Estimation of the Number of Nucleotide Substitutions When There Are Strong Transition-Transversion and G+C-Content Biases[1]

Koichiro Tamura[2]

Department of Biology, Tokyo Metropolitan University

A simple mathematical method is developed to estimate the number of nucleotide substitutions per site between two DNA sequences, by extending Kimura's (1980) two-parameter method to the case where a G+C-content bias exists. This method will be useful when there are strong transition-transversion and G+C-content biases, as in the case of *Drosophila* mitochondrial DNA.

Introduction

In the study of molecular evolution, it is important to know the number of nucleotide substitutions per site (d) between DNA sequences. There are many different methods for estimating this number (for reviews, see Nei 1987, chap. 5; Gojobori et al. 1990), but all of them depend on some simplifying assumptions and do not always give reliable estimates. Two important factors that should be considered in the estimation of d are the inequality of the rates of transitional and transversional nucleotide substitution (transition-transversion bias) and the deviation of the G+C content (θ) from 0.5 (G+C-content bias).

Kimura (1980) developed a method of estimating d for the case where the transition-transversion bias exists. In his method, however, the frequencies of the four nucleotides A, T, C, and G are all assumed to be equal to 0.25, throughout the evolutionary time. In practice, this assumption usually does not hold. For example, the G+C content at the third-codon positions of the coding region of *Drosophila* mitochondrial DNA (mtDNA) is $\sim\leq 0.1$ (Clary and Wolstenholme 1985). It is therefore useful to develop a method that will take care of both the transition-transversion and G+C-content biases. In the following, I present one such method and compare the method's reliability with that of others.

Theory

I consider the model of nucleotide substitution shown in table 1, where α and β denote the rates of transitional and transversional changes, respectively. This model extends Kimura's (1980) two-parameter model to the case where $\theta \neq 0.5$. This model also is a special case of Hasegawa et al.'s (1985) model, where the frequency of A plus G is assumed to be equal to that of T plus C. This model assumes that the pattern of

[1] Key words: nucleotide substitutions, transition-transversion bias, G+C-content bias.

[2] Present address: Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, 328 Mueller Laboratory, University Park, Pennsylvania 16802.

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

Mol. Biol. Evol. 9(4):678–687. 1992.

0737-4038/92/0904-0009$02.00

Table 1
Rates of Nucleotide Substitution in Present Model

Original Nucleotide	Rates of Substitution for Mutant Nucleotide			
	A	T	C	G
A	$1-(\theta\alpha+\beta)$	$(1-\theta)\beta$	$\theta\beta$	$\theta\alpha$
T	$(1-\theta)\beta$	$1-(\theta\alpha+\beta)$	$\theta\alpha$	$\theta\beta$
C	$(1-\theta)\beta$	$(1-\theta)\alpha$	$1-[(1-\theta)\alpha+\beta]$	$\theta\beta$
G	$(1-\theta)\alpha$	$(1-\theta)\beta$	$\theta\beta$	$1-[(1-\theta)\alpha+\beta]$

nucleotide substitution is the same for all nucleotide sites, irrespective of the location of the nucleotide. This assumption does not seem to hold in many cases, but we can extract particular nucleotide sites at which this assumption approximately holds. For example, third-codon positions or fourfold-degenerate sites in protein-coding sequences can be regarded as such sites.

In the present case, the rate of nucleotide substitution per site is given by $\theta\alpha+\beta$ for nucleotides A and T and by $(1-\theta)\alpha+\beta$ for C and G. The average rate of nucleotide substitution per site (λ) is therefore given by

$$\lambda=(1-\theta)(\theta\alpha+\beta)+\theta[(1-\theta)\alpha+\beta]=2\theta(1-\theta)\alpha+\beta\,, \tag{1}$$

whereas the total d since the time of divergence between two sequences is given by

$$d=4\theta(1-\theta)\alpha t+2\beta t\,. \tag{2}$$

Here t is the divergence time (measured in years) between the two sequences, generations, or any other time units. The estimate of d is obtained by the following equation:

$$\begin{aligned}\hat{d}&=2\hat{\theta}(1-\hat{\theta})\left[-\log_e\left(1-\frac{1}{2\hat{\theta}(1-\hat{\theta})}\hat{P}-\hat{Q}\right)+\tfrac{1}{2}\log(1-2\hat{Q})\right]-\tfrac{1}{2}\log_e(1-2\hat{Q})\\&=-2\hat{\theta}(1-\hat{\theta})\log_e\left(1-\frac{1}{2\hat{\theta}(1-\hat{\theta})}\hat{P}-\hat{Q}\right)-\frac{1-2\hat{\theta}(1-\hat{\theta})}{2}\log_e(1-2\mathrm{Q})\,,\end{aligned} \tag{3}$$

where $\hat{\theta}$ is the estimate of θ and where $\hat{P}$ and $\hat{Q}$ are the estimates of the proportions of the nucleotide sites showing, respectively, transitional and transversional differences between the two sequences (for the analytical solution, see the Appendix). We can easily calculate these estimates from the observed number of the different nucleotide matches between the two sequences compared. In this equation, $\hat{\theta}$ is the estimate of θ for the two sequences. In practice, however, the θ's may not be the same for the two sequences. In this case, the following equation seems to be useful:

$$\hat{d}=-(\hat{\theta}_1+\hat{\theta}_2-2\hat{\theta}_1\hat{\theta}_2)\log_e\left(1-\frac{1}{\hat{\theta}_1+\hat{\theta}_2-2\hat{\theta}_1\hat{\theta}_2}\hat{P}-\hat{Q}\right)-\frac{1-\hat{\theta}_1-\hat{\theta}_2+2\hat{\theta}_1\hat{\theta}_2}{2}\log_e(1-2\hat{Q})\,, \tag{4}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the estimates of the θ's for the two sequences, respectively. This modification is due to the work of Bulmer (1991).

The sampling variance of $\hat{d}$ can be obtained by

$$V(\hat{d}) = \left(\frac{\partial d}{\partial P}\right)^2 V(P) + \left(\frac{\partial d}{\partial Q}\right)^2 V(Q) + 2\left(\frac{\partial d}{\partial P}\right)\left(\frac{\partial d}{\partial Q}\right)\mathrm{cov}(P,Q)\,, \tag{5}$$

where $V(P) = \hat{P}(1-\hat{P})/n$, $V(Q) = \hat{Q}(1-\hat{Q})/n$, and $\mathrm{cov}(P,Q) = -\hat{P}\hat{Q}/n$. Here n denotes the number of nucleotide sites examined. It becomes

$$V(\hat{d}) = [a^2\hat{P} + b^2\hat{Q} - (a\hat{P} + b\hat{Q})^2]/n\,, \tag{6}$$

where

$$\begin{aligned} a &= \frac{1}{1 - \dfrac{\hat{P}}{2\hat{\theta}(1-\hat{\theta})} - \hat{Q}}\,, \\ b &= 2\hat{\theta}(1-\hat{\theta})a + \frac{1 - 2\hat{\theta}(1-\hat{\theta})}{1-2\hat{Q}}\,. \end{aligned} \tag{7}$$

Numerical Example

Satta et al. (1987) determined the DNA sequences of parts of the NADH dehydrogenase subunit 2 (ND2) and cytochrome oxidase subunit 1 (COI) genes of mtDNAs from *Drosophila melanogaster, D. simulans,* and *D. mauritiana.* The sequences for these regions of mtDNA are also available from *D. yakuba* (Clary and Wolstenholme 1985). I therefore compared the third-codon positions (n = 254) of these sequences, between the four species. Table 2 shows the observed number (n_{ij}) of the different nucleotide pairs, at the third-codon positions, between the *D. simulans*

Table 2
Observed Numbers (n_{ij}) of 10 Different Pairs of Nucleotides, between DNA Sequence of *Drosophila simulans* and Those of *D. mauritiana, D. melanogaster,* and *D. yakuba,* at Third-Codon Positions for Parts of ND2 and COI Genes

	No. of Differences for Nucleotide Pair										
Species Compared	AA	AT	AC	AG	TT	TC	TG	CC	CG	GG	Total
D. simulans vs. *D. mauritiana*	120	1	0	5	113	9	0	4	0	2	254
D. simulans vs. *D. melanogaster*	116	4	0	6	114	9	0	2	0	3	254
D. simulans vs. *D. yakuba*	111	12	2	9	101	13	1	3	0	2	254

and each of the *D. mauritiana*, *D. melanogaster*, and *D. yakuba* sequences. The relative frequency of nucleotide pair *i* and *j* (x_{ij}) can then be obtained by dividing n_{ij} by the total number of nucleotides used. For example, x_{AG} between *D. simulans* and *D. yakuba* in table 2 is given by 9/254 = 0.0354. Then $\hat{P}$, and transversional $\hat{Q}$, and $\hat{\theta}$ are given by $\hat{P} = x_{AG}+x_{TC} = 0.0866$, $\hat{Q} = x_{AT}+x_{AC}+x_{TG}+x_{GC} = 0.0591$, and $\hat{\theta} = (x_{AG}+x_{AC}+x_{TC}+x_{TG})/2 + x_{GC}+x_{GG}+x_{CC} = 0.0689$. Therefore, equation (3) gives an estimate of $\hat{d} = 0.225$, whereas the variance of $\hat{d}$ given by equations (6) and (7) is 0.00466. The standard error of $\hat{d}$ is then 0.068. Similar computations give $\hat{d} = 0.086 \pm 0.034$ between *D. simulans* and *D. mauritiana* and $\hat{d} = 0.112 \pm 0.043$ between *D. simulans* and *D. melanogaster.*

Table 3 shows the estimates of *d* that are obtained by the Jukes and Cantor (1969) (JC), Kimura (1980) two-parameter (K2), Tajima and Nei (1984) (TN), Takahata and Kimura (1981) four-parameter (TK), and Gojobori et al. (1982*a*) six-parameter (GIN) methods and by equation (3). The estimates obtained by equation (3) are largest for all pairwise comparisons of species. If we consider that there are strong transition-transversion and G+C-content biases in the nucleotide *Drosophila* mtDNA (see table 4), this result suggests that all methods other than equation (3) give serious underestimates of *d*.

Computer Simulations

To examine the efficiency of equation (3) relative to that of other methods, in obtaining reliable estimates of *d*, I conducted a computer simulation. In this simulation, a random nucleotide sequence was generated as the common ancestral sequence for the two sequences compared under the condition that the expected nucleotide frequency is $x_i = (1-\theta)/2$ for i = A or T and $i = \theta/2$ for i = G or C, as before. I assumed that $\theta = 0.1$ and $\alpha/\beta = 10$ in this simulation. An extremely biased condition such as this has actually been observed in *Drosophila* mtDNA (table 4; also see DeSalle et al. 1987; Satta et al. 1987). The two descendant sequences were then generated from this ancestral sequence by a Monte Carlo simulation, with either $d = 0.5$ or $d = 1.0$. The number of nucleotides examined was 300, 1,000, or 3000. The estimates of *d* between two descendant sequences were obtained by the JC, K2, TN, TK, and GIN methods and by equation (3). This process was repeated 100 times.

Table 5 shows the mean and standard deviation (SD) for each estimates obtained, as well as the number of cases in which a particular method was not applicable because of negative values of the argument of the logarithms involved. The JC, K2, and TN

Table 3
Estimates of $\hat{d}$ between Sequence of *Drosophila simulans* and Those of *D. mauritiana, D. melanogaster,* and *D. yakuba,* at Third-Codon Positions for Parts of ND2 and COI Genes

	$\hat{d}$, Calculated by Method					
Species Compared	JC	K2	TN	TK	GIN	Equation (3)
D. simulans vs. *D. mauritiana*	0.062	0.063	0.066	0.084	0.079	0.086
D. simulans vs. *D. melanogaster*	0.079	0.080	0.086	0.102	0.103	0.112
D. simulans vs. *D. yakuba*	0.162	0.164	0.186	0.201	0.216	0.225

Table 4
Estimated Percent of Relative Substitution in Third-Codon Positions of Cytochrome *b* and NADH Dehydrogenase Subunit 1 Genes in *Drosophila* mtDNA

Original Nucleotide	Estimated % Relative Substitution for Mutant Nucleotide			
	A	T	C	G
A		1.7	0	9.5
T	1.7		10.5	0.8
C	3.5	37.1		0
G	33.1	2.1	0	

NOTE.—These rates were obtained by the method of Gojobori et al. (1982*b*), from DNA sequence data for the *D. nasuta* species subgroup (six species and subspecies). The number of nucleotide sites examined is 698. Data are from Tamura (1991).

methods almost always give underestimates of *d*, although there are no inapplicable cases. The TK and GIN methods give better estimates, but there are many inapplicable cases when $d = 1.0$. Equation (3) gives even better estimates than do the TK and GIN methods, and there are fewer inapplicable cases. However, in all cases the GIN method gives a smaller SD than do the TK method and equation (3).

Discussion

From the present study, equation (3) is the best among the methods examined, for estimating the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Although these conditions may not exist widely, equation (3) is useful for analyzing *Drosophila* mtDNA. A high G+C-content bias has also been observed at the third-codon positions of the major histocompatibility complex class I loci (Hughes and Nei 1988). If the extent of these biases is not strong, the TK, GIN, and TN methods are useful, but equation (3) is much simpler than these methods, and the simplicity of a method is important for practical data analysis.

Estimates of *d* obtained by the present method have large SDs when both biases are strong and when *d* is large. The high frequency of inapplicable cases observed in the simulation (table 5) reflects the large sampling errors of the estimates of *d*. Table 6 shows the SDs calculated by equations (6) and (7), for various parameter values. In this table, we can see the large SDs in the case of $\theta = 0.1$, $\alpha/\beta = 10$, and $d = 1.0$. Because such large SDs are not found in the other cases, the cause of this large SD is attributable to the combined effect of the transition-transversion and G+C-content biases. Unfortunately, we cannot decrease this SD very much by increasing the number of the nucleotides examined, because the SD is inversely proportional to the square root of the number of nucleotides, as shown by equation (6). A quadrupling of the number of nucleotides decreases the SD by only half. When *d* increases, the SD increases rapidly. For example, the SD for $d = 1.0$ is about seven times greater than that for $d = 0.5$. This indicates that the applicability of the present method is limited to the case where *d* is not very large if there are strong transition-transversion and G+C-content biases.

Table 5
Means ± SDs of Estimates of *d*

	Mean ± SD of Estimates of d, Calculated by Method					
n	JC	K2	TN	TK	GIN	Equation (3)
$d \times 100 = 50$:						
300	35 ± 4 (0)	35 ± 5 (0)	43 ± 7 (0)	47 ± 16 (5)	49 ± 11 (14)	53 ± 14 (4)
1,000	35 ± 3 (0)	35 ± 3 (0)	43 ± 5 (0)	46 ± 8 (0)	47 ± 6 (0)	51 ± 8 (0)
3,000	35 ± 1 (0)	35 ± 1 (0)	42 ± 2 (0)	44 ± 5 (0)	46 ± 3 (0)	50 ± 4 (0)
$d \times 100 = 100$:						
300	54 ± 6 (0)	54 ± 6 (0)	70 ± 11 (0)	80 ± 37 (51)	84 ± 15 (60)	83 ± 21 (40)
1,000	54 ± 4 (0)	54 ± 4 (0)	70 ± 6 (0)	79 ± 22 (50)	87 ± 11 (53)	96 ± 26 (34)
3,000	54 ± 2 (0)	54 ± 2 (0)	70 ± 4 (0)	86 ± 23 (28)	87 ± 11 (27)	101 ± 21 (23)

NOTE.—Results are of 100 replications and have been multiplied by 100. $\theta = 0.1$; and $\alpha/\beta = 10$. The numbers in parentheses are the number of inapplicable cases.

Acknowledgments

I would like to thank Drs. O. Kitagawa and T. Aotsuka for their guidance during the course of this work at Tokyo Metropolitan University. I also thank Drs. Y. N. Tobari and Y. Fuyama for their advice and encouragement during the course of this work. I further thank Drs. M. Nei and Y. Tateno for their critical readings of the manuscript and for providing many useful comments. This work was partly supported by NSF grant BSR-9096248 and NIH grant GM-20293, both to M. Nei.

APPENDIX

Derivation of Equations (3)

Let us use the symbols given in table A1 to denote the frequencies of nucleotide pairs that are different between two homologous sequences. These frequencies at time t will be denoted by $R_1(t)$, $R_2(t)$, etc. We now derive equation (3) by considering the frequency change of each nucleotide pair during a short time period Δt. Let us first consider the change of $P_1(t)$, i.e., the frequency of nucleotide pair AG. The frequency of P_1 at time $t+\Delta t$, i.e., $P_1(t+\Delta t)$, is given by

$$\begin{aligned} P_1(t+\Delta t) = {} & [1-(\alpha+2\beta)\Delta t]P_1(t) \\ & +\{\theta[R_1(t)\alpha + Q_1(t)\beta + Q_2(t)\beta] \\ & +(1-\theta)[R_4(t)\alpha + Q_3(t)\beta + Q_4(t)\beta]\}\Delta t\,. \end{aligned} \tag{A1}$$

Here, the first term in the right-hand side of the equation represents the probability of nucleotide pair AG remaining unchanged during time Δt, if we neglect the rare event that both of A and G change simultaneously. In the second term, $R_1(t)\theta\alpha\Delta t$ represents the probability that nucleotide pair AA changes to AG during time Δt, again if we neglect the event of simultaneous changes of A and G. The remaining terms represent the contributions of other nucleotide pairs to the frequency of AG at time $t+\Delta t$. Similarly, the frequencies of other nucleotide pairs at time $t+\Delta t$ can be written as follows:

Table 6
Expected SDs of *d* for Various Parameter Values (n = 1,000)

PARAMETERS		SD FOR TRUE *d* OF				
θ	α/β	0.05	0.10	0.25	0.50	1.00
0.5	1	0.0073	0.0105	0.0181	0.0297	0.0597
0.1	1	0.0074	0.0108	0.0195	0.0349	0.0862
0.5	10	0.0074	0.0109	0.0196	0.0356	0.0913
0.1	10	0.0078	0.0121	0.0270	0.0762	0.5322

$$\begin{aligned} P_2(t+\Delta t) = {} & [1-(\alpha+2\beta)\Delta t]P_2(t) \\ & +\{\theta[R_2(t)\alpha + Q_1(t)\beta + Q_3(t)\beta] \\ & +(1-\theta)[R_3(t)\alpha + Q_2(t)\beta + Q_4(t)\beta]\}\Delta t\,, \end{aligned} \tag{A2}$$

$$\begin{aligned} Q_1(t+\Delta t) = {} & [1-2(\theta\alpha+\beta)\Delta t]Q_1(t) \\ & +(1-\theta)\{[Q_2(t)+Q_3(t)]\alpha \\ & +[R_1(t)+R_2(t)+P_1(t)+P_2(t)]\beta\}\Delta t\,, \end{aligned} \tag{A3}$$

$$\begin{aligned} Q_2(t+\Delta t) = {} & [1-(\alpha+2\beta)\Delta t]Q_2(t) \\ & +\{\theta[Q_1(t)\alpha + R_1(t)\beta + P_1(t)\beta] \\ & +(1-\theta)[Q_4(t)\alpha + R_3(t)\beta + P_2(t)\beta]\}\Delta t\,, \end{aligned} \tag{A4}$$

$$\begin{aligned} Q_3(t+\Delta t) = {} & [1-(\alpha+2\beta)\Delta t]Q_3(t) \\ & +\{\theta[Q_1(t)\alpha + R_2(t)\beta + P_2(t)\beta] \\ & +(1-\theta)[Q_4(t)\alpha + R_4(t)\beta + P_1(t)\beta]\}\Delta t\,, \end{aligned} \tag{A5}$$

$$\begin{aligned} Q_4(t+\Delta t) = {} & \{1-2[(1-\theta)\alpha + \beta]\Delta t\}Q_4(t) \\ & +\theta\{[Q_2(t)+Q_3(t)]\alpha+[R_3(t)+R_4(t)+P_1(t)+P_2(t)]\beta\}\Delta t\,. \end{aligned} \tag{A6}$$

Furthermore, we have

$$\theta = R_3(t)+R_4(t)+P_1(t)+P_2(t)+Q_2(t)+Q_3(t)+2Q_4(t)\,, \tag{A7}$$

$$(1-\theta) = R_1(t)+R_2(t)+P_1(t)+P_2(t)+2Q_1(t)+Q_2(t)+Q_3(t)\,. \tag{A8}$$

Let $P(t)$ and $Q(t)$, respectively, be the P and Q between the two sequences that diverged t time units ago. Then the P and Q at time $t+\Delta t$, i.e., $P(t+\Delta t)$ and $Q(t+\Delta t)$, respectively, can be obtained from equations (A1), . . . , (A8) as follows:

Table A1
Different Types of Nucleotide Matches between Two Homologous Sequences, and Their Frequencies

Type of Nucleotide Match	Frequency
Same:	
AA	R_1
TT	R_2
CC	R_3
GG	R_4
Total	R
Different:	
Transition:	
AG	P_1
GA	P_1
TC	P_2
CT	P_2
Total	P
Transversion:	
AT	Q_1
TA	Q_1
AC	Q_2
CA	Q_2
TG	Q_3
GT	Q_3
CG	Q_4
GC	Q_4
Total	Q

$$\begin{aligned} P(t+\Delta t) &= 2P_1(t)+2P_2(t) \\ &= [1-2(\alpha+\beta)\Delta t]P(t)+4\theta(1-\theta)\alpha\Delta t \\ &\quad -2\{\theta[2Q_1(t)+Q_2(t)+Q_3(t)] \\ &\qquad +(1-\theta)[Q_2(t)+Q_3(t)+2Q_4(t)]\}(\alpha-\beta)\Delta t \\ &= [1-2(\alpha+\beta)\Delta t]P(t)+4\theta(1-\theta)[\alpha-(\alpha-\beta)Q(t)]\Delta t\,, \end{aligned} \tag{A9}$$

$$Q(t+\Delta t) = 2Q_1(t)+2Q_2(t)+2Q_3(t)+2Q_4(t) = (1-4\beta\Delta t)Q(t)+2\beta\Delta t\,, \tag{A10}$$

where the last equality of equation (A9) can be obtained from the following two equations:

$$\begin{aligned} &2Q_1(t)+Q_2(t)+Q_3(t)-(1-\theta)Q(t) \\ &\quad = (1-\alpha-3\beta)^t[2Q_1(0)+Q_2(0)+Q_3(0)-(1-\theta)Q(0)] = 0\,, \end{aligned} \tag{A11}$$

$$Q_2(t)+Q_3(t)+2Q_4(t)-\theta Q(t) = (1-\alpha-3\beta)^t[Q_2(0)+Q_3(0)+2Q_4(0)-\theta Q(0)] = 0\,. \tag{A12}$$

These two equations can be approximated to the following set of differential equations, if Δt is very small:

$$\frac{dP(t)}{dt} = -2(\alpha+\beta)P(t)+4\theta(1-\theta)[\alpha-(\alpha-\beta)Q(t)]\,, \tag{A13}$$

$$\frac{dQ(t)}{dt} = 2\beta-4\beta Q(t)\,. \tag{A14}$$

Solution of these equations with the initial condition $P(0) = Q(0) = 0$ gives

$$P(t) = \theta(1-\theta)[1+e^{-4\beta t}-2e^{-2(\alpha+\beta)t}]\,, \tag{A15}$$

$$Q(t) = {}^1\!/\!_2-{}^1\!/\!_2 e^{-4\beta t}\,. \tag{A16}$$

$P(t)$ depends on the G+C content, θ, whereas $Q(t)$ does not. $P(t)$ is highest when $\theta = 0.5$ and decreases as θ deviates from 0.5. At time $t = \infty$, $P(t) = \theta(1-\theta)$ and $Q(t) = {}^1\!/\!_2$.

Equations (A15) and (A16) can be written as

$$2(\alpha+\beta)t = -\log_e\left[1-\frac{1}{2\theta(1-\theta)}P(t)-Q(t)\right], \tag{A17}$$

$$4\beta t = -\log_e[1-2Q(t)]\,. \tag{A18}$$

From these two equations and equation (2), we obtain equation (3).

LITERATURE CITED

BULMER, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. **8**:868–883.

CLARY, D. O., and D. R. WOLSTENHOLME. 1985. The mitochondrial DNA molecule of *Drosophila yakuba:* nucleotide sequence, gene organization and genetic code. J. Mol. Evol. **22**: 252–271.

DESALLE, R., T. FREEDMAN, E. M. PRAGER, and A. C. WILSON. 1987. Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian Drosophila. J. Mol. Evol. **26**:157–164.

GOJOBORI, T., K. ISHII, and M. NEI. 1982*a*. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. **18**:414–423.

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

GOJOBORI, T., E. N. MORIYAMA, and M. KIMURA. 1990. Statistical methods for estimating sequence divergence. Pp. 531–550 *in* R. F. DOOLITTLE, ed. Methods in enzymology. Vol. **183:** Molecular evolution: computer analysis of protein and nucleic acid sequences. Academic Press, San Diego.

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.

HUGHES, A. L., and M. NEI. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167–170.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 121–123 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

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

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

SATTA, Y., H. ISHIWA, and S. I. CHIGUSA. 1987. Analysis of nucleotide substitutions of mitochondrial DNAs in *Drosophila melanogaster* and its sibling species. Mol. Biol. Evol. **4**:638–650.

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

TAKAHATA, N., and M. KIMURA. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics **98**:641–657.

TAMURA, K. 1991. Molecular evolution of mitochondrial DNA sequences in *Drosophila*. Ph.D. thesis, Tokyo Metropolitan University, Tokyo.

WALTER M. FITCH, reviewing editor

Received July 16, 1991; revision received December 16, 1991

Accepted December 16, 1991