Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees[1]

Koichiro Tamura and Masatoshi Nei

Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University

Examining the pattern of nucleotide substitution for the control region of mitochondrial DNA (mtDNA) in humans and chimpanzees, we developed a new mathematical method for estimating the number of transitional and transversional substitutions per site, as well as the total number of nucleotide substitutions. In this method, excess transitions, unequal nucleotide frequencies, and variation of substitution rate among different sites are all taken into account. Application of this method to human and chimpanzee data suggested that the transition/transversion ratio for the entire control region was ~15 and nearly the same for the two species. The 95% confidence interval of the age of the common ancestral mtDNA was estimated to be 80,000–480,000 years in humans and 0.57–2.72 Myr in common chimpanzees.

Introduction

The control region of mitochondrial DNA (mtDNA) in humans and apes is known to evolve very rapidly, and for this reason a number of authors (e.g., Horai and Hayasaka 1990; Vigilant et al. 1991; Ward et al. 1991) have used this region to study various problems of human evolution. However, the pattern of nucleotide substitution in this region is quite complicated; the ratio of transitional to transversional nucleotide substitutions is very high (Brown et al. 1982; Aquadro and Greenberg 1983), and the rate of nucleotide substitution (λ) varies extensively among different sites (Kocher and Wilson 1991). Furthermore, the transitional changes between purines and between pyrimidines do not occur with equal frequency, as will be seen later. Therefore, it is necessary to develop an appropriate mathematical method to use this region of mtDNA for the study of human evolution.

The main purpose of this paper is to develop a mathematical method to estimate the number of nucleotide substitutions between sequences. However, before developing the mathematical method, we first investigate the pattern of nucleotide substitution in the control region of mtDNA in humans. We shall also apply the mathematical formula developed for estimating the number of nucleotide substitutions between humans and chimpanzees, as well as for the age of the common ancestral mtDNA in humans and chimpanzees.

1. Key words: mtDNA control region, nucleotide substitutions, transition/transversion bias, human evolution, chimpanzees.

Address for correspondence and reprints: Masatoshi Nei, Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, 328 Mueller Laboratory, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 10(3):512–526. 1993.

0737-4038/93/1003-0002$02.00

Pattern of Nucleotide Substitution

Vigilant et al. (1991) sequenced the control region of mtDNA for 135 humans sampled from around the world. Although the control region is highly variable at the sequence level, this variation is mainly confined to the hypervariable segments on the 5′ and 3′ sides of the control region (~630 nucleotide sites), and the central portion shows virtually no variation among human sequences. Vigilant et al. (1991) therefore did not sequence the central portion, except for 20 individuals. Furthermore, some sequences included unidentified nucleotides even in the hypervariable segments. We therefore used 95 sequences in which nucleotides are known for 625 sites.

To study the relative frequencies of substitutions between different nucleotides, we used Gojobori et al.'s (1982) method. In this method the nucleotide sequences of ancestral mtDNA are inferred by using the principle of maximum parsimony, and the directional changes of nucleotides are determined by comparing a sequence with its immediate ancestral sequence. When the nucleotide at a site of an ancestral sequence was ambiguous and two nucleotides were possible at the site, each of the nucleotides was considered as the ancestral nucleotide with a probability of ½. When three or more nucleotides were possible, the site was discarded from the analysis. In the present case, all nucleotide sequences were so closely related to each other that almost all nucleotide substitutions were probably detected by the parsimony method.

To apply the above method, it is necessary to have the phylogenetic tree for all the sequences used. This tree was inferred by using Saitou and Nei's (1987) neighbor-joining method, which is known generally to produce the minimum-evolution tree (Saitou and Imanishi 1989; Rzhetsky and Nei 1992). The root of the tree was determined by using the chimpanzee sequence C3 of Kocher and Wilson (1991) as an outgroup (fig. 1). According to the bootstrap test, the location of the root was not stable (see Hedges et al. 1992; Templeton 1992), and the root was sometimes located to the cluster of !Kungs, pygmies, or some other non-African groups. However, this change in root, as well as other changes in the tree topology, had little effect on the relative frequencies of nucleotide substitutions estimated.

The numbers of directional nucleotide changes observed were as follows: A→T = 2; A→C = 5; A→G = 64.5; T→A = 1; T→C = 112; T→G = 1; C→A = 5; C→T = 115; C→G = 3; G→A = 37.5; G→T = 2; and G→C = 2. To obtain the relative frequency of a class of nucleotide substitution, we must divide the above value by the frequency of the original nucleotide, for each substitution class. For example, the relative frequency of substitution A→T is obtained by dividing 2 by 0.321, which is the frequency of nucleotide A. In practice, however, it is more convenient to express the relative substitution frequencies (f_{ij}; change from nucleotide i to j), so as to make the total sum of f_{ij}'s equal to 100%. The f_{ij}'s in table 1 are obtained in this way. In this computation we used the average nucleotide frequencies of A, T, C, and G, over all sequences compared. The average frequencies of A, T, C, and G are 0.321, 0.233, 0.314, and 0.132, respectively.

Table 1 shows that the rate of transitional nucleotide substitution is much higher than that of transversional substitution. The transition/transversion ratio is usually defined as the ratio of the observed number of transitions to that of transversions. This was 329/21 = 15.7 in the present case. This ratio is close to the value (15.0) observed by Vigilant et al. (1991). It is interesting that the transitional rate between pyrimidines (T⇄C) is higher than that between purines (A⇄G). A similar purine-pyrimidine bias in the transitional substitutions has been observed in *Drosophila*

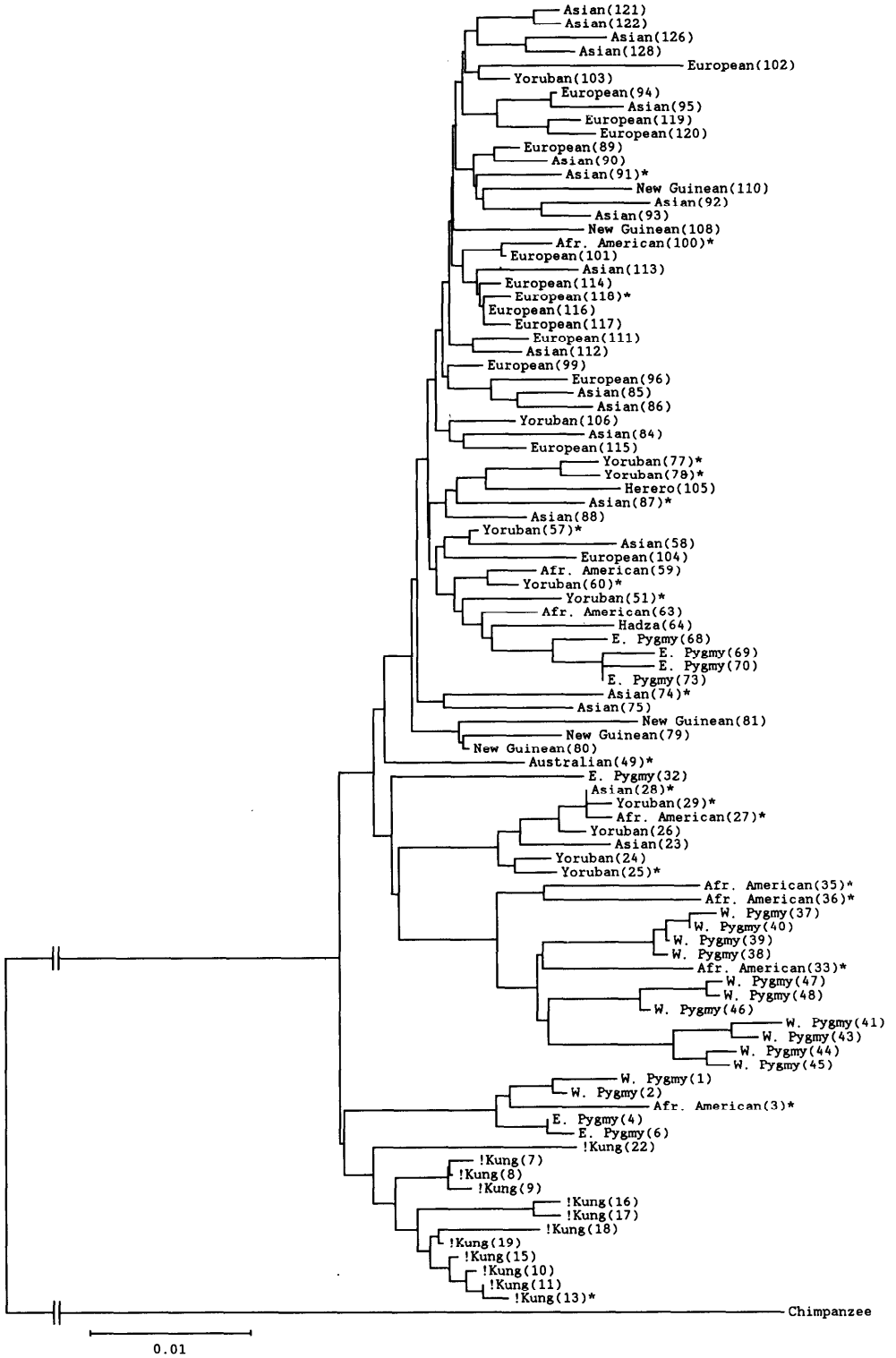

FIG. 1.—Phylogenetic tree for 95 human and one chimpanzee sequences of the mtDNA control region. This tree was constructed by the neighbor-joining method (Saitou and Nei 1987), from the distance matrix of the proportion of different nucleotides. The 20 sequences used for estimating the age of common ancestral mtDNA are indicated by asterisks. The numbers in parentheses refer to the sequence number in the original data from Vigilant et al. (1991).

Table 1
Esimated Relative Frequencies of Nucleotide Substitutions in the Hypervariable Segments of the mtDNA Control Region in Humans

Mutant Nucleotide	Original Nucleotide			
	A	T	C	G
A		0.3	1.1	20.0
T	0.4		25.8	1.1
C	1.1	33.8		1.6
G	14.1	0.3	0.5	

mtDNA (Tamura 1992*b*). Close examination of the substitution pattern also suggests that the substitution rate depends on the frequency of the mutant nucleotide. Thus, f_{GA} and f_{TC} are higher than f_{AG} and f_{CT}, respectively.

As mentioned earlier, the rate of nucleotide substitution λ in the control region is known to vary substantially from site to site. Kocher and Wilson (1991) showed that the distribution of the number of nucleotide substitutions per site approximately follows the negative binomial distribution rather than the Poisson distribution. Since we have a larger set of substitution data, we reexamined this problem. Table 2 shows the observed numbers of sites showing 0, 1, 2, and ≥ 3 substitutions, together with the expected numbers obtained for the Poisson and the negative binomial distributions. The observed data follow the negative binomial rather than the Poisson distribution.

The negative binomial distribution is known to be generated when Poisson parameter λ varies according to the following gamma distribution among sites (Johnson and Kotz 1973, pp. 124–125).

$$f(\lambda) = \frac{b^a}{\Gamma(a)} e^{-b\lambda} \lambda^{a-1}, \quad (1)$$

where $a = \bar{\lambda}^2/V(\lambda)$ and $b = a/\bar{\lambda}$, $\bar{\lambda}$ and $V(\lambda)$ being, respectively, the mean and variance of λ. $\Gamma(a)$ is the gamma function. Therefore, one can derive an equation for the average number of nucleotide substitutions for the entire control region by using the above gamma distribution, as will be shown later. However, to estimate this average number, we need an estimate of parameter a in equation (1). This parameter can be estimated by

$$a = m^2/(s^2 - m), \quad (2)$$

where m and s^2 are, respectively, the mean and variance of the number of nucleotide substitutions per site (Johnson and Kotz 1973, p. 131). In the present case, we have $a = 0.11$ from observed data in table 2. This estimate is virtually identical with that obtained by Kocher and Wilson (1991).

Mathematical Methods

On the basis of the pattern of nucleotide substitution observed in table 1, we propose that the mathematical model of nucleotide substitution presented in table 3

Table 2
Observed and Expected Distributions of the Number of Nucleotide Substitutions in the Control Region of Human mtDNA

No. of Substitutions per Site[a]	No. of Sites[b]		
	Observed	Poisson	Negative Binomial
0	1,028	987.0	1,028.0
1	58	121.2	59.3
2	21	7.4	17.4
≥3	9	0.4	11.3

[a] Estimated by parsimony analysis of 20 human sequences. The mean and variance of the estimated number of nucleotide substitutions per site are $m = 0.1228$ and $s^2 = 0.2582$, respectively.

[b] The χ^2 value for the difference between the observed and expected distributions is 97.8 ($P \ll 0.001$; 2 df) for the Poisson distribution and 1.24 ($0.5 < P < 0.7$; 2 df) for the negative binomial distribution. The number of degrees of freedom is 2 for the Poisson distribution because the substitution classes "2" and "≥3+" were pooled and one parameter was estimated, whereas the number for the negative binomial distribution is 2 because all the four classes were used and two parameters were estimated. The a value used for the latter distribution is 0.11.

be used for estimating the number of nucleotide substitutions in the control region of mtDNA. In this model, α_1, α_2, and β stand for the rates of transitional changes between purines and between pyrimidines and of transversional change, respectively, and g_A, g_T, g_C, and g_G represent the equilibrium frequencies of nucleotide A, T, C, and G, respectively. The latter parameters are incorporated into the substitution rates per unit evolutionary time, because the rates clearly depend on the frequencies of the mutant nucleotides (table 1). These parameters are also important to take care of the differences among the equilibrium values of g_A, g_T, g_C, and g_G (Tajima and Nei 1984). Note that when $\alpha_1 = \alpha_2$, the model presented in table 3 becomes identical with Hasegawa et al.'s (1985), though those authors did not derive any analytical formula for estimating the number of nucleotide substitutions.

We first consider a particular nucleotide site and derive an equation for the expected number of nucleotide substitutions between two sequences at that site. This equation obviously applies both to a set of nucleotide sites that have the same substitution pattern and the same substitution rate and to the case where the substitution pattern and the substitution rate are the same for all sites. When the nucleotide frequencies are in equilibrium, the average rate of nucleotide substitution per site for this model is given by $\lambda = 2g_Ag_G\alpha_1 + 2g_Tg_C\alpha_2 + 2g_Rg_Y\beta$, where $g_R = g_A + g_G$ and

Table 3
Rates of Nucleotide Substitution in the Model Used

Mutant Nucleotide	Original Nucleotide			
	A	T	C	G
A		$g_A\beta$	$g_A\beta$	$g_A\alpha_1$
T	$g_T\beta$		$g_T\alpha_2$	$g_T\beta$
C	$g_C\beta$	$g_C\alpha_2$		$g_C\beta$
G	$g_G\alpha_1$	$g_G\beta$	$g_G\beta$	

$g_Y = g_T + g_C$, whereas the expected number of nucleotide substitutions between two sequences that diverged t evolutionary time units ago ($d = 2\lambda t$) is

$$d = 4g_A g_G \alpha_1 t + 4g_T g_C \alpha_2 t + 4g_R g_Y \beta t . \tag{3}$$

To derive a formula for estimating d, one must know the expected proportions of nucleotide sites showing transitional differences between purines (P_1) and between pyrimidines (P_2) and of those showing transversional differences (Q), as expressed in terms of the substitution rates and evolutionary time. These proportions can be derived by the method described by Tamura (1992*a*) and become

$$P_1 = \frac{2g_A g_G}{g_R}\{g_R + g_Y \exp(-2\beta t) - \exp[-2(g_R\alpha_1 + g_Y\beta)t]\}, \tag{4}$$

$$P_2 = \frac{2g_T g_C}{g_Y}\{g_Y + g_R \exp(-2\beta t) - \exp[-2(g_Y\alpha_2 + g_R\beta)t]\}, \tag{5}$$

$$Q = 2g_R g_Y[1 - \exp(-2\beta t)] . \tag{6}$$

Since P_1, P_2, and Q are estimable from sequence comparison and since g_A, g_T, g_C, and g_G can be estimated by the average nucleotide frequencies of the two sequences compared, one can obtain the estimator of d as defined in equation (3). It becomes

$$\begin{aligned}\hat{d} = &-\frac{2g_A g_G}{g_R}\log_e\left(1 - \frac{g_R}{2g_A g_G}\hat{P}_1 - \frac{1}{2g_R}\hat{Q}\right)\\ &-\frac{2g_T g_C}{g_Y}\log_e\left(1 - \frac{g_Y}{2g_T g_C}\hat{P}_2 - \frac{1}{2g_Y}\hat{Q}\right)\\ &-2\left(g_R g_Y - \frac{g_A g_G g_Y}{g_R} - \frac{g_T g_C g_R}{g_Y}\right)\log_e\left(1 - \frac{1}{2g_R g_Y}\hat{Q}\right),\end{aligned} \tag{7}$$

where $\hat{P}_1$, $\hat{P}_2$, and $\hat{Q}$ are the estimates of P_1, P_2, and Q, respectively. Here, g_A, g_T, etc., represent the estimated nucleotide frequencies, but $\hat{}$ is eliminated for simplicity. It can be shown that equation (7) is a maximum-likelihood estimator of d under the model of nucleotide substitution given in table 3. The large-sample variance of $\hat{d}$ is given by

$$V(\hat{d}) = [(c_1^2\hat{P}_1 + c_2^2\hat{P}_2 + c_3^2\hat{Q}) - (c_1\hat{P}_1 + c_2\hat{P}_2 + c_3\hat{Q})^2]/n , \tag{8}$$

where n is the number of nucleotides examined, and

$$c_1 = \frac{\partial d}{\partial P_1} = \frac{2g_A g_G g_R}{2g_A g_G g_R - g_R^2\hat{P}_1 - g_A g_G\hat{Q}}, \tag{9}$$

$$c_2 = \frac{\partial d}{\partial P_2} = \frac{2g_T g_C g_Y}{2g_T g_C g_Y - g_Y^2\hat{P}_2 - g_T g_C\hat{Q}}, \tag{10}$$

$$c_3 = \frac{\partial d}{\partial Q} = \frac{2g_A^2 g_G^2}{g_R(2g_A g_G g_R - g_R^2 \hat{P}_1 - g_A g_G \hat{Q})} + \frac{2g_T^2 g_C^2}{g_Y(2g_T g_C g_Y - g_Y^2 \hat{P}_2 - g_T g_C \hat{Q})} + \frac{g_R^2(g_T^2 + g_C^2) + g_Y^2(g_A^2 + g_G^2)}{2g_R^2 g_Y^2 - g_R g_Y \hat{Q}}. \quad (11)$$

Our computer simulations have shown that equation (7) gives good estimates, unless $\hat{P}_1$, $\hat{P}_2$, and $\hat{Q}$ are very large and n is relatively small. In the latter case, equation (7) may become inapplicable because of negative arguments of logarithms. Therefore, the application of this equation should be confined to the case of relatively closely related sequences (say, $d < 1$) and a relatively large value of n.

Equation (7) was derived under the assumption that the rate of nucleotide substitution λ is the same for all sites considered. In the control region, however, λ is known to vary extensively (table 2) and approximately follows the gamma distribution, $f(\lambda)$. Therefore, following Nei and Gojobori (1986) and Jin and Nei (1990), we obtain the means of P_1, P_2, and Q as follows:

$$\bar{P}_1 = \frac{2g_A g_G}{g_R}\left\{g_R - \left[\frac{a}{a + 2(g_R\bar{\alpha}_1 + g_Y\bar{\beta})t}\right]^a + g_Y\left(\frac{a}{a + 2\bar{\beta}t}\right)^a\right\}, \quad (12)$$

$$\bar{P}_2 = \frac{2g_T g_C}{g_Y}\left\{g_Y - \left[\frac{a}{a + 2(g_Y\bar{\alpha}_2 + g_R\bar{\beta})t}\right]^a + g_R\left(\frac{a}{a + 2\bar{\beta}t}\right)^a\right\}, \quad (13)$$

$$\bar{Q} = 2g_R g_Y\left[1 - \left(\frac{a}{a + 2\bar{\beta}t}\right)^a\right], \quad (14)$$

where $\bar{\alpha}$ and $\bar{\beta}$ are the means of α and β, respectively. From these equations, we can derive the following formula for the average number of nucleotide substitutions per site between two sequences compared:

$$\hat{d} = 2a\left[\frac{g_A g_G}{g_R}\left(1 - \frac{g_R}{2g_A g_G}\hat{P}_1 - \frac{1}{2g_R}\hat{Q}\right)^{-1/a} + \frac{g_T g_C}{g_Y}\left(1 - \frac{g_Y}{2g_T g_C}\hat{P}_2 - \frac{1}{2g_Y}\hat{Q}\right)^{-1/a} + \left(g_R g_Y - \frac{g_A g_G g_Y}{g_R} - \frac{g_T g_C g_R}{g_Y}\right)\left(1 - \frac{1}{2g_R g_Y}\hat{Q}\right)^{-1/a} - g_A g_G - g_T g_C - g_R g_Y\right], \quad (15)$$

where $\hat{P}_1$, $\hat{P}_2$, and $\hat{Q}$ are the estimates of $\bar{P}_1$, $\bar{P}_2$, and $\bar{Q}$, respectively. The large-sample variance of $\hat{d}$ is approximately given by

$$V(\hat{d}) = [(c_1^2\hat{P}_1 + c_2^2\hat{P}_2 + c_3^2\hat{Q}) - (c_1\hat{P}_1 + c_2\hat{P}_2 + c_3\hat{Q})^2]/n\,, \tag{16}$$

where

$$c_1 = \frac{\partial d}{\partial \bar{P}_1} = \left[1 - \frac{g_R}{2g_Ag_G}\hat{P}_1 - \frac{1}{2g_R}\hat{Q}\right]^{-(1+1/a)}, \tag{17}$$

$$c_2 = \frac{\partial d}{\partial \bar{P}_2} = \left[1 - \frac{g_Y}{2g_Tg_C}\hat{P}_2 - \frac{1}{2g_Y}\hat{Q}\right]^{-(1+1/a)}, \tag{18}$$

$$\begin{aligned} c_3 = \frac{\partial d}{\partial \bar{Q}} = {} & \frac{g_Ag_G}{g_R^2}\left[1 - \frac{g_R}{2g_Ag_G}\hat{P}_1 - \frac{1}{2g_R}\hat{Q}\right]^{-(1+1/a)} \\ & + \frac{g_Tg_C}{g_Y^2}\left[1 - \frac{g_Y}{2g_Tg_C}\hat{P}_2 - \frac{1}{2g_Y}\hat{Q}\right]^{-(1+1/a)} \\ & + \left[\frac{g_A^2 + g_G^2}{2g_R^2} + \frac{g_T^2 + g_C^2}{2g_Y^2}\right]\left[1 - \frac{1}{2g_Rg_Y}\hat{Q}\right]^{-(1+1/a)}. \end{aligned} \tag{19}$$

With the present model, it is possible to estimate the average numbers of transitional (s) and transversional (v) substitutions separately. They are given by the following equations:

$$\begin{aligned} \hat{s} = 2a\Bigg\{ & \frac{g_Ag_G}{g_R}\left[1 - \frac{g_R}{2g_Ag_G}\hat{P}_1 - \frac{1}{2g_R}\hat{Q}\right]^{-1/a} \\ & + \frac{g_Tg_C}{g_Y}\left[1 - \frac{g_Y}{2g_Tg_C}\hat{P}_2 - \frac{1}{2g_Y}\hat{Q}\right]^{-1/a} \\ & - \left[\frac{g_Ag_Gg_Y}{g_R} + \frac{g_Tg_Cg_R}{g_Y}\right]\left[1 - \frac{1}{2g_Rg_Y}\hat{Q}\right]^{-1/a} - g_Ag_G - g_Tg_C\Bigg\}, \end{aligned} \tag{20}$$

$$\hat{v} = 2ag_Rg_Y\left\{\left[1 - \frac{1}{2g_Rg_Y}\hat{Q}\right]^{-1/a} - 1\right\}, \tag{21}$$

where $\hat{s}$ and $\hat{v}$ are the estimates of s and v, respectively. The variances for $\hat{s}$ and $\hat{v}$ are approximately given by

$$V(\hat{s}) = [(c_1^2\hat{P}_1 + c_2^2\hat{P}_2 + c_4^2\hat{Q}) - (c_1\hat{P}_1 + c_2\hat{P}_2 + c_4\hat{Q})^2]/n\,, \tag{22}$$

and

$$V(\hat{v}) = c_5^2\hat{Q}(1 - \hat{Q})/n\,, \tag{23}$$

respectively, where c_1 and c_2 are the same as those in equation (16) and

$$c_4 = \frac{\partial s}{\partial \bar{Q}} = \frac{g_A g_G}{g_R^2}\left[1 - \frac{g_R}{2g_A g_G}\hat{P}_1 - \frac{1}{2g_R}\hat{Q}\right]^{-(1+1/a)} + \frac{g_T g_C}{g_Y^2}\left[1 - \frac{g_Y}{2g_T g_C}\hat{P}_2 - \frac{1}{2g_Y}\hat{Q}\right]^{-(1+1/a)} + \left[\frac{g_A g_G}{g_R^2} + \frac{g_T g_C}{g_Y^2}\right]\left[1 - \frac{1}{2g_R g_Y}\hat{Q}\right]^{-(1+1/a)}, \quad (24)$$

and

$$c_5 = \frac{\partial v}{\partial \bar{Q}} = \left[1 - \frac{1}{2g_R g_Y}\hat{Q}\right]^{-(1+1/a)}. \quad (25)$$

We can also compute the variance of the $\hat{s}/\hat{v}$ ratio by the following formula:

$$V(\hat{s}/\hat{v}) = \frac{1}{\hat{v}^2}\left[(c_1^2\hat{P}_1 + c_2^2\hat{P}_2 + c_6^2\hat{Q}) - (c_1\hat{P}_1 + c_2\hat{P}_2 + c_6\hat{Q})^2\right]/n, \quad (26)$$

where

$$c_6 = c_4 - \frac{\hat{s}}{\hat{v}}c_5. \quad (27)$$

Estimation of the Number of Nucleotide Substitutions

To apply equation (15) to human and chimpanzee data, we must consider the entire control region, because the parameter a in the equation has been estimated for this region. There are 20 human sequences, 3 common chimpanzee sequences, and 1 pygmy chimpanzee sequence that are complete for the control region (Kocher and Wilson 1991; Vigilant et al. 1991). We therefore computed the average distance between each of the chimpanzee sequences and all human sequences, as well as the pairwise distances for all chimpanzee sequences, using equation (15). In this computation, $a = 0.11$ was used. The results obtained are presented in table 4. The $\hat{d}$ value for the human-chimpanzee comparison varies considerably with chimpanzee sequence, but the differences among different $\hat{d}$ values are not statistically different, and the average value becomes 0.752 ± 0.224. This value is substantially larger than the estimate (0.150) obtained by Kimura's (1980) two-parameter method, without taking into account unequal nucleotide frequencies and variation in λ among different nucleotide sites. This indicates the importance of using a proper mathematical model in estimating the number of nucleotide substitutions in the control region.

Table 4 also shows the $\hat{d}$ values obtained for various chimpanzee sequence comparisons. The $\hat{d}$ value between the common and pygmy chimpanzee sequences is 0.256–0.349, which is approximately one-third of the $\hat{d}$ value between human and chimpanzee sequences. In this case, too, Kimura's method gives substantial underestimates of nucleotide substitutions. The $\hat{d}$ values for the comparisons of common chimpanzee sequences are considerably lower than those for the comparison of common and pygmy chimpanzees, but the value for the comparison of C1 and C3 is as high as 0.153. For human sequences, the $\hat{d}$ value is given only for the deepest root of the sequences. That is, figure 1 shows that the human sequences can be divided into

Table 4
Estimates of the Number of Nucleotide Substitutions $\hat{d}$, Transition/Transversion Ratio $\hat{s}/\hat{v}$, and Their SEs

Groups Compared[a]	$\hat{d} \pm$ SE		$\hat{s}/\hat{v} \pm$ SE	
	Eq. (15)	K2[b]	Eqq. (20) and (21)	K2[b]
H-C1	0.586 ± 0.161	0.141 ± 0.013	9.4 ± 3.0	2.7 ± 0.5
H-C2	0.706 ± 0.213	0.144 ± 0.013	16.7 ± 5.8	3.8 ± 0.8
H-C3	0.840 ± 0.264	0.154 ± 0.013	16.9 ± 6.1	3.6 ± 0.7
H-P1	0.874 ± 0.246	0.162 ± 0.014	15.2 ± 4.9	3.3 ± 0.6
Average	0.752 ± 0.224	0.150 ± 0.013	14.6 ± 5.1	3.4 ± 0.7
C1-C2	0.118 ± 0.024	0.066 ± 0.008	4.7 ± 1.4	2.8 ± 0.8
C1-C3	0.153 ± 0.072	0.077 ± 0.009	5.2 ± 1.5	2.8 ± 0.7
C1-P1	0.317 ± 0.073	0.111 ± 0.011	7.7 ± 2.2	3.0 ± 0.7
C2-C3	0.072 ± 0.016	0.046 ± 0.007	14.3 ± 6.8	9.2 ± 4.4
C2-P1	0.265 ± 0.060	0.100 ± 0.016	10.4 ± 3.2	4.2 ± 1.0
C3-P1	0.349 ± 0.094	0.107 ± 0.011	14.1 ± 4.8	4.6 ± 1.1
Within humans[c]	0.024 ± 0.006	0.020 ± 0.004	15.7 ± 3.5[d]	

[a] C1–3 = common chimpanzees; P1 = pygmy chimpanzee; and H = humans. Data from Kocher and Wilson (1991) and Vigilant et al. (1991) were used.
[b] Estimates obtained by Kimura's (1980) two-parameter method.
[c] Divergence for the deepest root.
[d] Estimated by Nei's (1992) method.

two major groups, i.e., the !Kung cluster, including a few pygmies and an African-American, and the rest of the sequences. The first group includes 2 complete sequences (sequences 3 and 13), whereas the second group has 18 sequences. The $\hat{d}$ for the deepest human mtDNA root is obtained by computing the average distance between the sequences belonging to the two groups. It becomes 0.024. This value is substantially smaller than the $\hat{d}$ values between common chimpanzee sequences. In this case, Kimura's two-parameter method gives an estimate of $\hat{d}$ similar to that obtained by equation (15), because the number of nucleotide differences per site is very small.

Table 4 includes the ratio of transitional ($\hat{s}$) and transversional ($\hat{v}$) substitutions estimated by using equations (20) and (21) for various sequence comparisons. In the case of human and chimpanzee comparisons, the average ratio is 14.6, which is very close to the empirical $\hat{s}/\hat{v}$ ratio (15.7) obtained by parsimony analysis of human sequences. In the case of chimpanzee sequence comparisons, however, the ratio tends to be smaller than that for human sequences. In particular, all comparisons involving sequence C1 show a low $\hat{s}/\hat{v}$ ratio. Nevertheless, it is not clear whether this tendency is real, because the number of sequences used is small and the sampling errors of the $\hat{s}/\hat{v}$ ratio are very large. Kimura's method consistently gives underestimates of the ratio.

Rate of Nucleotide Substitution and the Age of the Common Ancestral mtDNA

Conducting a phylogenetic analysis of human mtDNA sequences, Vigilant et al. (1991) concluded that the age of the common ancestral mtDNA is 166,000–249,000 years. Nei (1992) reanalyzed their data and reached the conclusion that the age is

~110,000–504,000 years when the 95% confidence interval is considered. Let us now examine this problem, using our new mathematical method.

To estimate the age of the common ancestral mtDNA, it is necessary to know the average rate of nucleotide substitution for the control region. This rate can be obtained by assuming that humans and chimpanzees diverged 4–6 Mya (Vigilant et al. 1991). Since the mean $\hat{d}$ value ± standard error (SE) between humans and chimpanzees is 0.752 ± 0.224, the 95% confidence interval of the average number of substitutions becomes $\hat{d} \pm 2$ SE = 0.304–1.200. Therefore, the 95%-confidence upper and lower bounds of the average rate of substitution per site per year per lineage ($\bar{\lambda}$) are 2.5×10^{-8} and 1.5×10^{-7}, respectively. On the other hand, if we use a divergence time of 5 Myr between humans and chimpanzees, without considering the SE of $\hat{d}$, the rate becomes 7.5×10^{-8}. We call this the "modal rate" in the following.

In table 4, we have seen that the $\hat{d}$ for the deepest root of human mtDNA sequences is 0.024. Therefore, the 95%-confidence upper and lower bounds of the age of the common ancestral mtDNA ($\hat{d}/2\bar{\lambda}$) become 80,000 and 480,000 years, respectively, whereas the age when the modal rate is used is 160,000 years. (Here we have ignored the SE of $\hat{d}$ for the deepest root, because the error for the substitution rate λ has already been considered, and this error is not independent of the error of $\hat{d}$.) The 95% confidence interval of the age is similar to that obtained by Nei (1992) by a different method, but the range of the estimate is considerably wider than that given by Vigilant et al. (1991). This is because Vigilant et al. did not consider the SEs of their estimates.

The mtDNAs in common chimpanzees are known to be more variable than those in humans (Ferris et al. 1981; Kocher and Wilson 1991), but no one seems to have estimated the age of the common ancestral mtDNA. Although there are only three sequences for the mtDNA control region, it is interesting to estimate this age, because it is apparently much older than that of humans. For this purpose, we constructed a neighbor-joining tree for the three sequences, using the pygmy chimpanzee as an outgroup. The results obtained are presented in figure 2. From this figure, we can compute the $\hat{d}$ for the common ancestor of the three sequences by taking the average of $\hat{d}$'s for C1 versus C2 and C1 versus C3. It becomes 0.136. Therefore, if we use the modal rate, the age of the common ancestral mtDNA becomes 0.9 Myr. If we consider the 95% confidence interval, the age is estimated to be 0.57–2.72 Myr. These values are approximately six times higher than those for humans. These are minimum estimates, because only three sequences are used.

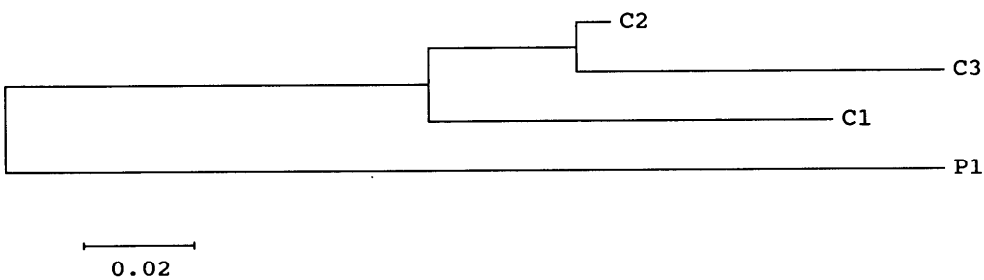

FIG. 2.—Phylogenetic tree for three common (C) and one pygmy (P) chimpanzee sequences of the mtDNA control region. This tree was constructed by the neighbor-joining method, from the distances computed by equation (15).

Discussion

It is well known that the rate of nucleotide (or amino acid) substitution generally varies extensively with nucleotide (or amino acid) site (Fitch and Margoliash 1967; Uzzell and Corbin 1971; Shoemaker and Fitch 1989). A simple method to estimate the total number of nucleotide substitutions in this case is to assume that there are two classes of sites and that the sites in the first class are invariable whereas those in the second class are subject to substitution with the same rate (Fitch and Margoliash 1967; Hasegawa et al. 1985). In practice, however, this model is less realistic than the one developed in the present paper, because variation in substitution rate is usually continuous (Uzzell and Corbin 1971; Kocher and Wilson 1991). This is clearly the case with the mtDNA control region (table 2). Actually, Wakeley (submitted) showed that, in the 250-bp 5′-side hypervariable segment of this region, even a model of two different rate classes does not fit the data, while the negative binomial distribution fits very well.

Hasegawa and Horai (1991) fitted the above two-class model to human and chimpanzee sequence data to estimate the age of the common ancestral mtDNA in humans. Their estimate of the 95% confidence interval of the age was 180,000–380,000 years. However, this estimate seems to be less reliable than ours, for three reasons. First, the two-class model that they used is less realistic than our model, as mentioned above. Second, they estimated the proportion f of variable sites from sequence data, but the reliability of their estimate of f is unclear. Actually, the estimate of the number of nucleotide substitutions obtained by using this model is known to be very sensitive to the value of f. Third, the number of human sequences used was much smaller than that used in the present study.

In recent years the age of the common ancestral mtDNA in humans has become a controversial issue in relation to the origin of *Homo sapiens*. Some authors have attempted to estimate the time of the origin of *H. sapiens* from the age of the common ancestral mtDNA. Theoretically, this attempt is doomed to failure because the age (or the coalescence time) is nothing but a function of long-term effective size of the population (Kingman 1982; Tajima 1983). If the effective size is large, the age of the common ancestor is expected to be large. However, the age has one important evolutionary implication. That is, the age of the common ancestor cannot be smaller than the time at which different populations diverged (Nei 1985). It is now generally agreed that the first major division of human populations occurred between Africans and non-Africans (Nei and Roychoudhury 1982; Cavalli-Sforza et al. 1988; Nei and Ota 1991). Therefore, our lower-bound estimate of the age of the common ancestral mtDNA suggests that the divergence between Africans and non-Africans occurred at least ~80,000 years ago [or 110,000 years ago, according to Nei's (1992) computation]. This estimate is close to the divergence time (115,000 years ago) estimated from data on nuclear genes (Nei and Roychoudhury 1982), though the latter estimate is very crude.

As mentioned earlier, our estimate of the s/v ratio (15.7) obtained from analysis of human sequences is very close to that obtained by Vigilant et al. (1991). It is also close to the estimate obtained by our model from the comparison of human and chimpanzee sequences. However, it is considerably smaller than Horai and Hayasaka's (1990) estimate (37.2) for the hypervariable segment on the 5′ side of the control region. Ward et al. (1991) also examined the s/v ratio for the 5′-side hypervariable segment, using an independent but small sample, and found no transversions. These

Table 5
Numbers of Transitional $\hat{s}$ and Transversional $\hat{v}$ Substitutions in the Hypervariable Segments of the mtDNA Control Region

Hypervariable Segment	$\hat{s}$	$\hat{v}$	Total	$\hat{s}/\hat{v}$
5′ Side	207	11	218	18.8
3′ Side	122	10	132	12.2
Overall	329	21	350	15.7

NOTE.—The number of nucleotide substitutions was estimated by parsimony analysis of 95 human sequences. The 5′- and 3′-side hypervariable segments are defined as positions 16028–16362 and 71–379, respectively, of Anderson et al.'s (1981) sequence. The χ^2 value for the 2×2 table is 0.95 ($0.3 < P < 0.5$).

observations suggest that the 5′-side hypervariable segment has a higher s/v ratio than does the 3′ side. Because we have more data for both 5′- and 3′-side segments, we examined this problem. The results obtained (table 5) show that the $\hat{s}/\hat{v}$ ratio certainly tends to be higher in the 5′-side segment than in the 3′ side, but the difference is not statistically different. Therefore, there is no need to treat the two segments separately.

However, the a value for the 5′ and 3′ hypervariable segment is not the same as that for the entire control region, because the former does not include a large segment of apparently invariable sites. For example, Wakeley (submitted) obtained $a = 0.47$ for the 250-bp 5′-side hypervariable segment. This value is higher than the value for the entire control region, because the invariable segment is not included.

The estimate of a is subject to sampling errors, and theoretically these errors are expected to affect the estimate of d. In practice, however, the effect of these errors is confounded with that of the sampling errors of $\hat{P}_1$, $\hat{P}_2$, and $\hat{Q}$, so that it is difficult to evaluate the effect. In the present case, our estimate of the age of the common ancestral mtDNA in humans was virtually the same as that obtained by Nei (1992) without using information on a. Furthermore, the $\hat{s}/\hat{v}$ ratio, which is a function of a, was nearly the same for the human-chimpanzee comparison and for the human sequence comparison. These observations suggest that our estimation of a is reliable.

Computer Program

A computer program for computing $\hat{d}$, $\hat{s}$, $\hat{v}$, and their variances is available on request.

Acknowledgments

We would like to thank Mark Stoneking for providing DNA sequence data and John Wakeley for sending an unpublished manuscript. This work was supported by research grants, from the National Institute of Health and the National Science Foundation, to M.N.

LITERATURE CITED

ANDERSON, S., A. T. BANKIER, B. G. BARRELL, M. H. L. DE BRUIJN, A. R. COULSON, J. DROUIN, I. C. EPERON, D. P. NIERLICH, B. A. ROE, F. SANGER, P. H. SCHREIER, A. J. H. SMITH, R. STADEN, and I. G. YOUNG. 1981. Sequence and organization of the human mitochondrial genome. Nature **290**:457–465.

AQUADRO, C. F., and B. D. GREENBERG. 1983. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics **103**:287–312.

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

CAVALLI-SFORZA, L. L., A. PIAZZA, P. MENOZZI, and J. MOUNTAIN. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc. Natl. Acad. Sci. USA **85**:6002–6006.

FERRIS, S. D., W. M. BROWN, W. S. DAVIDSON, and A. C. WILSON. 1981. Extensive polymorphism in the mitochondrial DNA of apes. Proc. Natl. Acad. Sci. USA **78**:6319–6323.

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

GOJOBORI, T., W.-H. LI, and D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:360–369.

HASEGAWA, M., and S. HORAI. 1991. Time of the deepest root for polymorphism in human mitochondrial DNA. J. Mol. Evol. **32**:37–42.

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., and K. HAYASAKA. 1990. Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am. J. Hum. Genet. **46**:828–842.

JIN, L., and M. NEI. 1990. Limitation of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

JOHNSON, N. L., and S. KOTZ. 1973. Distributions in statistics: discrete distributions. Houghton-Mifflin, Boston.

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

KINGMAN, J. F. C. 1982. On the genealogy of large populations. J. Appl. Probability **19A**:27–43.

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

NEI, M. 1985. Human evolution at the molecular level. Pp. 41–64 *in* T. OHTA and K. AOKI, eds. Population genetics and molecular evolution. Japan Scientific Societies, Tokyo.

———. 1992. Age of the common ancestor of human mitochondrial DNA. Mol. Biol. Evol. **9**:1176–1178.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

NEI, M., and T. OTA. 1991. Evolutionary relationship of human populations at the molecular level. Pp. 415–428 *in* S. OSAWA and T. HONJO, eds. Evolution of life: fossils, molecules, and culture. Springer, Tokyo.

NEI, M., and A. K. ROYCHOUDHURY. 1982. Genetic relationship and evolution of human races. Evol. Biol. **26**:421–443.

RZHETSKY, A., and M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. **9**:945–967.

SAITOU, N., and T. IMANISHI. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. **6**:514–525.

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

SHOEMAKER, J. S., and W. M. FITCH. 1989. Evidence from nuclear sequences that invariable

sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. **6**:270–289.

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

TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.

TAMURA, K. 1992*a*. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. **9**:678–687.

———. 1992*b*. The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol. Biol. Evol. **9**:814–825.

TEMPLETON, A. R. 1992. Human origins and analysis of mitochondrial DNA sequences. Science **255**:737.

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

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

WAKELEY, J. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. (submitted).

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

TAKASHI GOJOBORI, reviewing editor

Received June 2, 1992; revision received October 8, 1992

Accepted October 28, 1992