*P.t.v.*), and 40 from *Pan troglodytes schweinfurthii* (*P.t.s.*) (Morin et al. 1994; Horai et al. 1995; Goldberg and Ruvolo 1997; Wise et al. 1997). The data set was separated into one set of 100 humans and another set of 100 chimpanzees, the latter of which was further separated into data sets containing different chimpanzee subspecies. In our attempt to understand the effects of sequence sampling and possible differences in the rate distribution among populations, larger data sets were analyzed as well, including a data set of 200 humans and the complete data set that contains 839 human sequences and 224 chimpanzee sequences. The 839 human sequences represent all distinct sequences available in our database that do not contain many missing nucleotides.

The average nucleotide frequencies in the complete data set of 1,063 sequences are 22.63% (T), 32.82% (C), 33.93% (A), and 10.61% (G). Out of the 347 alignment positions, 173 sites involve alignment gaps or undetermined nucleotides; these are all treated as ambiguity nucleotides by ML.

The assumed phylogenetic relationships among those subspecies/populations are given in figure 1. The divergence times between subspecies/populations correspond approximately to those given in table 4 of Morin et al. (1994). The HVI region examined in this paper covers the segment between nucleotides 16,024 and 16,362 in the Cambridge numbering system (Anderson et al. 1981). Sequences were aligned manually. The alignment had 347 sites and included eight alignment gaps compared with the Cambridge sequence. The sequence data are available from the authors on request.

Simulations

Using a coalescent approach (Hudson 1990), we simulated genealogical trees for 200 DNA sequences, each of 300 nucleotide sites, distributed in six populations which have diverged according to the population tree of figure 1. All population sizes were fixed at 10,000 breeding females. The sample size within each population was the same as in the original combined data set of 100 humans and 100 chimpanzees. The phylogenetic tree for the populations (subspecies) (fig. 1) was fixed, while genealogical trees within populations (tree topology and coalescent times) were generated by simulating a stochastic coalescent process (Hudson 1990).

A random sequence was generated for the root of the tree, with the four nucleotides having equal proba-

bilities. Mutations (substitutions) were then added along branches of the genealogical tree according to the K80 substitution model (Kimura 1980). The transition/transversion ratio κ was fixed at 20. Three models of mutation rates among sites were used. Two variable-rates models assume a discrete gamma distribution with eight rate categories and with the shape parameter fixed either at $\alpha = 0.26$ (our initial estimate of the parameter for humans, which is also the estimate obtained by Meyer, Weiss, and von Haeseler 1999) or $\alpha = 0.40$ (our final estimate of the parameter). The third model assumes the same rate for all sites ($\alpha = \infty$). Four average mutation rates were used, that is, a slow, two intermediates, and a fast rate, fixed at 5×10^{-7}, 1.25×10^{-6}, 2.5×10^{-6}, and 1.25×10^{-5} substitutions per site per generation, respectively. Those mutation rates correspond to 5%, 12.5%, 25%, and 125% sequence divergence per Myr and cover the possible range of the mutation rates for HVI (Jazin et al. 1998; Parsons and Holland 1998). Possible generation time effects were examined by using an equal generation time for both humans and chimpanzees (10 years per generation) or unequal generation times (20 years per generation for humans and 10 years for chimpanzees).

Like the real data set analyzed in this paper, the 200 simulated sequences were also analyzed as three distinct data sets: one of four chimpanzee populations, one of two human populations, and a third of all six populations. In total, there are 48 different simulation conditions. For each simulation condition, 100 data sets were simulated using a program written by one of us (L.E.). Simulated DNA sequences as well as the true genealogical tree were collected into a file, to be subjected to ML analysis using PAUP* version 4.0.0d65 (David Swofford). Parameters α and κ were estimated under the K80+G model with eight rate categories (Yang 1994) using the true genealogical tree. PAUP* log files were processed to calculate the means and standard errors of parameter estimates over replicates. In a few cases, both the true genealogical tree and a neighbor-joining tree (Saitou and Nei 1987) were used. Estimates from the two trees were virtually identical, and only results from the true tree are presented below. The effect of the tree topology is further examined using real sequence data.

Real Data Analysis

We used the ML method to estimate the gamma shape parameter α for variable substitution rates among sites. The discrete gamma model is used with eight rate categories (Yang 1994). The method requires a genealogical tree linking the sequences. It would be best to use the true tree in the estimation. However, due to the low sequence divergence in the data, the probability of recovering the true genealogy of sequences is negligibly small (see, e.g., Hedges et al. 1992). We thus used the parsimony method to generate multiple candidate trees to obtain parameter estimates by the ML method and examine the robustness of the estimates to the tree topology used. Previous experience with phylogenetic analyses suggests that estimates of parameters do not change much among trees if the trees are not too wrong (Yang, Goldman, and Friday 1994; Sullivan, Holsinger, and Simon 1996). Our working hypothesis is that if the parameter estimates vary little over different parsimony trees, the estimates may be expected to be similar to those obtainable if the true tree were used. An alternative approach to accounting for the uncertainties in the genealogical tree is to average over all possible within-population genealogical trees using a coalescent model (Griffiths and Tavaré 1994; Kuhner, Yamato, and Felsenstein 1995). Such a likelihood approach is theoretically superior to the approach taken in this paper but has not been implemented yet. However, we may expect the two approaches to produce similar estimates of substitution rate parameters κ and α. Parsimony and likelihood are known to be highly correlated, especially at low sequence divergences (see, e.g., DeBry and Abele 1995). Thus, parsimony trees used in our estimation would make the greatest contributions to the coalescent likelihood calculation, which averages over all genealogical trees.

We used PAUP* to perform heuristic tree search using the parsimony criterion, with the starting tree generated by random addition of sequences followed by the tree-bisection-reconnection (TBR) perturbation. About five independent searches were carried out for each data set using different random number seeds. Candidate trees generated this way were used in ML estimation of parameters. Use of multiple parsimony trees provides an indication of the robustness of parameter estimates to inaccuracies in the recovered tree.

Three different Markov models of nucleotide substitution were used. The JC69+G model (Jukes and Cantor 1969) assumes equal substitution rates between any two nucleotides. The K80+G model (Kimura 1980) assumes different rates for transitions and transversions, while the HKY85+G model (Hasegawa, Kishino, and Yano 1985) accounts for both transition/transversion bias and nucleotide frequency biases. PAUP* is used to obtain ML estimates of parameters and branch lengths. The discrete gamma model of variable rates among sites is used with eight rate categories (Yang 1994).

After ML estimates of parameters were obtained, an empirical Bayes approach was used to estimate substitution rates at individual sites (Yang and Wang 1995). The substitution rate at a site is estimated by the conditional (posterior) mean of the rate distribution given the data at that site, and the resulting estimate has the highest correlation with the unknown true rate. The program BASEML in the PAML package (Yang 1998) is used for this calculation. Rates estimated from different data sets are compared to examine possible differences in the rate distribution among populations.

Results

Estimates of Parameters from Real Data

For each data set, parsimony was used to perform a heuristic tree search to generate candidate tree topologies, which were then used to estimate substitution parameters by ML. In the following, we use the combined

Table 1
Estimates of κ and α from Three Different Parsimony Trees for the Combined Data Set of 200 Humans and Chimpanzees

Tree and Maximum-Parsimony Score	ℓ	$\hat{\kappa}$	$\hat{\alpha}$	Tree length (S)
JC69+G				
Tree 1 (611 steps) . . .	−4,189.41		0.455	2.085
Tree 2 (613 steps) . . .	−4,194.36		0.453	2.107
Tree 3 (615 steps) . . .	−4,187.86		0.457	2.093
K80+G				
Tree 1	−3,739.27	20.811	0.417	2.330
Tree 2	−3,746.63	20.417	0.416	2.343
Tree 3	−3,741.50	20.388	0.419	2.336
HKY85+G				
Tree 1	−3,647.73	25.382	0.423	2.506
Tree 2	−3,659.74	24.914	0.422	2.502
Tree 3	−3,649.79	24.936	0.424	2.492

data set of 200 human and chimpanzee sequences as an example to illustrate the approach taken in this paper, while only summary results are presented for other data sets. Starting trees in the parsimony search were generated by stepwise addition with sequences added at random. The maximum number of trees retained was set at 3,000. Each search thus produced 3,000 locally best trees. Multiple local optima appear to exist in the tree space (Charleston 1995). In a preliminary analysis, many parsimony searches and many trees from each search were evaluated by ML. It was noted that trees generated in the same parsimony search produced virtually identical ML estimates of parameters and likelihood scores, indicating that those trees just represented different ways of resolving estimated polytomies. Trees generated from different parsimony searches gave more different parameter estimates and likelihood values. In later analyses, five parsimony searches were performed for each data set, and only one tree from each search was used in the likelihood estimation.

Table 1 presents parameter estimates obtained from three locally best trees in different parsimony searches for the combined human and chimpanzee data. The log-likelihood values are quite different among the three trees, while parameter estimates and the tree lengths (sum of branch lengths) are very similar. For example, under the HKY85+G model, the log-likelihood values range from −3,648 to −3,660, while estimates of α (0.42) and κ (25) are identical among the three trees at this level of accuracy. The tree lengths range from 2.49 to 2.51, with little difference. The results suggest that although the genealogical tree of the sequences cannot be estimated with any confidence, the uncertainty in the tree is unlikely to cause large errors in the parameter estimates. The three nucleotide substitution models provide quite different fits to data (table 1). The K80+G model accounts for the transition/transversion rate bias and involves one more parameter (κ) than the JC69+G model. The log-likelihood difference between the two models is about 450 (table 1), much greater than the critical value for the likelihood ratio test ($\frac{1}{2}\chi^2_{1\%} = 3.32$, df = 1). Thus, the K80+G model fits the data much better than the JC69+G model, and the transition rate is indeed much higher than the transversion rate. Similarly, the HKY85+G model is by far the best of the three substitution models used. Nevertheless, estimates of α obtained under the three substitution models are similar, despite drastic differences in their fits to data. It can be expected that minor violations of the HKY85+G model may not introduce large biases to the parameter estimates. For example, the TN93+G model (Tamura and Nei 1993) may provide a better fit than HKY85, but we do not expect much difference in estimates of α under the two models. In the following, we concentrate on parameter estimates obtained under the HKY85+G model, with the JC69+G and K80+G models being used to examine the robustness of the estimates.

Figure 2 shows the surface-contour plot of the log likelihood as a function of parameters κ and α under the HKY85+G model for the combined data set of 200 human and chimpanzee sequences. Tree 3 of table 1 is used in the calculation. The ML estimates are $\hat{\kappa} = 24.9$ and $\hat{\alpha} = 0.42$, with the log-likelihood value $\ell = -3{,}649.79$. The 95% likelihood (confidence) region for the two parameters (not shown, but see fig. 2) is surrounded by the contour at −3,652.79, that is, 3.00 ($=\frac{1}{2}\chi^2_{2,5\%}$) units worse than the optimum (see, e.g., Kalbfleisch 1985, pp. 113–114). Similarly, the 95% confidence interval for parameter α alone includes all α values at which the log-likelihood is at least −3,651.71 ($\frac{1}{2}\chi^2_{1,5\%} = 1.92$ units worse than the optimum); this interval is (0.34, 0.53).

The same approach was taken to analyze other data sets, and similar patterns were found concerning the stability of parameter estimates over parsimony trees and the fit of the three substitution models. Table 2 lists estimates obtained under the three substitution models from one tree topology only (usually the one with the highest likelihood among trees evaluated). The 200 sequences in the combined data set were separated into a data set of 100 humans and several data sets containing different chimpanzee subspecies/populations. To examine the effect of sequence sampling on the estimation, we also analyzed complete dat sets available for humans and for each chimpanzee subspecies. For example, the 98 *P.t.* sequences included in our combined human and chimpanzee data set and all 222 *P.t.* sequences available were analyzed as two data sets. Estimates range from 0.20 for the partial *P.t.v.* data set of 40 sequences to 0.45 for the complete data set of 839 humans and 224 chimpanzees (table 2 and fig. 1). For most data sets, the order of the α estimates under the three models is $\hat{\alpha}_{JC69} > \hat{\alpha}_{HKY85} > \hat{\alpha}_{K80}$. This order is different from that of previous analyses of between-species data, where simpler models (as well as parsimony) most often gave larger estimates of α, such that the order was $\hat{\alpha}_{JC69} > \hat{\alpha}_{K80} > \hat{\alpha}_{HKY85}$ (see, e.g., Yang, Goldman, and Friday 1994). Estimates of the transition/transversion rate ratio from large data sets are about 25. However, for small data sets with little sequence divergence (such as those of *P.t.v.*), estimates as large as 240 were obtained. Those large estimates appear to be overestimates, as the ML

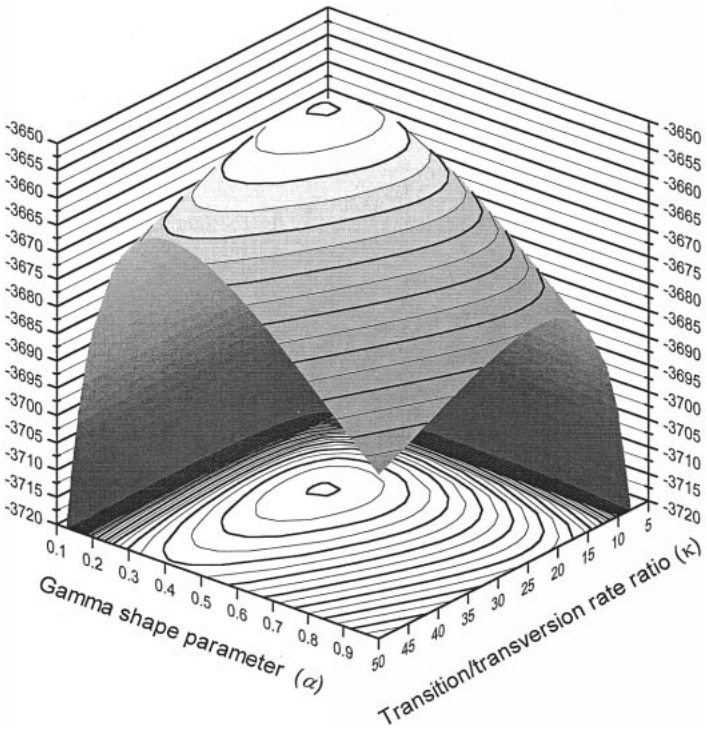

FIG. 2.—The log-likelihood surface-contour as a function of the transition/transversion rate ratio (κ) and the gamma parameter (α) for the combined data set of 200 human and chimpanzee sequences. For given values of κ and α, the log-likelihood is calculated by optimizing branch lengths in a parsimony tree (tree 3 in table 1) under the HKY85+G model. The maximum-likelihood estimates are $\hat{\kappa} = 24.9$ and $\hat{\alpha} = 0.42$, with the log-likelihood value $\ell = -3{,}649.79$. The 95% and 99% likelihood contours (not shown due to limitations of the graphics software) are at log-likelihood values of $-3{,}652.79$ and $-3{,}654.40$, respectively.

method tends to overestimate κ at low divergences (unpublished simulation results).

All estimates of α are much smaller than 1, consistent with earlier results indicating that substitution rates in HVI are highly variable (e.g., Kocher and Wilson 1991; Wakeley 1993). Rate constancy is rejected for all data sets by the likelihood ratio test comparing the constant-rate and gamma-rates models. For example, for the data set of 839 human sequences, the log-likelihood value under the gamma model is $\ell_1 = -9{,}309.42$, while the model with one rate for all sites ($\alpha = \infty$) gives $\ell_0 = -8{,}476.90$. Comparison of $2\Delta\ell = 2(\ell_1 - \ell_0) = 1{,}665.04$ with $\chi^2 = 6.63$ at 1% with df = 1 suggests that substitution rates are significantly different among sites. The same conclusion is reached with other data sets, including the *P.t.t.* data set with only 18 sequences (with $2\Delta\ell = 26.8$).

While the genealogical tree and the substitution model do not appear to affect the estimation of the α parameter much, the estimates are more sensitive to the choice of sequences in the data. The α values estimated under the HKY85+G model from different sequence samples are reported in figure 1 (see also table 2). The difference between the two estimates (0.20 for 40 sequences and 0.29 for 55 sequences) for the *P.t.t.* population appears to be largely due to the inclusion of sites with missing data in the larger data set. Estimates for other subspecies/populations appear more stable. The log-likelihood contours are shown in figure 3 for the data set of 839 human sequences and for that of 224 chimpanzee sequences. The 839 human sequences gave an estimate of 0.39, with a 95% confidence interval (CI) of (0.33, 0.46). The 224 chimpanzee sequences gave an estimate of 0.39, with a 95% CI of (0.31, 0.50). The estimate obtained from the complete data set of 1,063 human and chimpanzee sequences is 0.45, with a 95% CI of (0.39, 0.52).

It is noteworthy that larger estimates of α were consistently obtained for larger samples and for data covering earlier divergences in the tree (fig. 1). To explore this pattern further, we constructed 50 random subsamples of a fixed size from the 839 human sequences and examined how the average α estimate changes with the sample size. We considered sample sizes of 100, 200, 300, and 400 sequences. The average α estimate from 50 samples of sequences increased from 0.34 for 100 sequences to 0.39 for all 839 sequences, and the initial estimate (0.28) from our selected data set of 100 human sequences turned out to be unusually small. Two possible reasons may account for this effect of sequence sampling in humans and for the different estimates among chimpanzee and human populations. The first is possible differences in the rate distributions among sites in the different populations/subspecies, which may be due to shifts in the rate distributions along certain ancestral lineages caused by slight functional changes of the D-loop. If two populations with the same level of rate heterogeneity had a shift in rate distribution (implying that some fast sites in one population are slow sites in the other), combining data from the two popu-

Table 2
Estimates of κ and α from Different Data Sets

Data Set	*n*	MP Score	JC69+G		K80+G			HKY85+G		
			$\hat{\alpha}$	*S*	$\hat{\kappa}$	$\hat{\alpha}$	*S*	$\hat{\kappa}$	$\hat{\alpha}$	*S*
Homo and Pan	200	611	0.46	2.09	20.8	0.42	2.33	25.4	0.42	2.51
	1,063	1,848	0.41	8.77	20.9	0.40	9.10	24.7	0.45	7.70
Homo	100	263	0.27	0.90	23.3	0.25	1.00	25.7	0.28	0.88
	200	408	0.31	1.48	20.8	0.31	1.51	23.3	0.34	1.35
	839	1,338	0.34	6.46	21.3	0.34	6.65	24.0	0.39	5.21
Pan	100	334	0.43	1.10	19.7	0.38	1.22	25.2	0.38	1.33
	224	455	0.42	1.58	23.7	0.39	1.71	30.7	0.39	1.86
Pan troglodytes	98	257	0.34	0.83	39.3	0.31	0.90	49.6	0.32	0.94
	222	405	0.37	1.40	28.2	0.35	1.48	36.6	0.35	1.57
Pan troglodytes troglodytes and *Pan troglodytes schweinfurthii*	58	157	0.32	0.52	46.6	0.31	0.54	58.3	0.32	0.58
	167	290	0.30	1.07	26.0	0.29	1.11	36.0	0.28	1.20
Pan troglodytes verus	40	83	0.21	0.27	201.7	0.17	0.30	241.2	0.20	0.32
	55	98	0.30	0.30	109.4	0.26	0.33	133.5	0.29	0.35
P. troglodytes schweinfurthii	40	76	0.19	0.25	160.2	0.18	0.26	188.5	0.23	0.26
	149	208	0.23	0.80	26.8	0.23	0.85	34.5	0.23	0.85
P. troglodytes troglodytes	18	77	0.35	0.27	27.3	0.29	0.28	34.8	0.27	0.32

Note.—*n* is the number of sequences in the data set. *S* is the tree length calculated as the sum of branch lengths and measured by the average number of substitutions per site. The MP score is the best parsimony score found for the data set.

lations will lead to apparently less rate variation among sites and larger estimates of the α parameter. It is not clear whether and to what extent this explanation may account for the inferred rate differences between humans and chimpanzees. However, we cannot expect this interpretation to apply to samples of human sequences of different sizes. The second possible reason is an estimation bias, that is, a systematic underestimation of α for short tree lengths.

Simulation Results

Computer simulations were performed to examine possible sources of biases in ML estimates of the transition/transversion rate ratio (κ) and the gamma shape parameter (α) under different parameter combinations. The means and standard deviations of parameter estimates over simulated replicates are presented in table 3. The amount of sequence divergence, as measured by the tree length, appears to be very important for the accuracy of estimation. For example, the standard deviations of both α and κ are much smaller for the high mutation rate than for the low mutation rate (table 3). The chimpanzee data sets produced more accurate estimates than did the corresponding human data sets, as the former are more divergent than the latter.

Surprisingly enough, a positive bias is observed in estimates of α in the simulated human data using eight rate classes. For the intermediate mutation rate, the average α estimates are 0.288 and 0.427 when the true values are 0.26 and 0.40, respectively. For the low mutation rates (cases 28 and 32 in table 3), ML optimization by numerical iteration was sometimes problematic, and some replicates led to unreasonably large parameter estimates for α or κ (e.g., ∞). Nevertheless, it is clear that α becomes increasingly overestimated with shorter tree length. This estimation bias is in the opposite direction to the pattern observed for real data (table 2).

Apart from the bias at low divergences, the estimation procedure appears to be quite reliable. In particular, no large differences were noticed between estimates obtained from the human and chimpanzee sequences. Whether the sequences are analyzed jointly or separately, sensible estimates are obtainable as long as the data contain enough sequence variation. Neither is the estimation affected by the assumed generation time for humans (10 or 20 years). The results suggest that if the same mutation model (rate distribution among sites) applies to different parts of the tree, one should be able to recover the parameter of the rate distribution using data from different parts of the tree. We also examined possible effects of population growth on the estimation of the gamma parameter by simulating human sequence data assuming a sudden population expansion 10,000 or 50,000 years ago. The results (not shown) confirm our expectation that demographic processes affect the shape of the genealogical tree but not estimation of parameters concerning the mutation rate distribution.

Substitution Rates at Sites

Substitution rates at individual sites were estimated under the HKY85+G model using the posterior mean of the rates given the data (Yang and Wang 1995). ML estimates of branch lengths and model parameters (κ and α) were used in the calculation. Rates were estimated for different data sets to examine potential differences in the rate distributions among populations. The correlation coefficients between the estimated and the unknown true rates are about 0.70–0.73 for large data sets (with over 100 sequences) and are as low as 0.50 for small data sets.

Relative rates at fast sites identified from the complete data sets of 839 humans and 224 chimpanzees are presented in table 4. For comparison, fast sites identified by Wakeley (1993), Hasegawa et al. (1993), and Meyer, Weiss, and von Haeseler (1999) for the human HVI are also listed. Our results obtained from the 839 human sequences suggest the presence of 20 superfast sites with rates at least four times the average rate and 4 additional

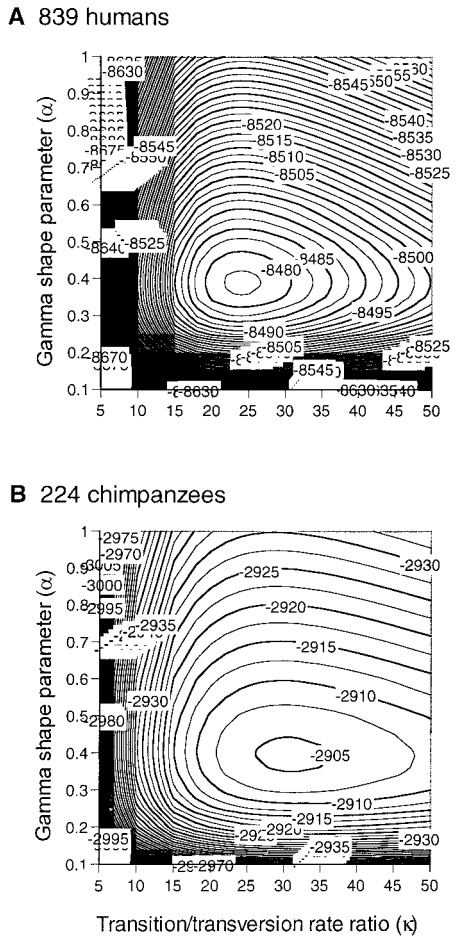

FIG. 3.—The log-likelihood contours as functions of parameters κ and α for the data set of 839 human sequences (*A*) and for the data set of 224 chimpanzee sequences (*B*). For each data set, a parsimony tree (not shown) is used to optimize the branch lengths for given values of κ and α. The optimum log-likelihood values are −8,476.90 and −2,904.51 for *A* and *B*, respectively. The 95% confidence (significance) regions are defined by contour lines (not shown) 3.00 ($=\frac{1}{2}\chi^2_{2,5\%}$) log-likelihood units worse than the optimum, that is, by contour lines at −8,479.90 and −2,907.51 for *A* and *B*, respectively.

fast sites with rates more than three times the average rate (table 4). Of the 20 superfast sites, 8 were identified as fast by all three previous studies (sites 16172, 16189, 16223, 16278, 16294, 16311, and 16302), 5 were identified as fast by at least two previous studies (sites 16093, 16129, 16209, 16291, and 16304), and 3 were identified as fast by Meyer, Weiss, and von Haeseler (1999) only (sites 16051, 16183, and 16309). Four superfast sites were not identified previously (sites 16182, 16192, 16265, and 16270). A few sites previously suggested to be moderately fast by at least two studies do not even show rates of twice the average in our study (sites 16126, 16148, 16187, 16274, 16292, 16298, and 16319).

The discrepancies between these lists of fast sites can be due to several factors. The first is stochastic errors, which appear to be substantial, since the correlation between estimated and true rates is not very high even with our data set of 839 sequences. Second, different methods have been used to identify fast sites. Wakeley (1993) and Hasegawa et al. (1993) counted minimum numbers of mutations at sites using parsimony, while Meyer, Weiss, and von Haeseler (1999) used another intuitive variability measure. The Bayes approach used in this paper (Yang and Wang 1995) is statistically optimal. Third, different sequences were used in these analyses, affecting both estimation of the α parameter (table 2) and identification of fast sites. Finally, different regions of HVI were considered. Wakeley considered only 250 nt and discarded the first 100 nt of HVI, while in our study and in the studies of Hasegawa et al. (1993) and Meyer, Weiss, and von Haeseler (1999), about 350 nt are considered for HVI. Among those factors, sampling of sequences and sites appears to be most influential.

Even greater differences in estimated substitution rates for sites exist between humans and chimpanzees (table 4). While the correlation between estimated rates from two human data sets, one of 100 sequences and another of 839 sequences (see fig. 1 and table 2), is as high as 0.84, the correlation of estimated rates between the human data set of 839 sequences and the chimpanzee data set of 224 sequences is only 0.52, and other chimpanzee data sets have rates even more different from those of the humans (correlation coefficients of 0.29, 0.36, and 0.45 between data of 839 humans and those of *P.t.s.*, *P.t.t.*, and *P.t.v.*, respectively). Among the 20 superfast sites identified in humans, only 12 are found to be at least twice as fast as the average for chimpanzees. Conversely, among the 16 superfast sites identified in the global chimpanzee data set, only 5 are identified as at least twice as fast as the average for humans. Relative substitution rates for sites estimated for the data of 839 human sequences and 224 chimpanzee sequences are plotted in figure 4. Estimates of the α parameter are identical in the two data sets (0.39), but there seem to be differences in the rate distribution among sites between the two species. Large differences in rate estimates between the two data sets are not restricted to particular regions of HVI and are scattered over the whole segment.

Discussion

Estimates of the Gamma Parameter for Humans and Chimpanzees

In a study of rate heterogeneity in HVI, Wakeley (1993) obtained an α estimate of 0.47 for HVI from 135 human sequences (Vigilant et al. 1991). This estimate is larger than our estimate of 0.39 from 839 human sequences obtained under the HKY85+G model. Although the parsimony method used by Wakeley (1993) may involve biases (Yang and Kumar 1996), his simulations suggested that the method is reliable when applied to a large number of sequences. We thus investigated the difference by analyzing the data of Vigilant et al. (1991) using the likelihood method. We note that Wakeley (1993) used only 250 bp from HVI (nucleotides 16130–16379), while 347 bp (nucleotides 16024–16362) are used in this study. We found that when the complete sequence of 347 bp from the data of Vigilant

Table 3
Estimates of κ and α in Computer Simulations

Case	No. of Populations	Mutation Rate[a]	Tree Length[b]	True α = 0.26				True α = 0.40			
				$\hat{\kappa}$	SE	$\hat{\alpha}$	SE	$\hat{\kappa}$	SE	$\hat{\alpha}$	SE
Chimpanzees and humans (200 sequences), 10 years per generation for chimpanzees and humans											
1 . .	6	Fast	22.841	20.152	0.192	∞	—				
2 . .	6	Intermediate 1	4.568	20.617	0.292	∞	—				
3 . .	6	Intermediate 2	2.736	20.987	0.411	∞	—				
4 . .	6	Slow	0.913	20.870	0.568	∞	—				
5 . .	6	Fast	22.836	20.497	0.200	0.259	0.001	20.642	0.214	0.398	0.002
6 . .	6	Intermediate 1	4.571	20.318	0.367	0.266	0.003	20.188	0.343	0.397	0.004
7 . .	6	Intermediate 2	2.731	20.906	0.420	0.268	0.003	20.889	0.408	0.398	0.005
8 . .	6	Slow	0.915	21.673	0.644	0.272	0.005	20.449	0.532	0.419	0.009
Chimpanzees and humans (200 sequences), 20 years per generation for humans											
9 . .	6	Fast	20.189	20.283	0.155	∞	—				
10 . .	6	Intermediate 1	4.038	20.635	0.326	∞	—				
11 . .	6	Intermediate 2	2.419	20.708	0.342	∞	—				
12 . .	6	Slow	0.808	21.324	0.636	∞	—				
13 . .	6	Fast	20.191	20.277	0.188	0.256	0.001	20.559	0.189	0.395	0.002
14 . .	6	Intermediate 1	4.039	21.069	0.391	0.260	0.002	20.218	0.325	0.400	0.004
15 . .	6	Intermediate 2	2.420	20.624	0.387	0.263	0.003	20.677	0.384	0.395	0.005
16 . .	6	Slow	0.807	20.544	0.606	0.274	0.005	20.913	0.534	0.432	0.011
Chimpanzees (100 sequences), 10 years per generation											
17 . .	4	Fast	11.707	20.167	0.230	∞	—				
18 . .	4	Intermediate 1	2.333	20.311	0.330	∞	—				
19 . .	4	Intermediate 2	1.393	20.587	0.380	∞	—				
20 . .	4	Slow	0.467	21.985	0.700	∞	—				
21 . .	4	Fast	11.639	20.513	0.244	0.256	0.002	20.647	0.283	0.394	0.003
22 . .	4	Intermediate 1	2.338	20.683	0.412	0.269	0.003	20.670	0.365	0.406	0.005
23 . .	4	Intermediate 2	1.394	21.259	0.473	0.270	0.005	20.510	0.478	0.416	0.008
24 . .	4	Slow	0.467	23.141	0.837	0.292	0.009	21.687	0.802	0.447	0.026
Humans (100 sequences), 20 years per generation											
25 . .	2	Fast	1.769	19.994	0.329	∞	—				
26 . .	2	Intermediate 1	0.355	21.822	0.793	∞	—				
27 . .	2	Intermediate 2	0.215	26.188	2.048	∞	—				
28 . .	2	Slow[c]	0.070	22.518	1.589	∞	—				
29 . .	2	Fast	1.784	21.772	0.450	0.268	0.002	20.370	0.309	0.402	0.004
30 . .	2	Intermediate 1	0.354	22.735	0.868	0.288	0.007	22.609	0.994	0.427	0.012
31 . .	2	Intermediate 2	0.219	25.672	1.662	0.341	0.023	24.757	1.583	0.841	0.239
32 . .	2	Slow[d]	0.071	23.461	1.795	0.442	0.062	21.023	1.596	0.705	0.129

[a] Rates are 5×10^{-7}, 1.25×10^{-6}, 2.5×10^{-6}, and 1.25×10^{-5} substitutions per site per generation for slow, intermediate 2, intermediate 1, and fast rates, respectively.

[b] Average number of substitutions per site along the tree.

[c] Only 87 estimates out of 100 led to plausible values for κ (i.e., $\kappa < 1{,}000$).

[d] Only 80 and 64 estimates out of 100 led to plausible values for simulated cases with $\alpha = 0.26$ and with $\alpha = 0.40$, respectively.

et al. (1991) is used, the likelihood estimates of α are 0.32, 0.32, and 0.34 under the JC69+G, K80+G, and HKY85+G models, respectively. These estimates are similar to those we obtained from different samples of human sequences of comparable sizes (table 2). When only the 250 nt used by Wakeley (1993) are used, however, the likelihood estimates of α are 0.44, 0.43, and 0.46 under the JC69+G, K80+G, and HKY85+G models, respectively. These estimates are similar to Wakeley's (0.47). Therefore, the differences between parameter estimates seem to be mainly due to the use of different sites (segments in the D-loop) in the data. The estimated tree length (the sum of branch lengths along the tree, measured as the average number of substitutions per site) is 0.77 when all 347 sites are used, and it is 0.86 for the partial sequence of 250 sites. The 100 sites at the 5′ end of HVI are less variable than the rest of the segment, and including them led to more variable rates among sites for the entire segment and, thus, smaller estimates of the α parameter.

More recently, Meyer, Weiss, and von Haeseler (1999) analyzed a large database of HVI and HVII sequences. These authors averaged estimates of the α parameter over 150 subsamples, each of 80 sequences, and obtained an estimate of 0.26 for the α parameter for HVI. This estimate is much smaller than our estimate (0.39) for the complete data set of 839 human sequences but is close to our estimates from samples of similar sizes. Meyer, Weiss, and von Haeseler (1999) included 20 sites at the 3′ end of the HVI (nucleotides 16362–16383) which appear to evolve slowly. To examine whether the different estimates are due to the different sites included in the data sets, we gathered a set of 848 sequences covering the same region (nucleotides 16024–16383) studied by Meyer, Weiss, and von Haeseler (1999). This data set leads to an α estimate of 0.39

Table 4
Relative Substitution Rates at Fast Sites from Different Studies

Site	*Homo* (839)	*Pan* (224)	W	H	M	Site	*Homo* (839)	*Pan* (224)	W	H	M
16051....	4				4	16242.....				5	
16074....		2				16243.....			5		
16092*...						16249.....				6	
16093*...	4			8	2	16256*....	2	2	6	5	2
16111....	2	4			2	16259.....		4			
16126....				6	4	16260.....	2	2			
16129....	4	3		15	4	16261.....					2
16136....			5			16265.....	4				
16143....		3				16266.....		2	6		
16145....			5	5		16269.....					
16148....				5	2	16270.....	4				
16156....		4				16273.....		3			
16157....		2				16274.....			6	5	2
16163....	2	2			3	16275.....		2			
16166....	2				2	16278.....	4	2	5	8	4
16167....		4				16284.....		4			
16168....		3				16289.....		3			
16172....	4	3	9	8	3	16290.....	2	3	6	7	
16175....		2				16291.....	4		8	6	
16179....		3				16292.....				6	2
16181....		4				16293.....	4	4	6	8	3
16182....	4					16294.....	4	3	11	9	4
16183....	4	4			2	16295.....		4			
16184....			5			16297.....					
16186....			5			16298.....			5	6	
16187....			7	7	3	16299.....		3			
16188....		3	5			16300.....		2			
16189....	4	3	13	15	4	16303.....					
16191....						16304.....	4	2	8	6	
16192....	4					16305.....					
16193....		3				16309.....	4				3
16194....						16311.....	4	4	17	14	4
16207....		4				16319.....		4	6	5	2
16209....	4	3	6	8		16320.....	3		5		2
16213....			5			16323.....		2			
16216....						16325.....				8	
16217....			5			16327.....		4			
16218....						16330.....					
16220....		2				16335.....		3			
16221....	2					16343.....	3	3			2
16223....	4	2	19	9	4	16352.....					
16230....					3	16355.....	3	3	7	6	2
16233....	2	3				16357.....		4			
16234....	3		7	7		16360.....		3			
16235....	2					16362.....	4	4	19	13	4
16241....		4									

Note.—The entries are estimated relative rates for the data set of 839 humans and that of 224 chimpanzes; for example, a 3 means that the estimated rate for that site is three times the average rate. Only sites at which at least one data set suggests a relative rate of at least 3 are listed, while for the listed sites, relative rates greater than 2 are also indicated. Fast sites reported by Wakeley (1993), Hasegawa et al. (1993), and Meyer, Weiss, and von Haeseler (1999) are shown in the columns headed ''W,'' ''H,'' and ''M,'' respectively. Entries in the M column are estimated relative rates, whereas those in the W and H columns are the minimum numbers of substitutions inferred by parsimony on the tree used; these are not comparable with each other or with our estimates of relative rates. Three sites (16092, 16093, and 16256) indicated by asterisks are found to be polymorphic in human pedigree studies (Howell, Kubacha, and Mackey 1996; Parsons et al. 1997).

under the HKY85+G model, identical to the estimate obtained from our 839 sequences with 20 fewer sites. Thus, the difference between our estimate and that of Meyer, Weiss, and von Haeseler is not due to the different sites included in the data sets. Instead, the small estimate of Meyer, Weiss, and von Haeseler appears to be due to those authors' use of a small number of sequences in their subsamples. It thus appears important to use a large number of sequences to obtain reliable estimates of the α parameter in within-species comparisons. We also note that drastically different estimates were obtained in previous studies from data sets covering different sites in the D-loop. For example, Tamura and Nei (1993) obtained an estimate of 0.11 for the entire control region (HVI and HVII and the middle conserved region), while Yang and Kumar (1996) obtained an estimate of 0.27 using only 25 selected divergent sequences from the same data but with the middle conserved region removed.

At any rate, a value of 0.4 for the gamma parameter appears appropriate for both humans and chimpanzees for the first 360 or 380 nucleotide sites in human HVI.

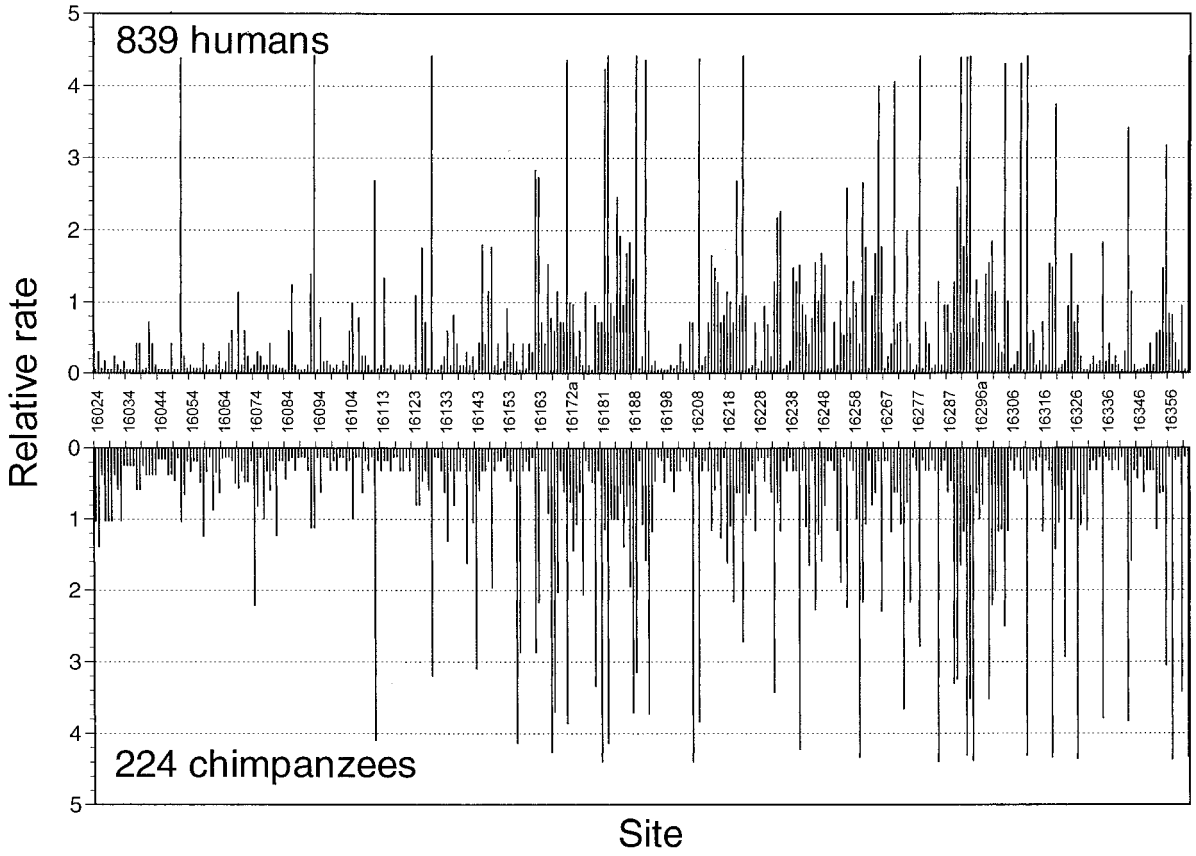

FIG. 4.—Substitution rates along the D-loop HVI sequence for humans (839 sequences) and chimpanzees (224 sequences). The HKY85+G model is applied to a parsimony tree to estimate branch lengths and model parameters from each data set, which are used to obtain empirical Bayes estimates of substitution rates for sites (Yang and Wang 1995). The rates are relative rates with a mean of 1.

The estimates do not seem to be affected much by the estimation procedure or the nucleotide substitution model if enough sequences are included in the analysis. Nevertheless, caution may be needed regarding the nucleotide sites included in the data when this estimate is applied. Among the factors considered in this paper, choice of nucleotide sites and sampling of sequences seem to have the greatest effects on the estimates. Including sequences containing missing nucleotides is found to affect the estimates too (results not shown), but we tried to avoid this problem when compiling the human sequences used in this study. The genealogical trees do not seem to have a major effect on the estimation as long as the data involve sufficient sequence divergence.

Possible Shifts in Substitution Rate Distribution Since the Divergence of Humans and Chimpanzees

The different estimates of the α parameter obtained when chimpanzees and humans are analyzed separately (0.39) and when they are analyzed together (0.45) may indicate a shift in the rate distribution since the divergence of humans and chimpanzees. The differences between the estimated rates for humans and chimpanzees shown in table 4 and figure 4 are consistent with this hypothesis. However, we note that our analyses are ad hoc, since the estimated rates have only moderately high correlation with the unknown true rates. The approach of examining estimates of α over data sets is also an intuitive one, as the assumption of a single rate distribution for all branches of the tree is unrealistic if rate shifts have occurred. Our analyses do not provide direct evidence for this hypothesis, although the results are compatible with it.

It is possible to construct a likelihood model that allows for different rate distributions for different parts of the genealogical tree. For example, two gamma distributions with different shape parameters may be used for branches of the tree before and after the presumed rate shift. The likelihood calculation would proceed similarly to the discrete gamma model of Yang (1994), but an average has to be taken over the two distributions to calculate the probability of observing data at a site. While a discrete-gamma model with k rate classes requires about k times the computation of the single-rate model (Yang 1994), the rate-shift model with two gamma distributions each of k rate classes would require k^2 times the computation of the single-rate model. The rate-shift model could then be compared with the model of one rate distribution for the entire tree to construct a likelihood ratio test of rate shift.

Discrete Versus Continuous Gamma Distribution of Mutation Rates

We simulated and analyzed DNA sequence data using a discrete gamma model with eight rate classes (ta-

ble 3). However, the real distribution of the mutation rates may differ substantially from our discrete model or may be more continuous. To explore the effect of an underlying continuous gamma distribution, we simulated data using 100 rate categories and analyzed them using 8 categories. A negative bias in the estimate of α is observed, apparently because a smaller α is needed for the discrete gamma to match the extent of variability in the continuous gamma. However, the negative bias is much more pronounced at low than at high divergences. For case 7 of table 3 (with a tree length of 2.42 substitutions per site), the average estimates of α (using eight rate categories) are 0.242 and 0.369 when the true values are 0.26 and 0.40, respectively. However, for case 31 of table 3 (with a tree length of 0.22), the average estimates are 0.213 and 0.342 for true values of 0.26 and 0.40, respectively. This negative bias in small samples is consistent with the observed patterns of table 2 (see also fig. 1). The mutation rate in the real data may be close to the small mutation rate used in the simulations. If the underlying distribution of mutation rates is continuous, the negative bias at low mutation rates could account for most, if not all, of the dependence of the α estimates on the sequence sample size.

Acknowledgments

We thank D. Swofford for permission to use PAUP* test version 4.0d65, and S. Meyer, G. Weiss, and A. von Haeseler for making their unpublished manuscript available. This study was supported by Swiss National Foundation grant no. 32-047053.96 to L.E. and a BBSRC grant to Z.Y.

Literature Cited

ANDERSON, S., A. T. BANKIER, B. G. BARRELL et al. (14 coauthors). 1981. Sequence and organization of the human mitochondrial genome. Nature **290**:457–465.

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.

CHARLESTON, M. A. 1995. Toward a characterization of landscapes of combinatorial optimization problems, with special attention to the phylogeny problem. J. Comput. Biol. **2**: 439–450.

DEBRY, R. W., and L. G. ABELE. 1995. The relationship between parsimony and maximum-likelihood analyses: tree scores and confidence estimates for three real data sets. Mol. Biol. Evol. **12**:291–297.

GOLDBERG, T. L., and M. RUVOLO. 1997. The geographic apportionment of mitochondrial genetic diversity in east African chimpanzees, *Pan troglodytes schweinfurthii*. Mol. Biol. Evol. **14**:976–984.

GRIFFITHS, R. C., and S. TAVARÉ. 1994. Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. Biol. Sci. **344**:403–410.

HANDT, O., S. MEYER, and A. VON HAESELER. 1998. Compilation of human mtDNA control region sequences. Nucleic Acids Res. **26**:126–129.

HASEGAWA, M., A. DIRIENZO, T. D. KOCHER, and A. C. WILSON. 1993. Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. **37**:347–354.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HEDGES, S. B., S. KUMAR, K. TAMURA, and M. STONEKING. 1992. Human origins and analysis of mitochondrial DNA sequences. Science **255**:737–739.

HORAI, S., K. HAYASAKA, R. KONDO, K. TSUGANE, and N. TAKAHATA. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. USA **92**:532–536.

HOWELL, N., I. KUBACHA, and D. A. MACKEY. 1996. How rapidly does the human mitochondrial genome evolve? Am. J. Hum. Genet. **59**:501–509.

HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. J. FUTUYMA and J. D. ANTONOVICS, eds. Oxford surveys in evolutionary biology. Oxford University Press, New York.

JAZIN, E., H. SOODYALL, P. JALONEN, E. LINDHOLM, M. STONEKING, and U. GYLLENSTEN. 1998. Mitochondrial mutation rate revisited: hot spots and polymorphism. Nat. Genet. **18**: 109–110.

JUKES, T., and C. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KALBFLEISCH, J. G. 1985. Probability and statistical inference, Vol. 2. Statistical inference. Springer-Verlag, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and protein-coding region. Pp. 391–413 *in* S. OSAWA and T. HONJO, eds. Evolution of life: fossils, molecules and culture. Springer-Verlag, Tokyo.

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.

LUNDSTROM, R., S. TAVARÉ, and R. H. WARD. 1992*a*. Estimating substitution rates from molecular data using the coalescent. Proc. Natl. Acad. Sci. USA **89**:5961–5965.

———. 1992*b*. Modeling the evolution of the human mitochondrial genome. Math. Biosci. **112**:319–335.

MEYER, S., G. WEISS, and A. VON HAESELER. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics **152**: 1103–1110.

MORIN, P. A., J. J. MOORE, R. CHAKRABORTY, L. JIN, J. GOODALL, and D. S. WOODRUFF. 1994. Kin selection, social structure, gene flow, and the evolution of chimpanzees. Science **265**:1193–1201.

PARSONS, T. J., and M. M. HOLLAND. 1998. Reply to Jazin et al. Nat. Genet. **18**:110.

PARSONS, T. J., D. S. MUNIEC, K. SULLIVAN et al. (11 coauthors). 1997. A high observed substitution rate in the human mitochondrial DNA control region. Nat. Genet. **15**: 363–368.

ROGERS, A. R., A. E. FRALEY, M. J. BAMSHAD, W. S. WATKINS, and L. B. JORDE. 1996. Mitochondrial mismatch analysis is insensitive to the mutational process. Mol. Biol. Evol. **13**:895–902.

SAITOU, N., and M. NEI. 1987. The neighbor joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SCHNEIDER, S., and L. EXCOFFIER. 1999. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152**: 1079–1089

SULLIVAN, J., K. E. HOLSINGER, and C. SIMON. 1996. The effect of topology on estimates of among-site rate variation. J. Mol. Evol. **42**:308–312.

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.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distribution to evolutionary events. Science **172**: 1089–1096.

VIGILANT, L. A. 1986. Control region sequences from African populations and the evolution of human mitochondrial DNA. Ph.D. thesis, University of California, Berkeley.

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

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

WEISS, G., and A. VON HAESELER. 1998. Inference of population history using a likelihood approach. Genetics **149**: 1539–1546.

WISE, C. A., M. SRAML, D. C. RUBINSZTEIN, and S. EASTEAL. 1997. Comparative nuclear and mitochondrial genome diversity in humans and chimpanzees. Mol. Biol. Evol. **14**: 707–716.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–341.

———. 1996*a*. Among-site variation and its impact on phylogenetic analyses. TREE **11**:367–372.

———. 1996*b*. Statistical properties of a DNA sample under the finite-sites model. Genetics **144**:1941–1950.

———. 1998. PAML: phylogenetic analysis by maximum likelihood. Version 1.4. University College London.

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.

YANG, Z., N. GOLDMAN, and A. E. FRIDAY. 1994. Comparison of models for nucleotide substitution in maximum likelihood phylogenetic estimation. Mol. Evol. Biol. **11**:316–324.

YANG, Z., and T. WANG. 1995. Mixed model analysis of DNA sequence evolution. Biometrics **51**:552–561.

MASAMI HASEGAWA, reviewing editor

Accepted June 24, 1999