J Mol Evol (1995) 40:190–226

JOURNAL OF MOLECULAR EVOLUTION

New Methods for Estimating the Numbers of Synonymous and Nonsynonymous Substitutions

Yasuo Ina

National Institute of Genetics, Mishima 411, Japan

Received: 13 October 1993 / Accepted: 5 March 1994

Abstract. New methods for estimating the numbers of synonymous and nonsynonymous substitutions per site were developed. The methods are unweighted pathway methods based on Kimura's two-parameter model. Computer simulations were conducted to evaluate the accuracies of the new methods, Nei and Gojobori's (NG) method, Miyata and Yasunaga's (MY) method, Li, Wu, and Luo's (LWL) method, and Pamilo, Bianchi, and Li's (PBL) method. The following results were obtained: (1) The NG, MY, and LWL methods give overestimates of the number of synonymous substitutions and underestimates of the number of nonsynonymous substitutions. The major cause for the biased estimation is that these three methods underestimate the number of synonymous sites and overestimate the number of nonsynonymous sites. (2) The PBL method gives better estimates of the numbers of synonymous and nonsynonymous substitutions than those obtained by the NG, MY, and LWL methods. (3) The new methods also give better estimates of the numbers of synonymous and nonsynonymous substitutions than those obtained by the NG, MY, and LWL methods. In addition, estimates of the numbers of synonymous and nonsynonymous sites obtained by the new methods are reasonably accurate. (4) In some cases, the new methods and the PBL method give biased estimates of substitution numbers. However, from the number of nucleotide substitutions at the third position of codons, we can examine whether estimates obtained by the new methods are good or not, whereas we cannot make an examination of estimates obtained by the PBL method. (5) When there are strong transition/transversion and nucleotide-frequency biases like mitochondrial genes, all of the above methods give biased estimates of substitution numbers. In such cases, Kondo et al.'s method is recommended to be used for estimating the number of synonymous substitutions, although their method cannot estimate the number of nonsynonymous substitutions and is time-consuming. These results, particularly result (1), call for reexaminations of some genes. This is because evolutionary pictures of genes have often been discussed on the basis of results obtained by the NG, MY, and LWL methods, which are favorable for the neutral theory of molecular evolution.

Key words: Synonymous and nonsynonymous substitutions — Miyata and Yasunaga's method — Li, Wu, and Luo's method — Nei and Gojobori's method — Pamilo, Bianchi, and Li's method

Introduction

The number of substitutions per site between nucleotide or amino acid sequences is one of the most fundamental quantities for molecular evolutionary studies. In particular, it is of great importance to estimate the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site separately, because estimates of d_S and d_N are used not only for reconstruction of molecular phylogenetic trees but also for a statistical test for the neutral theory of molecular evolution (Kimura 1968, 1983). Furthermore, evolutionary rates have been controversial, and many authors have discussed this topic using esti-

mates of d_S and d_N (Wu and Li 1985; Kikuno et al. 1985; Easteal 1985, 1990; Li and Tanimura 1987; Li et al. 1987; Moriyama 1987; Miyata et al. 1987a, 1987b, 1990; Shields et al. 1988; Sharp and Li 1989; Wolfe et al. 1989; Moriyama and Gojobori 1992; Ohta 1993; for review, Ohta 1992). For these reasons, various methods for estimating d_S and d_N have been proposed (for review, Nei 1987, pp. 73–79). Among them, Miyata and Yasunaga's (MY) method (1980), Li, Wu, and Luo's (LWL) method (1985), and Nei and Gojobori's (NG) method (1986) have been widely used. All of these three methods are based on the following simple assumptions: (1) All the methods assume Jukes and Cantor's (1969) one-parameter model as a mutation matrix. (2) The MY and NG methods assume that substitutions follow Jukes and Cantor's model, and the LWL method assumes Kimura's (1980) two-parameter model as a substitution matrix. However, these assumptions do not necessarily hold. Furthermore, although these methods give formulae for estimating d_S and d_N, all of the formulae are approximate because they are not directly obtained from the above assumptions. Thus, the accuracies of these methods should be evaluated by computer simulations. (It is difficult to analytically examine the accuracies of these methods.)

Nei and Gojobori (1986) conducted a computer simulation by Gojobori's (1983) method and concluded that the MY, LWL, and NG methods give good estimates of d_S and d_N. However, their computer simulation had the following problems: (1) They did not study the accuracies of these methods in various situations in terms of mutation scheme, selection scheme, and sequence length. Thus, the applicability and inapplicability of these methods are unclear. (2) The expectations of d_S and d_N themselves were probably biased in the computer simulation, because Nei and Gojobori computed the expectations of the numbers of synonymous (S) and nonsynonymous (N) sites from Jukes and Cantor's model, not from a given mutation matrix. This indicates that Nei and Gojobori's computer simulation was favorable for the MY, LWL, and NG methods. (3) Nei and Gojobori conducted only one replication of computer simulation. Because of these problems, particularly problem (2), their conclusion should be reexamined. Thus, I conducted a large-scale series of computer simulations and reevaluated the accuracies of the MY, LWL, and NG methods. Furthermore, I developed new methods for estimating d_S and d_N. The methods assume Kimura's two-parameter model as a mutation matrix. Recently, Pamilo and Bianchi (1993) and Li (1993) proposed a new method for estimating d_S and d_N, which does not assume that mutations follow Jukes and Cantor's model. In this paper, I call their new method the PBL method for short. By computer simulations, in addition to the MY, LWL, and NG methods, I evaluated the accuracies of the PBL method and the new methods developed in this study. In this paper, I present the results obtained by computer simulations.

Simulation Method

Theory. Gojobori (1983) developed a simulation method to evaluate the accuracies of methods for estimating the numbers of synonymous and nonsynonymous substitutions between a pair of homologous nucleotide sequences. In his method, the equal probability of nucleotide changes (Jukes and Cantor's [1969] one-parameter model) by mutation is assumed when the expected numbers of synonymous and nonsynonymous sites are computed. In actual nucleotide sequences, however, this assumption does not always hold. It is observed (Gojobori et al. 1982b; Li et al. 1984; for review, Nei 1987, pp. 27–29) that transitional mutations occur more frequently than transversional ones and that the mutation rates among the four nucleotides are not equal. Taking into account these observations, I modified Gojobori's method so that the unequal probability of nucleotide changes by mutation can be incorporated.

We denote by λ_{ij} the mutation rate from nucleotide i to nucleotide j during one evolutionary time unit (e.g., 1 year, one generation). From these values of λ_{ij}, we can compute the number of (s_i) of synonymous sites for codon i, which is defined as the sum of the proportion of synonymous mutations to allowable total mutations at each position of the codon. For example, the number of synonymous sites for codon TTT, which encodes phenylalanine, is given by

$$s_{TTT} = 0 + 0 + \frac{\lambda_{TC}}{\lambda_{TC} + \lambda_{TA} + \lambda_{TG}} = \frac{\lambda_{TC}}{\lambda_{TC} + \lambda_{TA} + \lambda_{TG}}$$

because only a mutation from T to C at the third position is synonymous. The values of s_i for the universal genetic code are listed in Table 1. In Jukes and Cantor's model [$\lambda_{ij} = \alpha$ ($i \neq j$)], Table 1 reduces to Gojobori's (1983) Table 1. The number (n_i) of nonsynonymous sites for codon i is obtained by $n_i = 3 - s_i$ if the codon is a sense codon. The expected total numbers of synonymous (S) and nonsynonymous (N) sites for a nucleotide sequence of L codons are given by

$$\mathrm{E}(S) = L \sum_{i=TTT}^{GGG} s_i q_i \tag{1}$$

$$\mathrm{E}(N) = L \sum_{i=TTT}^{GGG} n_i q_i = 3L - \mathrm{E}(S) \tag{2}$$

where $\mathrm{E}(\cdot)$ is the expectation operator and q_i is the equilibrium frequency of codon i.

Let s_{dij} and n_{dij} be the numbers of synonymous and nonsynonymous substitutions, respectively, during one evolutionary time unit between codons i and j. We define s_{dij} as the number of nucleotide differences between codons i and j which encode the same amino acid, and n_{dij} as the number of nucleotide differences between codons i and j which code for different amino acids. The expected total number of synonymous (S_d) and nonsynonymous (N_d) substitutions during one evolutionary time unit between two nucleotide sequences of L codons are given by

$$\mathrm{E}(S_d) = L \sum_{i=TTT}^{GGG} \sum_{j=TTT}^{GGG} s_{dij} p_{ij} q_i \tag{3}$$

$$\mathrm{E}(N_d) = L \sum_{i=TTT}^{GGG} \sum_{j=TTT}^{GGG} n_{dij} p_{ij} q_i \tag{4}$$

where p_{ij} is the substitution rate from codon i to codon j during one evolutionary time unit.

We can now compute the expected numbers of synonymous (d_S)

Table 1. Number (s_i) of synonymous sites for codon i[a]

Codon	s_i	Codon	s_i	Codon	s_i	Codon	s_i
TTT	$\frac{\lambda_{TC}}{\lambda_{TC}+\lambda_{TA}+\lambda_{TG}}$	TCT	1	TAT	1	TGT	$\frac{\lambda_{TC}}{\lambda_{TC}+\lambda_{TG}}$
TTC	$\frac{\lambda_{CT}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}$	TCC	1	TAC	1	TGC	$\frac{\lambda_{CT}}{\lambda_{CT}+\lambda_{CG}}$
TTA	$\frac{\lambda_{TC}}{\lambda_{TC}+\lambda_{TA}+\lambda_{TG}}+\frac{\lambda_{AG}}{\lambda_{AT}+\lambda_{AC}+\lambda_{AG}}$	TCA	1	TAA	—	TGA	—
TTG	$\frac{\lambda_{TC}}{\lambda_{TC}+\lambda_{TA}+\lambda_{TG}}+\frac{\lambda_{GA}}{\lambda_{GT}+\lambda_{GC+GA}}$	TCG	1	TAG	—	TGG	0
CTT	1	CCT	1	CAT	$\frac{\lambda_{TC}}{\lambda_{TC}+\lambda_{TA}+\lambda_{TG}}$	CGT	1
CTC	1	CCC	1	CAC	$\frac{\lambda_{CT}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}$	CGC	1
CTA	$\frac{\lambda_{CT}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}+1$	CCA	1	CAA	$\frac{\lambda_{AG}}{\lambda_{AT}+\lambda_{AC}+\lambda_{AG}}$	CGA	$\frac{\lambda_{CA}}{\lambda_{CA}+\lambda_{CG}}+1$
CTG	$\frac{\lambda_{CT}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}+1$	CCG	1	CAG	$\frac{\lambda_{GA}}{\lambda_{GT}+\lambda_{GC}+\lambda_{GA}}$	CGG	$\frac{\lambda_{CA}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}+1$
ATT	$\frac{\lambda_{TC}+\lambda_{TA}}{\lambda_{TC}+\lambda_{TA}+\lambda_{TG}}$	ACT	1	AAT	$\frac{\lambda_{TC}}{\lambda_{TC}+\lambda_{TA}+\lambda_{TG}}$	AGT	$\frac{\lambda_{TC}}{\lambda_{TC}+\lambda_{TA}+\lambda_{TG}}$
ATC	$\frac{\lambda_{CT}+\lambda_{CA}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}$	ACC	1	AAC	$\frac{\lambda_{CT}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}$	AGC	$\frac{\lambda_{CT}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}$
ATA	$\frac{\lambda_{AT}+\lambda_{AC}}{\lambda_{AT}+\lambda_{AC}+\lambda_{AG}}$	ACA	1	AAA	$\frac{\lambda_{AG}}{\lambda_{AT}+\lambda_{AC}+\lambda_{AG}}$	AGA	$\frac{\lambda_{AC}}{\lambda_{AC}+\lambda_{AG}}+\frac{\lambda_{AG}}{\lambda_{AT}+\lambda_{AC}+\lambda_{AG}}$
ATG	0	ACG	1	AAG	$\frac{\lambda_{GA}}{\lambda_{GT}+\lambda_{GC}+\lambda_{GA}}$	AGG	$\frac{\lambda_{AC}}{\lambda_{AT}+\lambda_{AC}+\lambda_{AG}}+\frac{\lambda_{GA}}{\lambda_{GT}+\lambda_{GC}+\lambda_{GA}}$
GTT	1	GCT	1	GAT	$\frac{\lambda_{TC}}{\lambda_{TC}+\lambda_{TA}+\lambda_{TG}}$	GGT	1
GTC	1	GCC	1	GAC	$\frac{\lambda_{CT}}{\lambda_{CT}+\lambda_{CA}+\lambda_{CG}}$	GGC	1
GTA	1	GCA	1	GAA	$\frac{\lambda_{AG}}{\lambda_{AT}+\lambda_{AC}+\lambda_{AG}}$	GGA	1
GTG	1	GCG	1	GAG	$\frac{\lambda_{GA}}{\lambda_{GT}+\lambda_{GC}+\lambda_{GA}}$	GGG	1

[a] $i = TTT, TTC, TTA, TTG, \ldots, GGG$

and nonsynonymous (d_N) substitutions per site between two nucleotide sequences. That is,

$$E(d_S) = \frac{E(S_d)}{E(S)}t = \frac{\sum_{i=TTT}^{GGG}\sum_{j=TTT}^{GGG} s_{dij}p_{ij}q_i}{\sum_{i=TTT}^{GGG} s_i q_i}t \quad (5)$$

$$E(d_N) = \frac{E(N_d)}{E(N)}t = \frac{\sum_{i=TTT}^{GGG}\sum_{j=TTT}^{GGG} n_{dij}p_{ij}q_i}{\sum_{i=TTT}^{GGG} n_i q_i}t \quad (6)$$

where t is the total time that passed in the two lineages since the divergence of the two nucleotide sequences.

In the above formulation, we did not distinguish transitional and transversional changes. However, if we consider these changes, we can compute $E(d_{S,Ts})$ and $E(d_{S,Tv})$ separately, where $d_{S,Ts}$ and $d_{S,Tv}$ are the numbers of synonymous transitional and transversional substitutions per site, respectively, between two nucleotide sequences of L codons. The expected numbers of synonymous transitional and transversional substitutions per site between the two nucleotide sequences are given by

$$E(d_{S,Ts}) = \frac{E(S_{Ts})}{E(S)}t \quad (7)$$

$$E(d_{S,Tv}) = \frac{E(S_{Tv})}{E(S)}t \quad (8)$$

where S_{Ts} and S_{Tv} are the total numbers of synonymous transitional and transversional substitutions, respectively, during one evolutionary time unit between the two nucleotide sequences. The expectations of S_{Ts} and S_{Tv} can be computed by equations similar to equation (3) if the values of s_{dij} are further divided into transitional and transversional substitu-

tions. Similarly, we can compute $E(d_{N,Ts})$ and $E(d_{N,Tv})$ separately, where $d_{N,Ts}$ and $d_{N,Tv}$ are the numbers of nonsynonymous transitional and transversional substitutions per site, respectively, between the two nucleotide sequences. $d_S = d_{S,Ts} + d_{S,Tv}$ and $d_N = d_{N,Ts} + d_{N,Tv}$ always hold. In this study, I computed $E(d_{S,Ts})$, $E(d_{S,Tv})$, $E(d_{N,Ts})$, and $E(d_{N,Tv})$ separately, and then obtained $E(d_S)$ and $E(d_N)$ without using equations (5) and (6). This is because such computations enable us to examine numerically the transition/transversion ratios at synonymous and nonsynonymous sites.

As described by Gojobori (1983), we can compute the expected number of nucleotide substitutions at each position of codons. In this study, however, I computed the expected numbers of transitional and transversional substitutions separately, at each position of codons, and then obtained the expected number of nucleotide substitutions at each position of codons. Such computations are useful for a theoretical examination of the transition/transversion ratio at each position of codons.

Let $l_{k,Ts,ij}$ and $l_{k,Tv,ij}$ be the numbers of transitional and transversional substitutions, respectively, at position k during one evolutionary time unit between codons i and j. The expected total numbers of transitional ($L_{k,Ts}$) and transversional ($L_{k,Tv}$) substitutions at position k during one evolutionary time unit between two nucleotide sequences of L codons are given by

$$E(L_{k,Ts}) = L \sum_{i=TTT}^{GGG} \sum_{j=TTT}^{GGG} l_{k,Ts,ij} p_{ij} q_i \tag{9}$$

$$E(L_{k,Tv}) = L \sum_{i=TTT}^{GGG} \sum_{j=TTT}^{GGG} l_{k,Tv,ij} p_{ij} q_i \tag{10}$$

The expected numbers of transitional ($d_{k,Ts}$) and transversional ($d_{k,Tv}$) substitutions per site at position k between the two nucleotide sequences are given by

$$E(d_{k,Ts}) = \frac{E(L_{k,Ts})}{L} t = \sum_{i=TTT}^{GGG} \sum_{j=TTT}^{GGG} l_{k,Ts,ij} p_{ij} q_i t \tag{11}$$

$$E(d_{k,Tv}) = \frac{E(L_{k,Tv})}{L} t = \sum_{i=TTT}^{GGG} \sum_{j=TTT}^{GGG} l_{k,Tv,ij} p_{ij} q_i t \tag{12}$$

We can compute the expected number of nucleotide substitutions per site (d_k) at position k between the two nucleotide sequences by $d_k = d_{k,Ts} + d_{k,Tv}$.

Equilibrium Codon Frequencies. Let $\mathbf{q}(t)$ be the column vector of $q_{TTT}(t)$, $q_{TTC}(t)$, $q_{TTA}(t)$, $q_{TTG}(t)$, . . . , $q_{GGG}(t)$, where $q_i(t)$ is the frequency of codon i at time t. The codon frequencies at time t can be represented in a vector form,

$$\mathbf{q}(t) = \mathbf{P}\mathbf{q}(t-1) = \ldots \mathbf{P}^t\mathbf{q}(0) \tag{13}$$

where $\mathbf{P}$ is the codon substitution matrix (transition matrix in terms of Markov chain), and the element (p_{ij}) of the matrix is the substitution rate from codon i to codon j during one evolutionary time unit. The values of p_{ij} are determined by the mutation rate and the fixation probability (f_{ij}) of codon i to codon j. For example, the value $p_{TCT,TCG}$ is given by

$$p_{TCT,TCG} = f_{TCT,TCG} \lambda_{TT} \lambda_{CC} \lambda_{TG}$$

where λ_{TT} and λ_{CC} are given by $1 - \sum_{j \neq T} p_{Tj}$ and $1 - \sum_{j \neq C} p_{Cj}$, respectively. The diagonal element p_{ii} of the substitution matrix is given by $1 - \sum_{j \neq i} p_{ij}$.

The equilibrium codon frequencies [$q_i(\infty) = q_i$] were numerically obtained by the power method. Multiplication of the codon substitution matrix by the codon frequency vector was iterated until $\sum_{i=TTT}^{GGG} \{q_i(t) - q_i(t-1)\}^2 \leq 10^{-30}$. The initial values of $q_i(0)$ used in this study were 0 for stop codons and $1/n_S$ for sense codons, where n_S is the number of sense codons for a given codon table (e.g., 61 for the universal genetic code).

Computer Simulation. From the equilibrium codon frequencies (q_i), an ancestral sequence was generated by pseudo-random numbers. Following the codon substitution matrix ($\mathbf{P}^t$) and pseudo-random numbers, substitutions occurred on the ancestral sequence. This procedure (generation of an ancestral sequence and occurrence of substitutions) was repeated.

Table 2. Mutation matrices used in this study

#	A	T	C	G
Influenza virus gene mutation scheme				
A	0.9910225	0.00149625	0.0005985	0.00688275
T	0.001368	0.98832925	0.00902025	0.0012825
C	0.0024795	0.00688275	0.99003925	0.0005985
G	0.0108585	0.00072675	0.00055575	0.987859
Pseudogene mutation scheme				
A	0.989563	0.002303	0.002548	0.005586
T	0.002205	0.992503	0.003038	0.002254
C	0.004067	0.01078	0.98285	0.002303
G	0.00784	0.00343	0.002695	0.986035
Mitochondrial gene mutation scheme				
A	0.995626	0.000486	0.001458	0.00243
T	0.001134	0.976834	0.02187	0.000162
C	0.001134	0.00729	0.991414	0.000162
G	0.01701	0.000486	0.001458	0.981046

Simulation Scheme

Mutation Scheme

In this study, three different mutation schemes were used. They are (1) influenza virus gene mutation scheme, (2) pseudogene mutation scheme, and (3) mitochondrial gene mutation scheme. The mutation matrices for these schemes are shown in Table 2.

1. Influenza virus gene mutation scheme: A substitution matrix at the third position of codons for human influenza A virus gene was estimated by Saitou (1987). In the present study, his substitution matrix was used as a mutation matrix, because it is thought that substitution patterns at the third position of codons reflect mutation patterns to some extent. The matrix obtained by Saitou was modified so that $E(d_S) = 0.01$ during one evolutionary time unit.
2. Pseudogene mutation scheme: A mutation matrix for mammalian pseudogenes was estimated by Gojobori et al. (1982b). The matrix was modified so that $E(d_S) = 0.01$ during one evolutionary time unit. This muta-

tion scheme was used by Nei and Gojobori's computer simulation.

3. Mitochondrial gene mutation scheme. The mutation matrix for this scheme was made by using Hasegawa et al.'s (1985) model. In their model, the values of π_A, π_T, π_C, and π_G were set so as to be nearly equal to those at the third position of codons in the ND4 gene of human (Anderson et al. 1981), chimpanzee, and gorilla (Brown et al. 1982). In this case, $\pi_A = 0.35$, $\pi_T = 0.15$, $\pi_C = 0.45$, and $\pi_G = 0.05$. The α/β ratio for primate mitochondrial genes was estimated to be from ten to 20 (Brown et al. 1982; Hayasaka et al. 1988; Horai et al. 1992; Kondo et al. 1993). In this study, the values of α and β were set to be 0.0486 and 0.00324, respectively, where $\alpha/\beta = 15$. Under this parameter set, $E(d_S) = 0.01$ during one evolutionary time unit.

Characteristics of these mutation matrices (i.e., transition/transversion ratio and equilibrium nucleotide frequencies) will be shown later.

Selection Scheme

Three different types of selection against nonsynonymous changes were incorporated.

1. No selection scheme: The fixation probability of nonsynonymous changes was assumed to be 1 for all pairs of amino acids. Although such flexible genes have rarely been found, this selection scheme helps us examine whether a method for estimating d_S and d_N has a necessary property; estimates of d_N should be nearly equal to those of d_S without any selection.
2. Moderate selection scheme: Miyata et al. (1979) have proposed an index which represents physicochemical similarity between a pair of amino acids. They reported that the index is strongly correlated with frequencies of amino acid substitutions. Thus, it seems to be appropriate to use the index for simulation studies which incorporate negative selection against amino acid changes. Under this selection scheme, the fixation probability of nonsynonymous changes was assumed to depend on the index. As pointed out by Nei and Gojobori (1986), this selection scheme is favorable for the MY method, because in the method, the weight for pathways between a pair of codons is based on Miyata et al.'s index.
3. Strong selection scheme: The fixation probability of nonsynonymous changes was assumed to be 0.2 irrespective of physicochemical similarity between each amino acid pair interchanged. This selection scheme was proposed by Gojobori (1983) to simulate the evolution of hemoglobin genes. However, this selection scheme can be applied to other genes because Li et al.'s (1985) analyses of various genes indicated that the average ratio of d_N to d_S was about 0.2.

Under all of the selection schemes, the fixation probabilities were assumed to be 1 for synonymous changes and 0 for stop codons. The latter two selection schemes were used by Nei and Gojobori's computer simulation.

Sequence Length

Nucleotide sequences of L = 1,000, 290, and 50 codons were used.

1. Nucleotide sequences of L = 1,000 codons: Nucleotide sequences of 1,000 codons were used in order to minimize the effect of stochastic fluctuations due to the small number of codons compared.
2. Nucleotide sequences of L = 290 codons: I examined the lengths of 52,257 amino acid sequences registered in the PIR protein database (release 36.00). The average length of the sequences was calculated to be 296. On the assumption that this value represents approximately the number of codons in typical genes analyzed in molecular evolutionary studies, nucleotide sequences of 290 codons were used.
3. Nucleotide sequences of L = 50 codons: Occasionally, the number of codons analyzed can be much smaller than 290. This is the case especially when we focus on a particular region of genes (e.g., exon, domain). For example, the number of codons compared was 57 when Hughes and Nei (1988) analyzed antigen recognition sites of MHC class I genes. Thus, nucleotide sequences of 50 codons were used in order to examine the effect of the small number of codons compared.

Expectation of the Number of Synonymous Substitutions

The total times that passed in two lineages since the divergence of two nucleotide sequences were assumed to be t = 10, 20, 30, . . . , 100 evolutionary time units. These values correspond to $E(d_S)$ = 0.1, 0.2, 0.3, . . . , 1.0, respectively.

Number of Replications

Computer simulations were conducted 100 times for each set of mutation scheme, selection scheme, sequence length, and substitution numbers. The total number of replications in this study was 27,000 ($= 3^3 \times 10 \times 100$).

Transition/Transversion Ratio and Equilibrium Nucleotide Frequencies

By numerical calculations, I examined the transition/transversion ratios and the equilibrium nucleotide frequencies under the mutation and selection schemes described above.

Table 3. Transition/transversion ratios at the third position of codons, synonymous, and nonsynonymous sites

	3rd position	Synonymous	Nonsynonymous
		Influenza virus gene mutation scheme	
No selection scheme	3.86	7.23	3.24
Moderate selection scheme	5.09	7.85	3.93
Strong selection scheme	6.36	7.30	3.25
		Pseudogene mutation scheme	
No selection scheme	1.18	2.51	0.92
Moderate selection scheme	1.44	2.43	0.93
Strong selection scheme	1.96	2.50	0.92
		Mitochondrial gene mutation scheme	
No selection scheme	5.57	8.95	4.36
Moderate selection scheme	6.25	8.95	3.95
Strong selection scheme	7.82	8.95	4.36

Table 3 shows the transition/transversion ratios under each set of the mutation and selection schemes. In the case of no selection, it is expected that the transition/transversion ratio at the third position of codons reflects directly the intrinsic transition/transversion ratio for a mutation matrix. Thus, this table shows that a transition/transversion bias is the strongest under the mitochondrial gene mutation scheme, followed by the influenza virus gene mutation scheme. Under the pseudogene mutation scheme, transversional mutations are expected to occur as frequently as transitional ones.

Table 4 shows the equilibrium nucleotide frequencies under each set of the mutation and selection schemes. The standard deviation (SD) of the nucleotide frequencies is also shown in this table because the SD value represents a measure of nucleotide-frequency biases. The nucleotide-frequency bias is the strongest under the mitochondrial gene mutation scheme, followed by the pseudogene mutation scheme. Under the influenza virus gene mutation scheme, the nucleotide-frequency bias is the weakest, although it exists. This tendency is observed under all of the selection schemes used in this study.

These results indicate that all of the mutation matrices used in this study violate the assumptions on which the MY, LWL, and NG methods are based. Furthermore, none of the mutation matrices satisfies the assumptions on which the new methods described later and the PBL method are based, because the equilibrium nucleotide frequencies deviate from equality (0.25).

Evaluation of the Accuracies of the MY, LWL, and NG Methods

By computer simulations, I evaluated the accuracies of the MY, LWL, and NG methods. I examined not only estimates of d_S and d_N but also those of S and N obtained by these three methods.

Table 4. Equilibrium nucleotide frequencies at the third position of codons

	T	C	A	G	SD[a]
			Influenza virus gene mutation scheme		
No selection scheme	0.191	0.214	0.355	0.240	0.063
Moderate selection scheme	0.185	0.215	0.359	0.241	0.066
Strong selection scheme	0.184	0.218	0.358	0.240	0.066
			Pseudogene mutation scheme		
No selection scheme	0.413	0.148	0.249	0.190	0.101
Moderate selection scheme	0.419	0.145	0.245	0.191	0.104
Strong selection scheme	0.422	0.138	0.249	0.191	0.107
			Mitochondrial gene mutation scheme		
No selection scheme	0.154	0.460	0.337	0.049	0.160
Moderate selection scheme	0.154	0.460	0.337	0.049	0.160
Strong selection scheme	0.154	0.460	0.337	0.049	0.160

[a] Standard deviation. The standard deviation of the nucleotide frequencies at the third position was calculated by $SD = \sqrt{\Sigma_{i=T}^{G}(\pi_{3i} - 0.25)^2/4}$, where π_{3i} is the frequency of nucleotide i at the third position.

Estimates of d_S and d_N

Table 5 shows the expectations and estimates of d_S and d_N under the simulation scheme of influenza virus gene mutation, no selection, and $L = 1{,}000$. It is clear that all of the NG, MY, and LWL methods give overestimates of d_S and underestimates of d_N. The extents of the biased estimation of d_S and d_N are substantial. When $t = 100$, the NG, MY, and LWL methods overestimate d_S by 23%, 46%, and 51%, respectively. The extents of underestimation of d_N by the NG, MY, and LWL methods are 30%, 32%, and 24%, respectively. Overestimation of d_S and underestimation of d_N are not specific for large values of t. For example, when $t = 10$, the NG, MY, and LWL methods overestimate d_S by 38%, 45%, and 50%, respectively. The extents of underestimation of d_N are 11%, 13%, and 12% by the NG, MY, and LWL methods, respectively. Note that under this simulation scheme, un-

Table 5. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by the NG, MY, and LWL methods under the simulation scheme of influenza virus gene mutation, no selection, and $L = 1000$[a]

	Expectation	NG	MY	LWL
		$t = 10$		
n	—	100	100	100
d_S	0.100	0.138 ± 0.017	0.145 ± 0.018	0.150 ± 0.018
d_N	0.100	0.089 ± 0.007	0.087 ± 0.007	0.088 ± 0.007
d_N/d_S	1.000	0.644	0.602	0.585
		$t = 20$		
n	—	100	100	94
d_S	0.201	0.270 ± 0.022	0.286 ± 0.024	0.297 ± 0.024
d_N	0.201	0.173 ± 0.009	0.170 ± 0.009	0.173 ± 0.009
d_N/d_S	1.000	0.642	0.596	0.582
		$t = 30$		
n	—	100	100	82
d_S	0.301	0.398 ± 0.035	0.425 ± 0.039	0.442 ± 0.038
d_N	0.301	0.252 ± 0.013	0.247 ± 0.013	0.254 ± 0.015
d_N/d_S	1.000	0.633	0.582	0.576
		$t = 40$		
n	—	100	100	73
d_S	0.402	0.524 ± 0.040	0.567 ± 0.045	0.586 ± 0.040
d_N	0.402	0.326 ± 0.016	0.320 ± 0.016	0.337 ± 0.019
d_N/d_S	1.000	0.622	0.564	0.575
		$t = 50$		
n	—	100	100	58
d_S	0.502	0.656 ± 0.053	0.716 ± 0.059	0.735 ± 0.058
d_N	0.502	0.396 ± 0.018	0.387 ± 0.018	0.412 ± 0.021
d_N/d_S	1.000	0.603	0.541	0.560
		$t = 60$		
n	—	100	100	47
d_S	0.602	0.785 ± 0.060	0.870 ± 0.071	0.890 ± 0.062
d_N	0.602	0.463 ± 0.019	0.452 ± 0.018	0.488 ± 0.022
d_N/d_S	1.000	0.591	0.520	0.548
		$t = 70$		
n	—	100	100	36
d_S	0.703	0.891 ± 0.071	0.999 ± 0.088	1.041 ± 0.090
d_N	0.703	0.528 ± 0.024	0.515 ± 0.023	0.560 ± 0.026
d_N/d_S	1.000	0.593	0.515	0.537
		$t = 80$		
n	—	100	100	35
d_S	0.803	0.998 ± 0.076	1.137 ± 0.095	1.160 ± 0.092
d_N	0.803	0.592 ± 0.022	0.575 ± 0.021	0.643 ± 0.026
d_N/d_S	1.000	0.593	0.506	0.555
		$t = 90$		
n	—	100	100	25
d_S	0.904	1.130 ± 0.112	1.317 ± 0.156	1.371 ± 0.168
d_N	0.904	0.646 ± 0.028	0.627 ± 0.026	0.713 ± 0.036
d_N/d_S	1.000	0.572	0.476	0.520
		$t = 100$		
n	—	100	100	19
d_S	1.004	1.233 ± 0.102	1.466 ± 0.145	1.518 ± 0.177
d_N	1.004	0.702 ± 0.032	0.679 ± 0.030	0.764 ± 0.043
d_N/d_S	1.000	0.569	0.463	0.503

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases. n = number of applicable cases

derestimation of d_N is not due to varying substitution rates among nonsynonymous sites. This is because the fixation probabilities are the same (i.e., 1) for all pairs of nonsynonymous changes under this simulation scheme.

Table 6 shows the expectations and estimates of d_S and d_N under the simulation scheme of pseudogene mutation, no selection, and $L = 1{,}000$. This table also clearly shows that the NG, MY, and LWL methods overestimate d_S and underestimate d_N. Overestimation of d_S and underestimation of d_N are observed again from small to large values of t. However, the extent of overestimation of d_S by the NG method decreases as t increases. When $t = 80$, estimates of d_S obtained by the NG method is close to the expectation of d_S. When $t = 90$ or 100, the NG method underestimates d_S slightly (2–6%). For these cases ($t = 80$, 90, and 100), the MY and LWL methods overestimate d_S by 20–22% and 40–44%, respectively. The extents of underestimation of d_N are 13–15%, 17–20%, and 19–22% by the NG, MY, and LWL methods, respectively. When $t = 10$, the NG, MY, and LWL methods overestimate d_S by 17%, 27%, and 36%, respectively. The extents of underestimation of d_N by the NG, MY, and LWL methods are 6%, 8%, and 10%, respectively. Note that underestimation of d_N is also not due to varying substitution rates among nonsynonymous sites under this simulation scheme.

Under the other simulation schemes of selection and sequence length, the NG, MY, and LWL methods always overestimated d_S and underestimated d_N if the mutation scheme of influenza virus gene or pseudogene was used. These results suggest that these methods give systematically biased estimates of d_S and d_N. The extents of biased estimation of d_S and d_N are dependent on the simulation schemes of mutation, selection, and t as seen above.

The mitochondrial gene mutation scheme was an exceptional case. Under this mutation scheme, d_S was overestimated when the expectation of d_S was small, whereas d_S was underestimated when the expectation of d_S was large. For example, under the simulation scheme of strong selection, $L = 1{,}000$, and $t = 10$ [$E(d_S) = 0.100$], the means of $\hat{d}_S$ obtained by the NG, MY, and LWL methods were 0.120, 0.125, and 0.128, respectively. On the other hand, when $t = 100$ [$E(d_S) = 0.999$], the means of $\hat{d}_S$ obtained by the NG, MY, and LWL methods were 0.718, 0.761, and 0.764, respectively. This underestimation of d_S for large t is probably caused by the strong transition/transversion and nucleotide-frequency biases, particularly nucleotide-frequency bias, under the mitochondrial gene mutation scheme. When the nucleotide frequencies are not equal and the divergence of nucleotide sequences is large, Jukes and Cantor's method and Kimura's two-parameter method underestimate the number of nucleotide substitutions (Kimura 1981; Takahata and Kimura 1981; Gojobori et al. 1982a; Tajima and Nei 1984; Tamura 1992). This applies to the NG, MY, and LWL methods because these methods use Jukes and Cantor's formula or Kimura's two-parameter formula to correct multiple substitutions. On the other hand, d_N was always underestimated.

Similar results for $L = 1{,}000$ were also obtained for $L = 290$ and 50. These results show that unless there is a strong nucleotide-frequency bias—say, SD > 0.11—the NG, MY, and LWL methods give overestimates of d_S and underestimates of d_N. The extent of the biased estimation depends on the simulation schemes of mutation and selection. Thus, we cannot correct the bias by certain methods, e.g., multiplying $\hat{d}_S$ and $\hat{d}_N$ by certain factors. This is because we do not know a mutation matrix or type of selection (fixation probabilities of nonsynonymous changes) for actual nucleotide sequences analyzed; we do not know the extents of biased estimation of d_S and d_N for these sequences.

Estimates of S and N

The number of substitutions per site is defined as the ratio of the total number of substitutions to the total number of sites compared. [See equations (5) and (6).] Thus, the biased estimation of d_S and d_N observed above can be explained by the following three possibilities: (1) Estimation of the number of sites (i.e., S and N) is biased. (2) Estimation of the total number of substitutions (i.e., $S_d t$ and $N_d t$) is biased. In other words, correction of multiple substitutions does not work well. (3) Estimation of the number of sites and estimation of the total number of substitutions both are biased. To clarify which of these possibilities causes the biased estimation of d_S and d_N, I examined estimates of S and N.

Table 7 clearly shows that all of the NG, MY, and LWL methods underestimate S and thus overestimate N. Under the simulation scheme of influenza virus gene mutation, overestimation of S is substantial. Under the simulation scheme of influenza virus gene mutation and moderate selection, errors of $\hat{S}$ to $E(S)$ obtained by the NG, MY, and LWL methods are 25%, 28%, and 30%, respectively. Errors of $\hat{N}$ are smaller than those of $\hat{S}$ but are 11%, 12%, and 13%, respectively, by the NG, MY, and LWL methods. Under the other selection schemes, by these methods, errors of $\hat{S}$ are larger than 23% and those of $\hat{N}$ are about 10%. Under the simulation scheme of pseudogene mutation, the extent of biased estimation of S and N is the smallest in Table 7. However, irrespective of type of selection, errors of $\hat{S}$ obtained by the NG, MY, and LWL methods are about 11%, 18%, and 22%, respectively. Errors of $\hat{N}$ are about 4%, 6%, and 8%, respectively, by the NG, MY, and LWL methods. Under the simulation scheme of mitochondrial gene mutation, the extent of biased estimation of $\hat{S}$ and $\hat{N}$ lies between those under the simulation schemes of influenza virus gene mutation and pseudogene mutation. For all the three types of selection, errors of $\hat{S}$ are about 19%, 21%, and 23%, respectively, by the NG, MY, and LWL methods. Errors of $\hat{N}$ obtained by the NG, MY, and LWL methods

Table 6. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by the NG, MY, and LWL methods under the simulation scheme of pseudogene mutation, no selection, and $L = 1{,}000$[a]

	Expectation	NG	MY	LWL
		$t = 10$		
n	—	100	100	100
d_S	0.102	0.119 ± 0.014	0.130 ± 0.016	0.139 ± 0.016
d_N	0.105	0.099 ± 0.007	0.097 ± 0.007	0.095 ± 0.007
d_N/d_S	1.030	0.835	0.743	0.687
		$t = 20$		
n	—	100	100	99
d_S	0.204	0.231 ± 0.020	0.256 ± 0.023	0.276 ± 0.024
d_N	0.210	0.196 ± 0.012	0.191 ± 0.012	0.188 ± 0.012
d_N/d_S	1.030	0.849	0.743	0.679
		$t = 30$		
n	—	100	100	93
d_S	0.306	0.333 ± 0.026	0.375 ± 0.029	0.409 ± 0.029
d_N	0.316	0.293 ± 0.014	0.283 ± 0.013	0.278 ± 0.014
d_N/d_S	1.030	0.879	0.755	0.679
		$t = 40$		
n	—	100	100	82
d_S	0.409	0.451 ± 0.032	0.514 ± 0.039	0.563 ± 0.038
d_N	0.421	0.382 ± 0.017	0.368 ± 0.016	0.362 ± 0.016
d_N/d_S	1.030	0.847	0.715	0.642
		$t = 50$		
n	—	100	100	76
d_S	0.511	0.547 ± 0.036	0.636 ± 0.044	0.703 ± 0.046
d_N	0.526	0.474 ± 0.020	0.455 ± 0.018	0.448 ± 0.019
d_N/d_S	1.030	0.865	0.716	0.637
		$t = 60$		
n	—	100	100	63
d_S	0.613	0.648 ± 0.048	0.766 ± 0.061	0.848 ± 0.064
d_N	0.631	0.561 ± 0.022	0.537 ± 0.022	0.526 ± 0.023
d_N/d_S	1.030	0.866	0.701	0.620
		$t = 70$		
n	—	100	100	64
d_S	0.715	0.735 ± 0.056	0.882 ± 0.075	1.000 ± 0.079
d_N	0.737	0.651 ± 0.027	0.621 ± 0.025	0.609 ± 0.026
d_N/d_S	1.030	0.886	0.703	0.609
		$t = 80$		
n	—	100	100	53
d_S	0.817	0.819 ± 0.058	0.997 ± 0.079	1.144 ± 0.084
d_N	0.842	0.734 ± 0.025	0.698 ± 0.023	0.682 ± 0.025
d_N/d_S	1.030	0.896	0.700	0.596
		$t = 90$		
n	—	100	100	44
d_S	0.919	0.901 ± 0.070	1.125 ± 0.103	1.316 ± 0.119
d_N	0.947	0.815 ± 0.033	0.771 ± 0.031	0.754 ± 0.034
d_N/d_S	1.030	0.904	0.686	0.573
		$t = 100$		
n	—	100	100	47
d_S	1.022	0.965 ± 0.074	1.228 ± 0.115	1.470 ± 0.173
d_N	1.052	0.896 ± 0.035	0.844 ± 0.032	0.821 ± 0.033
d_N/d_S	1.030	0.928	0.687	0.5

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases. n = number of applicable cases

Table 7. Means and standard deviations of $\hat{S}$ and $\hat{N}$ obtained by the NG, MY, and LWL methods under the simulation scheme of $L = 1{,}000$ and $t = 10$[a]

	Expectation	NG	MY	LWL
		Influenza virus gene mutation scheme		
		No selection scheme		
n	—	100	100	100
S	897.0	688.8 ± 10.6	658.2 ± 10.2	640.1 ± 10.4
N	2,103.0	2,311.2 ± 10.6	2,341.8 ± 10.2	2,359.9 ± 10.4
		Moderate selection scheme		
n	—	100	100	100
S	889.0	664.1 ± 12.6	638.3 ± 12.3	622.5 ± 12.0
N	2111.0	2,335.9 ± 12.6	2,361.7 ± 12.3	2,377.5 ± 12.0
		Strong selection scheme		
n	—	100	100	100
S	893.3	686.0 ± 12.4	656.8 ± 12.0	638.0 ± 11.7
N	2106.7	2,314.0 ± 12.4	2,343.2 ± 12.0	2,362.0 ± 11.7
		Pseudogene mutation scheme		
		No selection scheme		
n	—	100	100	100
S	776.1	691.1 ± 11.4	640.0 ± 11.4	608.5 ± 11.8
N	2,223.9	2,308.9 ± 11.4	2,360.0 ± 11.4	2,391.5 ± 11.8
		Moderate selection scheme		
n	—	100	100	99
S	781.7	701.6 ± 11.8	643.3 ± 12.2	611.0 ± 12.5
N	2,218.3	2298.4 ± 11.8	2,356.7 ± 12.2	2,389.0 ± 12.5
		Strong selection scheme		
n	—	100	100	100
S	772.7	692.4 ± 8.5	640.2 ± 8.9	609.2 ± 8.5
N	2,227.3	2,307.6 ± 8.5	2,359.8 ± 8.9	2,390.8 ± 8.5
		Mitochondrial gene mutation scheme		
		No selection scheme		
n	—	100	100	100
S	946.0	768.3 ± 10.1	743.2 ± 10.8	725.8 ± 10.8
N	2,054.0	2,231.7 ± 10.1	2,256.8 ± 10.8	2,274.2 ± 10.8
		Moderate selection scheme		
n	—	100	100	100
S	946.0	768.7 ± 10.3	744.0 ± 10.4	727.4 ± 10.6
N	2,054.0	2,231.3 ± 10.3	2256.0 ± 10.4	2,272.6 ± 10.6
		Strong selection scheme		
n	—	100	100	100
S	946.0	767.9 ± 10.1	743.3 ± 10.7	726.4 ± 11.0
N	2,054.0	2,232.1 ± 10.1	2,256.7 ± 10.7	2,273.6 ± 11.0

[a] Means and standard deviations of $\hat{S}$ and $\hat{N}$ were calculated by excluding inapplicable cases. n = number of applicable cases

are about 9%, 10%, and 11%, respectively. These results are consistent with possibilities (1) and (3), not possibility (2).

The underestimation of S and the overestimation of N seen above are due to the assumption of the equal probability of nucleotide changes (Jukes and Cantor's model) by mutation. The number of synonymous sites at the third position of a fourfold degenerate codon is always one. The number of synonymous sites at the third position of a nondegenerate codon is always zero. The frequency of threefold degenerate codons is lower than those of twofold and fourfold degenerate codons. The number of synonymous sites at the first position of codons is much smaller than that at the third position of codons. Thus, the accuracy of estimation of S largely depends on that of estimation of the number of synonymous sites at the third position of twofold degenerate codons. The number of synonymous sites at the third position of a twofold degenerate codon is the proportion of transitional mutations to total mutations at this position. For example, the number of synonymous sites at the third position of codon TTT is $\lambda_{TC}/(\lambda_{TC} + \lambda_{TA} + \lambda_{TG})$. In

general, since transitional mutations occur more frequently than transversional ones, the number of synonymous sites at the third position of a twofold degenerate codon is larger than one-third, which is expected in Jukes and Cantor's model [$\lambda_{ij} = \alpha$ $(i \neq j)$]. The NG, MY, and LWL methods give, therefore, underestimates of S and overestimates of N. Underestimation of S by the NG method has been reported by Kondo et al. (1993) for primate mitochondrial genes.

Furthermore, I computed the "expectations" of S and N as Nei and Gojobori did in their computer simulation; the "expectations" of S and N were computed from Jukes and Cantor's model irrespective of a given mutation matrix. For example, under the simulation scheme of influenza virus gene mutation, no selection, and L = 1,000, the "expectations" of S and N were 689.1 and 2310.9, respectively. Interestingly, the corresponding estimates of S and N obtained by the NG, MY, and LWL methods were much closer to the biased "expectations" than to the true expectations (Table 7). Putting these values (689.1 and 2310.9) into equations (5) and (6) as E(S) and E(N), respectively, I computed the "expectations" of d_S and d_N. The "expectations" of d_S and d_N for $t = 10$ were 0.131 and 0.091, respectively. Note that these values are not equal. The corresponding estimates of d_S and d_N obtained by the NG, MY, and LWL methods were in agreement with the biased "expectations" rather than the true expectations (Table 5). Similarly, under the simulation scheme of pseudogene mutation, no selection, and L = 1,000, the "expectations" of S and N were computed to be 692.4 and 2,307.6, respectively. The "expectations" of d_S and d_N for $t = 10$ were also computed to be 0.115 and 0.100, respectively. Note that these values are also not equal. Again, the corresponding estimates of S, N, d_S, and d_N obtained by the NG, MY, and LWL methods were much closer to the biased "expectations" than to the true expectations (Tables 6 and 7). Under the other simulation schemes, the "expectations" of d_S, d_N, S, and N were computed. The results obtained also showed that estimates of S, N, d_S, and d_N by the NG, MY, and LWL methods were in agreement with the "expectations" rather than the true expectations. These results indicate that correction of multiple substitutions used in the NG, MY, and LWL methods are not so biased. If the correction methods are seriously biased, estimates of d_S and d_N obtained by these methods cannot be in agreement with the "expectations" of d_S and d_N even when estimates of S and N are close to the "expectations" of S and N. Thus, the correction methods used in the NG, MY, and LWL methods do not lead to serious errors, although they appear to be rough approximations. Possibilities (2) and (3) are, therefore, ruled out.

From these considerations, I conclude that the major cause for biased estimation of d_S and d_N by the NG, MY, and LWL methods is a bias in estimation of S and N, which results from the assumption of the equal probability of nucleotide changes (Jukes and Cantor's model) by mutation. Thus, the NG, MY, and LWL methods give biased estimates of d_S and d_N not only when the divergence of nucleotide sequences is large but also when the divergence of nucleotide sequences is small. The proportions of synonymous and nonsynonymous differences obtained by the NG, MY, and LWL methods are also biased although some researchers (e.g., Tanaka and Nei 1989) use these quantities instead of d_S and d_N. Note that the NG, MY, and LWL methods given biased estimates of S and N.

In relation to Nei and Gojobori's computer simulation, I would like to show results of some calculations. Under the strong selection scheme, the ratio of E(d_N) to E(d_S) should be nearly equal to 0.2 because the fixation probabilities are assumed to be 0.2 for nonsynonymous changes and 1 for synonymous changes. In Nei and Gojobori's (1986) computer simulation, E(d_N)/E(d_S) = 0.174 under the pseudogene mutation scheme. From Jukes and Cantor's model, I computed the "expectations" of d_S and d_N, as Nei and Gojobori did, under the mutation schemes of influenza virus gene and mitochondrial gene, and then obtained the ratios of E(d_N) to E(d_S). The ratios were 0.137 and 0.138 under the mutation schemes of influenza virus gene and mitochondrial gene, respectively. Furthermore, as mentioned earlier, the "expectations" of d_S and d_N were not equal under the simulation scheme of no selection. These results indicate that in Nei and Gojobori's computer simulation, the expectations of d_S and d_N themselves were biased. The biases of the expectations result from the fact that Nei and Gojobori computed the expectations of S and N from Jukes and Cantor's model, not from a given mutation matrix.

New Methods for Estimating d_S and d_N

We have seen above that the assumption of the equal probability of nucleotide changes by mutation results in biased estimation of d_S and d_N but that approximations involved in the NG, MY, and LWL formulae for estimating d_S and d_N do not lead to such serious estimation errors. Taking into account these results, I developed new methods for estimating d_S and d_N. The methods were proposed to estimate S and N accurately and thus to give good estimates of d_S and d_N.

New Method 1

We assume that (1) mutations follow Kimura's two-parameter model ($\lambda_{TC} = \lambda_{CT} = \lambda_{AG} = \lambda_{GA} = \alpha$ and $\lambda_{TA} = \lambda_{TG} = \lambda_{CA} = \lambda_{CG} = \lambda_{AT} = \lambda_{AC} = \lambda_{GT} = \lambda_{GC} = \beta$) and that (2) substitutions also follow Kimura's two-parameter model. Matrices of mutation and substitution assumed here are not necessarily the same. Furthermore, we as-

sume that (3) substitution patterns at the third position of codons reflect mutation patterns. Namely, $\alpha/\beta \approx \alpha_3/\beta_3$ is assumed, where α_3 and β_3 are the transitional and transversional substitution rates, respectively, at the third position of codons.

Under assumption (1), the number (s_i) of synonymous sites for codon i becomes much simpler than that shown in Table 1. For example, the number of synonymous sites for codon TTT is given by

$$s_{TTT} = \frac{\lambda_{TC}}{\lambda_{TC} + \lambda_{TA} + \lambda_{TG}} = \frac{\alpha}{\alpha + 2\beta} = \frac{\alpha/\beta}{\alpha/\beta + 2}$$

The value of s_i for codon TTT is dependent on the α/β ratio alone. Similarly, it can be shown that the values of s_i for the other codons are also dependent on the α/β ratio alone. Thus, we can estimate the total numbers of synonymous and nonsynonymous sites for a given nucleotide sequence if we obtain the estimate of the α/β ratio. For this purpose, we have only to estimate the α_3/β_3 ratio under assumption (3). Under assumption (2), the transitional and transversional substitution rates at the third position of codons can be estimated by

$$\hat{\alpha}_3 = -\frac{1}{t}\left\{\frac{1}{4}\ln(1 - 2\hat{P}_3 - \hat{Q}_3) - \frac{1}{8}\ln(1 - 2\hat{Q}_3)\right\} \tag{14}$$

and

$$\hat{\beta}_3 = -\frac{1}{t}\left\{\frac{1}{8}\ln(1 - 2\hat{Q}_3)\right\} \tag{15}$$

respectively, where t is the divergence time of two nucleotide sequences under comparison, and $\hat{P}_3$ and $\hat{Q}_3$ are the estimates of the proportions of transitional and transversional differences, respectively, at the third position of codons. Although t is unknown in general, we do not have to obtain t in the present method to estimate d_S and d_N. This is because t's are canceled out each other in $\hat{\alpha}_3/\hat{\beta}_3$. Thus, from the resulting value of $\hat{\alpha}_3/\hat{\beta}_3$ and on the assumption of $\hat{\alpha}/\hat{\beta} \approx \hat{\alpha}_3/\hat{\beta}_3$, the number of synonymous sites for codon i can be estimated. The total numbers of synonymous (S) and nonsynonymous (N) sites for a given nucleotide sequence of L codons are also estimated by $\hat{S} = \sum \hat{s}_i$ and $\hat{N} = 3L - \hat{S}$, respectively, where $\sum$ stands for the summation of $\hat{s}_i$ over all codons in the nucleotide sequence. In practice, the estimates of S or N for two nucleotide sequences under comparison are not always the same. In such a case, the average of the estimates for the two nucleotide sequences is used as $\hat{S}$ or $\hat{N}$.

The numbers of synonymous and nonsynonymous differences are estimated in a manner similar to the NG method. However, the present method considers transitional and transversional changes separately, whereas the NG method does not distinguish these changes.

We denote by $s_{Ts,ij}$ and $s_{Tv,ij}$ the numbers of synonymous transitional and transversional differences, respectively, between codons i and j. Furthermore, we denote by $n_{Ts,ij}$ and $n_{Tv,ij}$ the numbers of nonsynonymous transitional and transversional differences, respectively, between codons i and j. When only one nucleotide difference is observed between a pair of codons, we can immediately assign it to one synonymous or nonsynonymous difference. The difference can be further divided into one transitional or transversional difference. For example, let us consider codons TTT and TTC, which encode phenylalanine. The nucleotide difference at the third position is assigned to one synonymous transitional difference. Thus, we have $\hat{s}_{Ts,TTT,TTC} = 1$, $\hat{s}_{Tv,TTT,TTC} = 0$, $\hat{n}_{Ts,TTT,TTC} = 0$, and $\hat{n}_{Tv,TTT,TTC} = 0$.

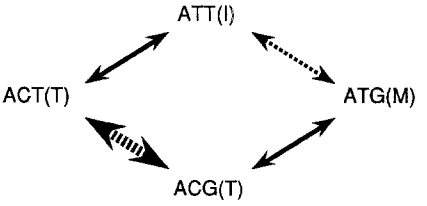

Fig. 1. Example illustrating the method for estimating the numbers of synonymous and nonsynonymous differences between a pair of codons when two nucleotide differences are observed between the codons. A synonymous change is indicated by a *thick line,* whereas nonsynonymous changes are indicated by *thin lines.* Transitional changes are indicated by *straight lines,* whereas transversional changes are indicated by *broken lines.* Amino acids encoded by each codon are shown in *parentheses.*

When two or three nucleotide differences are observed between a pair of codons, assignment of the differences is complicated because two or more possible pathways between the codons are involved. Let us consider codons ATT and ACG as an example and compute the numbers of synonymous and nonsynonymous differences between the codons. As shown in Fig. 1, there are two possible pathways between these codons. In pathway 1 (ATT ↔ ACT ↔ ACG), one synonymous transversional difference and one nonsynonymous transitional difference are involved. In pathway 2 (ATT ↔ ATG ↔ ACG), one nonsynonymous transitional difference and one nonsynonymous transversional difference are involved. Since we do not know which of these pathways appears in the evolutionary process, we assume, as done in the NG method, that these pathways appear with equal probability (1/2). Thus, we have $\hat{s}_{Ts,ATT,ACG} = (0 + 0)/2 = 0$, $\hat{s}_{Tv,ATT,ACG} = (1 + 0)/2 = 1/2$, $\hat{n}_{Ts,ATT,ACG} = (1 + 1)/2 = 1$, and $\hat{n}_{Tv,ATT,ACG} = (0 + 1)/2 = 1/2$. When three nucleotide differences are observed between a pair of codons, assignment of the differences is similarly done although it is much more complicated. If a stop codon is involved in a pathway, such a pathway is eliminated from assignment of nucleotide differences. This is because substitution of stop codons does not occur and thus such a pathway does not appear in the evolutionary process.

The total numbers of synonymous transitional (S_{Ts}) and transversional (S_{Tv}) differences between two nucle-

otide sequences are estimated by $\hat{S}_{Ts} = \sum \hat{s}_{Ts,ij}$ and $\hat{S}_{Tv} = \sum \hat{s}_{Tv,ij}$, respectively, where $\sum$ stands for the summation of $\hat{s}_{Ts,ij}$ and $\hat{s}_{Tv,ij}$ over all codon pairs between the two nucleotide sequences. Similarly, the total numbers of nonsynonymous transitional (N_{Ts}) and transversional (N_{Tv}) differences between the two nucleotide sequences are estimated by $\hat{N}_{Ts} = \sum \hat{n}_{Ts,ij}$ and $\hat{N}_{Tv} = \sum \hat{n}_{Tv,ij}$, respectively. Note that $\hat{S}_{Ts} + \hat{N}_{Ts} = \hat{L}_{1,Ts} + \hat{L}_{2,Ts} + \hat{L}_{3,Ts}$ and $\hat{S}_{Tv} + \hat{N}_{Tv} = \hat{L}_{1,Tv} + \hat{L}_{2,Tv} + \hat{L}_{3,Tv}$, where $\hat{L}_{k,Ts}$ and $\hat{L}_{k,Tv}$ are the estimates of the total numbers of transitional and transversional differences, respectively, at position k of codons between the two nucleotide sequences. The proportions of synonymous transitional (P_S) and transversional (Q_S) differences are estimated by

$$\hat{P}_S = \frac{\hat{S}_{Ts}}{\hat{S}}, \qquad \hat{Q}_S = \frac{\hat{S}_{Tv}}{\hat{S}} \tag{16}$$

To correct multiple substitutions, we use Kimura's (1980) formula. The estimate of d_S is obtained by

$$\hat{d}_S = -\frac{1}{2}\ln\{(1 - 2\hat{P}_S - \hat{Q}_S)\sqrt{1 - 2\hat{Q}_S}\} \tag{17}$$

Since this formula cannot be directly obtained from the above assumptions, it is an approximate formula. The accuracy of formula (17) will be evaluated later by computer simulations.

It seems to be difficult to derive rigorously a formula for the sampling variance of $\hat{d}_S$, because not only estimation of P_S and Q_S but also estimation of α/β are involved in the process of estimation of d_S. However, if neglect of an error in estimation of α/β does not lead to serious problems, we can derive an approximate formula for the sampling variance of $\hat{d}_S$ as Kimura (1980) did. Neglecting the error of $\hat{\alpha}/\hat{\beta}$, we have the following approximate formula for the variance of $\hat{d}_S$:

$$\begin{aligned} V(\hat{d}_S) &\approx \left(\frac{\partial \hat{d}_S}{\partial \hat{P}_S}\right)^2 V(\hat{P}_S) + \left(\frac{\partial \hat{d}_S}{\partial \hat{Q}_S}\right)^2 V(\hat{Q}_S) \\ &\quad + 2\frac{\partial \hat{d}_S}{\partial \hat{P}_S}\frac{\partial \hat{d}_S}{\partial \hat{Q}_S}\mathrm{Cov}(\hat{P}_S, \hat{Q}_S) \\ &= \frac{1}{\hat{S}}\{a^2\hat{P}_S + b^2\hat{Q}_S - (a\hat{P}_S + b\hat{Q}_S)^2\} \end{aligned} \tag{18}$$

where

$$\begin{aligned} a &= \frac{1}{1 - 2\hat{P}_S - \hat{Q}_S} \\ b &= \frac{1}{2}\left\{\frac{1}{1 - 2\hat{P}_S - \hat{Q}_S} + \frac{1}{1 - 2\hat{Q}_S}\right\} \end{aligned} \tag{19}$$

To derive formula (18), $V(\hat{P}_S) = \hat{P}_S(1 - \hat{P}_S)/\hat{S}$, $V(\hat{Q}_S) = \hat{Q}_S(1 - \hat{Q}_S)/\hat{S}$, and $\mathrm{Cov}(\hat{P}_S, \hat{Q}_S) = -\hat{P}_S\hat{Q}_S/\hat{S}$ were used. In the above derivation, $\hat{S}$ was regarded as a constant. In other words, the sampling error of the $\hat{\alpha}/\hat{\beta}$ ratio was neglected. In reality, this is not the case. The accuracy of formula (18) will be evaluated later by computer simulations.

We can obtain the estimate of d_N replacing $\hat{P}_S$ and $\hat{Q}_S$ with $\hat{P}_N$ and $\hat{Q}_N$, respectively, in formula (17), where $\hat{P}_N = \hat{N}_{Ts}/\hat{N}$ and $\hat{Q}_N = \hat{N}_{Tv}/\hat{N}$. We can also obtain the sampling variance of $\hat{d}_N$ replacing $\hat{S}$, $\hat{P}_S$, and $\hat{Q}_S$ with $\hat{N}$, $\hat{P}_N$, and $\hat{Q}_N$, respectively, in formulae (18) and (19).

New Method 2

The difference between the present method and new method 1 lies in estimation of the α/β ratio. The other parts (i.e., estimation of the numbers of synonymous and nonsynonymous sites and estimation of the numbers of synonymous and nonsynonymous differences) are the same for the present method and new method 1. Here I explain estimation of the α/β ratio in the present method.

Unlike new method 1, a problem immediately arises if we try to use equations similar to equations (14) and (15). To use the similar equations, we have to know the number of synonymous sites, but we do not know it. However, the following iterative procedure enables us to estimate the number of synonymous sites.

Let $\hat{\alpha}_{S,r}$ and $\hat{\beta}_{S,r}$ be the estimates of the synonymous transitional and transversional substitution rates, respectively, at the r-th iteration cycle. And let $\hat{S}_r$ be the estimate of the number of synonymous sites at the r-th iteration cycle. Furthermore, let $\hat{P}_{S,r}$ and $\hat{Q}_{S,r}$ be the estimates of the proportions of synonymous transitional and transversional differences, respectively, at the r-th iteration cycle. We use the number (L) of codons compared as $\hat{S}_0$, so $\hat{P}_{S,1} = \hat{S}_{Ts}/\hat{S}_0$ and $\hat{Q}_{S,1} = \hat{S}_{Tv}/\hat{S}_0$. Note that $\hat{S}_{Ts}$ and $\hat{S}_{Tv}$ are estimated in the same way as in new method 1. $\hat{\alpha}_{S,1}$ and $\hat{\beta}_{S,1}$ are obtained by

$$\hat{\alpha}_{s,1} = -\frac{1}{t}\left\{\frac{1}{4}\ln(1 - 2\hat{P}_{S,1} - \hat{Q}_{S,1}) - \frac{1}{8}\ln(1 - 2\hat{Q}_{S,1})\right\} \tag{20}$$

and

$$\hat{\beta}_{S,1} = -\frac{1}{t}\left\{\frac{1}{8}\ln(1 - 2\hat{Q}_{S,1})\right\} \tag{21}$$

respectively, where t is the divergence time of two nucleotide sequences under comparison. Unlike new method 1, we cannot use the $\hat{\alpha}_{S,1}/\hat{\beta}_{S,1}$ ratio itself as the $\hat{\alpha}/\hat{\beta}$ ratio to estimate the values of s_i. This is because $\alpha_S/\beta_S > \alpha/\beta$. Note that although the nucleotide changes at the first position between codons CGA and AGA and between codons CGG and AGG (twofold degenerate sites) are synonymous and transversional, synonymous changes at most of twofold degenerate sites are transitional alone. Note also that although synonymous trans-

versional changes occur at the third position of codons ATT, ATC, and ATA (threefold degenerate sites), the frequencies of these codons are much less than those of twofold degenerate codons. This is why $\alpha_S/\beta_S > \alpha/\beta$.

We introduce a weighting factor (W) so that $\alpha/\beta \approx W\alpha_S/\beta_S$ can hold. Since most synonymous changes occur at the third position of codons, here we consider only the third position of codons. Since the frequency of threefold degenerate codons is lower than those of twofold and fourfold degenerate codons, we treat threefold degenerate codons as twofold degenerate codons, as in the LWL and PBL methods. At the third position of twofold (+ threefold) degenerate codons, only transitional substitutions are synonymous. At the third position of fourfold degenerate codons, both transitional and transversional substitutions are synonymous. Thus, $\alpha_S/\beta_S \approx (q_2\alpha + q_3\alpha + q_4\alpha)/(q_4\beta)$, where q_2, q_3, and q_4 are the frequencies of twofold, threefold, and fourfold degenerate codons, respectively. We have, therefore, $\alpha/\beta \approx q_4/(q_2 + q_3 + q_4)\alpha_S/\beta_S$. The weighting factor, W, is given by

$$W = \frac{q_4}{q_2 + q_3 + q_4} \tag{22}$$

Replacing q_2, q_3, and q_4 with their estimates in two nucleotide sequences under comparison, we estimate W.

Let $\hat{\alpha}_r$ and $\hat{\beta}_r$ be the estimates of the transitional and transversional mutation rates at the r-th iteration cycle. From $\hat{\alpha}_{S,1}$, $\hat{\beta}_{S,1}$, and $\hat{W}$, $\hat{\alpha}_1/\hat{\beta}_1$ is computed by

$$\frac{\hat{\alpha}_1}{\hat{\beta}_1} = \hat{W}\frac{\hat{\alpha}_{S,1}}{\hat{\beta}_{S,1}} \tag{23}$$

As in new method 1, t's in equations (20) and (21) are canceled out each other in $\hat{\alpha}_{S,1}/\hat{\beta}_{S,1}$. Thus, from the resulting value of $\hat{\alpha}_1/\hat{\beta}_1$, the number of synonymous sites (s_i) for codon i can be estimated. The total number of synonymous sites for a given nucleotide sequence is estimated at the first iteration cycle by $\hat{S}_1 = \Sigma\hat{s}_{i,1}$, where $\hat{s}_{i,1}$ is the estimate of s_i at the first iteration cycle and Σ stands for the summation of $\hat{s}_{i,1}$ over all codons in the nucleotide sequence. When $\hat{S}_1$'s are not the same for two nucleotide sequences under comparison, the average of the estimates for the two nucleotide sequences is used as $\hat{S}_1$. Similarly, at the r-th iteration cycle, $\hat{S}_r$ is computed from $\hat{P}_{S,r} = \hat{S}_{Ts}/\hat{S}_{r-1}$, $\hat{Q}_{S,r} = \hat{S}_{Tv}/\hat{S}_{r-1}$, and $\hat{W}$. This iterative procedure is continued until $\hat{S}_r$ converges to $\hat{S}_\infty$ ($\hat{S}_r = \hat{S}_{r+1} = \ldots = \hat{S}_\infty$). The number ($N$) of nonsynonymous sites for a given nucleotide sequence of L codons is given by $\hat{N}_\infty = 3L - \hat{S}_\infty$.

To correct multiple substitutions, we use Kimura's (1980) formula again. The estimate of d_S is obtained by

$$\hat{d}_S = -\frac{1}{2}\ln\{(1 - 2\hat{P}_{S,\infty} - \hat{Q}_{S,\infty})\sqrt{1 - 2\hat{Q}_{S,\infty}}\} \tag{24}$$

Since this formula cannot be directly obtained, it is an approximate formula. The accuracy of formula (24) will be evaluated later by computer simulations.

As in new method 1, neglecting the error of $\hat{\alpha}/\hat{\beta}$, we have the following approximate formula for the variance of $\hat{d}_S$:

$$V(\hat{d}_S) \approx \frac{1}{\hat{S}_\infty}\{a^2\hat{P}_{S,\infty} + b^2\hat{Q}_{S,\infty} - (a\hat{P}_{S,\infty} + b\hat{Q}_{S,\infty})^2\} \tag{25}$$

$$a = \frac{1}{1 - 2\hat{P}_{S,\infty} - \hat{Q}_{S,\infty}}$$

$$b = \frac{1}{2}\left\{\frac{1}{1 - 2\hat{P}_{S,\infty} - \hat{Q}_{S,\infty}} + \frac{1}{1 - 2\hat{Q}_{S,\infty}}\right\} \tag{26}$$

In the above treatment, $\hat{S}_\infty$ was regarded as a constant. In other words, the sampling errors of the $\hat{\alpha}/\hat{\beta}$ ratio and $\hat{W}$ [$= \hat{q}_4/(\hat{q}_2 + \hat{q}_3 + \hat{q}_4)$] were neglected. In reality, this is not the case. The accuracy of formula (25) will be evaluated later by computer simulations.

We can obtain the estimate of d_N replacing $\hat{P}_{S,\infty}$ and $\hat{Q}_{S,\infty}$ with $\hat{P}_{N,\infty}$ and $\hat{Q}_{N,\infty}$, respectively, in formula (24), where $\hat{P}_{N,\infty} = \hat{N}_{Ts}/\hat{N}_\infty$ and $\hat{Q}_{N,\infty} = \hat{N}_{Tv}/\hat{N}_\infty$. We can also obtain the sampling variance of $\hat{d}_N$ replacing $\hat{S}_\infty$, $\hat{P}_{S,\infty}$, and $\hat{Q}_{S,\infty}$ with $\hat{N}_\infty$, $\hat{P}_{N,\infty}$, and $\hat{Q}_{N,\infty}$, respectively, in formulae (25) and (26).

Comparison Among New Methods 1 and 2 and the MY, LWL, and NG Methods

The difference between new methods 1 and 2 lies only in estimation of the α/β ratio. If the α/β ratio estimated by a certain method is given, the new methods give the same estimates of d_S, d_N, S, and N. In the special case of $\hat{\alpha}/\hat{\beta} = 1$, the new methods give the same estimates of S and N as the NG method does. Furthermore, in this case, if we use Jukes and Cantor's formula instead of Kimura's formula to correct multiple substitutions, the new methods reduce to the NG method. In this sense, the new methods may be characterized as extensions of the NG method. Note that if we do not distinguish transitional and transversional differences, the estimates of the numbers of synonymous and nonsynonymous differences are always the same for the NG method and the new methods.

The most notable difference between previously developed methods (e.g., the MY method, the LWL method, the NG method) and new methods 1 and 2 lies in estimation of the numbers of synonymous and nonsynonymous sites. In previously developed methods, the values of s_i are fixed; they are computable even if a pair of homologous nucleotide sequences are not given. This is because in Jukes and Cantor's model [$\lambda_{ij} = \alpha$ $(i \neq j)$], the values of s_i are independent of α. For example, the

Table 8. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by new methods 1 (NEW1) and 2 (NEW2) under Kimura's two-parameter model

	Expectation	NEW1	NEW2
		No selection	
		$t = 10$	
n^a	—	100	100
d_S	0.1	0.104 ± 0.012	0.105 ± 0.013
d_N	0.1	0.101 ± 0.008	0.101 ± 0.008
d_N/d_S	1.0	0.972	0.965
		$t = 30$	
n	—	100	100
d_S	0.3	0.305 ± 0.025	0.308 ± 0.025
d_N	0.3	0.305 ± 0.015	0.303 ± 0.015
d_N/d_S	1.0	1.000	0.985
		$t = 50$	
n	—	100	100
d_S	0.5	0.515 ± 0.035	0.518 ± 0.035
d_N	0.5	0.508 ± 0.024	0.507 ± 0.024
d_N/d_S	1.0	0.987	0.978
		$t = 100$	
n	—	100	100
d_S	1.0	1.008 ± 0.079	1.015 ± 0.082
d_N	1.0	1.006 ± 0.061	1.003 ± 0.061
d_N/d_S	1.0	0.998	0.988
		Moderate selection	
		$t = 10$	
n	—	100	100
d_S	0.100	0.104 ± 0.013	0.106 ± 0.013
d_N	0.051	0.051 ± 0.005	0.050 ± 0.005
d_N/d_S	0.512	0.488	0.477
		$t = 30$	
n	—	100	100
d_S	0.300	0.301 ± 0.021	0.306 ± 0.022
d_N	0.154	0.153 ± 0.009	0.152 ± 0.009
d_N/d_S	0.512	0.507	0.495
		$t = 50$	
n	—	100	100
d_S	0.500	0.492 ± 0.034	0.502 ± 0.036
d_N	0.256	0.246 ± 0.013	0.245 ± 0.012
d_N/d_S	0.512	0.500	0.487
		$t = 100$	
n	—	100	100
d_S	1.000	0.973 ± 0.091	0.999 ± 0.096
d_N	0.512	0.470 ± 0.023	0.467 ± 0.023
d_N/d_S	0.512	0.483	0.467
		Strong selection	
		$t = 10$	
n	—	100	100
d_S	0.10	0.101 ± 0.011	0.104 ± 0.011
d_N	0.02	0.020 ± 0.003	0.020 ± 0.003
d_N/d_S	0.2	0.199	0.191
		$t = 30$	
n	—	100	100
d_S	0.30	0.298 ± 0.024	0.307 ± 0.025
d_N	0.06	0.060 ± 0.005	0.059 ± 0.005
d_N/d_S	0.2	0.202	0.193
		$t = 50$	
n	—	100	100
d_S	0.50	0.495 ± 0.033	0.510 ± 0.034
d_N	0.10	0.106 ± 0.008	0.105 ± 0.007
d_N/d_S	0.2	0.213	0.205
		$t = 100$	
n	—	100	100
d_S	1.00	1.018 ± 0.103	1.062 ± 0.110
d_N	0.20	0.211 ± 0.010	0.210 ± 0.010
d_N/d_S	0.2	0.208	0.198

[a] Number of applicable cases

value of s_i for codon TTT is always 1/3 [= $\alpha/(3\alpha)$]. On the other hand, in new methods 1 and 2, s_i's are variables; the values of s_i can be estimated only when a pair of homologous nucleotide sequences are given. This is because in Kimura's two-parameter model, the values of s_i depend on the α/β ratio, as seen earlier, which can be estimated only when a pair of homologous nucleotide sequences are given.

Evaluation of the Estimation Formulae

The new methods have the same theoretical problem as the NG and MY methods do; at twofold and threefold degenerate sites, nucleotide substitutions do not follow Kimura's two-parameter model (Jukes and Cantor's model for the NG and MY methods). So, to examine whether Kimura's formula is appropriate for correction of multiple substitutions, I conducted computer simulations in which Kimura's two-parameter model ($\alpha/\beta = 10$) was used as a mutation matrix. The number of replications was 100 for each set of simulation schemes. Table 8 shows the results for $L = 1{,}000$ codons. We can see from this table that estimates of d_S and d_N obtained by the new methods are in good agreement with their expectations. In the case of $L = 290$ codons, the same results were obtained. Thus, the new estimation formulae are appropriate for multiple-hit correction, although they appear to be rough approximations.

Evaluation of the Accuracies of the PBL and the New Methods

By computer simulations, I evaluated the accuracies of the PBL method and the new methods presented above in terms of estimates of d_S, d_N, S, and N. It should be noted that all of the mutation matrices used in this study are unfavorable for the new methods and the PBL method. This is because the mutation matrices violate the assumptions on which these methods are based, as described earlier.

Table 9. Means and standard deviations of $\hat{S}$ and $\hat{N}$ obtained by the PBL method and new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of $L = 1000$ and $t = 10$[a]

	Expectation	PBL[b]	NEW1[c]	NEW2
		Influenza virus gene mutation scheme		
		No selection scheme		
n[d]	—	100	100	100
S	897.0	948.9 ± 48.3 (914.6 ± 41.4)	898.7 ± 22.2	882.4 ± 30.9
N	2,103.0	2051.1 ± 48.3 (2,085.4 ± 41.4)	2,101.3 ± 22.2	2,117.6 ± 30.9
		Moderate selection scheme		
n	—	100	100	100
S	889.0	957.0 ± 45.2 (993.3 ± 35.5)	912.9 ± 18.4	879.3 ± 29.9
N	2,111.0	2,043.0 ± 45.2 (2,006.7 ± 35.5)	2,087.1 ± 18.4	2,120.7 ± 29.9
		Strong selection scheme		
n	—	100	100	100
S	893.3	954.3 ± 50.9 (1,050.5 ± 21.6)	929.3 ± 17.7	883.0 ± 31.9
N	2,106.7	2,045.7 ± 50.9 (1,949.5 ± 21.6)	2,070.7 ± 17.7	2,117.0 ± 31.9
		Pseudogene mutation scheme		
		No selection scheme		
n	—	100	100	100
S	776.1	752.1 ± 58.6 (789.9 ± 44.3)	801.1 ± 29.7	771.9 ± 34.9
N	2,223.9	2,247.9 ± 58.6 (2,210.1 ± 44.3)	2,198.9 ± 29.7	2,228.1 ± 34.9
		Moderate selection scheme		
n	—	99	100	100
S	781.7	741.8 ± 62.7 (861.9 ± 46.4)	831.4 ± 28.9 (778.3 ± 10.9)	777.8 ± 32.3
N	2,218.3	2,258.2 ± 62.7 (2,138.1 ± 46.4)	2,168.6 ± 28.9 (2,221.7 ± 10.9)	2,222.2 ± 32.3
		Strong selection scheme		
n	—	100	100	100
S	772.7	752.3 ± 61.1 (992.2 ± 39.9)	862.7 ± 29.6 (774.1 ± 7.5)	775.9 ± 34.4
N	2,227.3	2,247.7 ± 61.1 (2,007.8 ± 39.9)	2,137.3 ± 29.6 (2,225.9 ± 7.5)	2,224.1 ± 34.4
		Mitochondrial gene mutation scheme		
		No selection scheme		
n	—	100	100	100
S	946.0	970.3 ± 25.1 (985.4 ± 24.0)	961.8 ± 14.3	957.1 ± 18.1
N	2,054.0	2,029.7 ± 25.1 (2,014.6 ± 24.0)	2,038.2 ± 14.3	2,042.9 ± 18.1
		Moderate selection scheme		
n	—	100	100	100
S	946.0	972.3 ± 27.7 (1,000.6 ± 19.3)	967.7 ± 15.8	959.6 ± 20.4
N	2,054.0	2,027.7 ± 27.7 (1,999.4 ± 19.3)	2,032.3 ± 15.8	2,040.4 ± 20.4
		Strong selection scheme		
n	—	100	100	100

Table 9. Continued

	Expectation	PBL[b]	NEW1[c]	NEW2
S	946.0	968.9 ± 24.4 (1,035.9 ± 13.2)	973.7 ± 13.6	953.8 ± 17.8
N	2,054.0	2031.1 ± 24.4 (1,964.1 ± 13.2)	2,026.3 ± 13.6	2,046.2 ± 17.8

[a] Means and standard deviations of $\hat{S}$ and $\hat{N}$ were calculated by excluding inapplicable cases
[b] S and N were estimated by formulae (27) and (28), respectively. Values in parentheses were estimated by formulae (29) or (30)
[c] Values in parentheses were estimated when the $\hat{\alpha}/\hat{\beta}$ ratio was given. The ratio was estimated from the geometric mean of the estimated transition/transversion ratio at synonymous sites
[d] Number of applicable cases

Estimates of S and N

Pamilo and Bianchi (1993) and Li (1993) did not show any formulae for estimating S and N. However, if we assume that substitution patterns at fourfold degenerate sites reflect mutation patterns, we can estimate S and N by

$$\hat{S} = \frac{\hat{A}_4}{\hat{K}_4}\hat{L}_2 + \hat{L}_4 \tag{27}$$

and

$$\hat{N} = \frac{\hat{B}_4}{\hat{K}_4}\hat{L}_2 + \hat{L}_0 \tag{28}$$

respectively, where $\hat{K}_i$, $\hat{A}_i$, and $\hat{B}_i$ are the estimates of the numbers of nucleotide substitutions, transitional substitutions, and transversional substitutions at i-fold degenerate sites, respectively, and $\hat{L}_i$ is the estimate of the number of i-fold degenerate sites. Since mutations at fourfold degenerate sites are always synonymous, all of these sites are synonymous sites. Similarly, since mutations at nondegenerate sites are always nonsynonymous, all of these sites are nonsynonymous sites. Transitional mutations at twofold degenerate sites are synonymous and transversional ones at these sites are nonsynonymous. (In the PBL method, the nucleotide changes at the first position between codon CGA or CGG and codon AGA or AGG are regarded as transitional changes. Furthermore, the third positions of codons ATT, ATC, and ATA are regarded as twofold degenerate sites, and the nucleotide changes at the third position between codons ATT and ATA and between codons ATC and ATA are regarded as transitional changes.) Thus, under the above assumption, the proportion of synonymous mutations to total mutations at twofold degenerate sites can be estimated as $\hat{A}_4/\hat{K}_4$ and the proportion of nonsynonymous mutations to total mutations at these sites can be estimated as $\hat{B}_4/\hat{K}_4$ $(= 1 - \hat{A}_4/\hat{K}_4)$. This is why formulae (27) and (28) give estimates of S and N, respectively. In this study, these formulae were used to estimate S and N.

In addition, formulae similar to formulae (27) and (28) were used. They are

$$\hat{S} = \frac{\hat{A}_2}{\hat{K}_2}\hat{L}_2 + \hat{L}_4 \tag{29}$$

and

$$\hat{N} = \frac{\hat{B}_2}{\hat{K}_2}\hat{L}_2 + \hat{L}_0 \tag{30}$$

In these formulae, the proportion of transitional substitutions to total substitutions at twofold degenerate sites and that of transversional substitutions to total substitutions at these sites are used as a factor of $\hat{L}_2$ to estimate S and N, respectively.

A computer program which estimates d_S and d_N by the LWL and PBL methods was provided by Dr. Wen-Hsiung Li. I modified the program so that it can estimate S and N by formulae (27) and (28) or formulae (29) and (30).

Table 9 shows estimates of S and N obtained by the PBL method [formulae (27) and (28) or formulae (29) and (30)] and the new methods. It is clear that estimates of S and N obtained by these methods are much better than those obtained by the NG, MY, and LWL methods (Table 7). Particularly, the accuracy of new method 2 is noticeable. This method gives the best estimates of S and N for all cases but the simulation scheme of influenza-virus gene mutation and no selection; even under this simulation scheme, estimates obtained by new method 2 and the best method (new method 1) are essentially the same. Errors of $\hat{S}$ to E(S) and those of $\hat{N}$ to E(N) are less than 2% and less than 0.7%, respectively, for all cases listed in Table 9. The largest errors of $\hat{S}$ and $\hat{N}$ were observed under the simulation scheme of influenza virus gene mutation and no selection. It is also noteworthy to point out that estimates of S and N obtained by new method 2 are insensitive to selection. This is because synonymous substitutions are not affected by selection and thus the transition/transversion ratio at synonymous sites is insensitive to selection, as seen in Table 3.

New method 1 also gives good estimates of S and N unless the transition/transversion ratio is small and neg-

ative selection is strong. Under the simulation scheme of no selection, this method estimates S and N accurately. This is because substitution patterns reflect directly mutation patterns without selection. Errors of $\hat{S}$ to E(S) and those of $\hat{N}$ to E(N) are less than 3% and less than 1%, respectively. Particularly, under the simulation scheme of influenza virus gene mutation, errors of $\hat{S}$ and $\hat{N}$ are 0.2% and 0.08%, respectively. Under this simulation scheme, new method 1 gives the best estimates of S and N.

In the presence of negative selection, new method 1 gives overestimates of S and underestimates of N because of inflation of the transition/transversion ratio at the third position of codons as seen in Table 3. If negative selection operates against amino acid changes, transversional substitutions at the third position of codons decrease, compared with the case of no selection, because transversional changes at the third position of twofold (+ 3-fold) degenerate codons lead to amino acid changes. For nondegenerate codons, both transitional and transversional substitutions decrease because of negative selection against amino acid changes. However, these decreases do not much affect the transition/transversion ratio at the third position of codons, because the frequency of nondegenerate codons is lower than those of twofold (+ 3-fold) and fourfold degenerate codons. As a result, the transition/transversion ratio at the third position of codons increases, compared with the case of no selection. Under the simulation scheme of pseudogene mutation and strong selection, overestimation of S and underestimation of N are substantial. Under this simulation scheme, errors of $\hat{S}$ and $\hat{N}$ are 11% and 4%, respectively. On the other hand, under the simulation scheme of influenza virus gene mutation and strong selection or mitochondrial gene mutation and strong selection, errors of $\hat{S}$ and $\hat{N}$ are not so large. Under these simulation schemes, errors of $\hat{S}$ and $\hat{N}$ are less than 4% and 1.7%, respectively. Under the simulation scheme of influenza virus gene mutation or mitochondrial gene mutation, the intrinsic transition/transversion ratio is large, as seen in Table 3. In such a case, inflation of the transition/transversion ratio at the third position of codons does not lead to such serious errors of $\hat{S}$ and $\hat{N}$. For example, under the simulation scheme of influenza virus gene mutation and no selection, we compute $\alpha/\beta = 2 \times 3.86 = 7.72$ from the transition/transversion ratio at the third position of codons in Table 3. On the other hand, under the simulation scheme of influenza virus gene mutation and strong selection, we compute $\hat{\alpha}/\hat{\beta} \approx \hat{\alpha}_3/\hat{\beta}_3 = 2 \times 6.36 = 12.72$. From these values, we estimate the extent of overestimation of the number of synonymous sites at the third position of twofold degenerate codons as $\{12.72/(12.72 + 2)\}/\{7.72/(7.72 + 2)\} = 1.09$—namely, 9%. The number of synonymous sites at the third position of fourfold degenerate codons is always one. Under the simulation scheme of influenza virus gene mutation and strong selection, we have the equilibrium frequencies of twofold, threefold, and fourfold degenerate codons by equation (13). They are computed to be 0.511, 0.056, and 0.404, respectively. Note that the frequency of threefold degenerate codons is lower than those of twofold and fourfold degenerate codons and that the number of synonymous sites at the first position of codons is much smaller than that at the third position of codons. Neglecting synonymous sites for threefold degenerate codons and those at the first position of codons, we obtain a rough estimate of the extent of overestimation of S as $0.511/(0.511 + 0.404) \times 1.09 + 0.404/(0.511 + 0.404) \times 1 = 1.05$—namely, 5%—which is in good agreement with the corresponding error of $\hat{S}$ (4%) as described earlier. Similarly, the extent of underestimation of N can be evaluated as $[3 - \{0.511 \times 12.72/(12.72 + 2) + 0.404 \times 1\}]/[3 - \{0.511 \times 7.72/(7.72 + 2) + 0.404 \times 1\}] = 0.984$—namely, 1.6%—which is in good agreement with the corresponding error of $\hat{N}$ (1.7%) as described earlier. If we treat threefold degenerate codons as twofold degenerate codons, as in new method 2, we obtain the same estimate of the extent of overestimation of S. These results indicate that when there is a strong transition/transversion bias, new method 1 gives reasonably good estimates of S and N even if strong negative selection operates against amino acid changes.

By formulae (27) and (28), the PBL method also gives much better estimates of S and N than those obtained by the NG, MY, and LWL methods. These formulae are also insensitive to selection because they use $\hat{A}_4/\hat{K}_4$ and $\hat{B}_4/\hat{K}_4$ as a factor of $\hat{L}_2$ to estimate S and N, respectively. However, for all cases in Table 9, new method 2 gives better estimates S and N than those obtained by the PBL method. Under the simulation scheme of influenza virus gene mutation or mitochondrial gene mutation, estimates of S and N obtained by new method 1 are better than those obtained by the PBL method for almost all cases; under the simulation scheme of mitochondrial gene mutation and strong selection, the PBL method gives slightly better estimates S and N than those obtained by new method 1, although the difference between these estimates is small (0.5% for $\hat{S}$ and 0.2% for $\hat{N}$). Under the simulation scheme of pseudogene mutation and no selection or moderate selection, the PBL method and new method 1 give essentially the same estimates of S and N. Only under the simulation scheme of pseudogene mutation and strong selection does the PBL method estimate S and N more accurately than new method 1 does, although new method 2 estimates S and N much more accurately than the PBL method does.

Standard deviations of $\hat{S}$ and $\hat{N}$ obtained by the PBL method are about twice as large as those obtained by new methods 1 and 2 for all cases in Table 9. Since in the PBL method nucleotide sites are divided into nondegenerate, twofold degenerate, and fourfold degenerate sites,

the sampling errors due to the small numbers of sites are large. In particular, the sampling errors of $\hat{A}_4/\hat{K}_4$ and $\hat{B}_4/\hat{K}_4$ are crucial. Since the sampling errors of $\hat{L}_0$, $\hat{L}_2$, and $\hat{L}_4$ are small, the major errors in formulae (27) and (28) are due to the sampling errors of $\hat{A}_4/\hat{K}_4$ and $\hat{B}_4/\hat{K}_4$. On the other hand, in new methods 1 and 2, nucleotide sites are divided into nucleotide sites at the third position of codons (new method 1) or synonymous and nonsynonymous sites (new method 2). Since the numbers of these sites are larger than that of fourfold degenerate sites, the sampling errors are smaller by new methods 1 and 2 than by the PBL method.

Formulae (29) and (30) are sensitive to selection because they use $\hat{A}_2/\hat{K}_2$ and $\hat{B}_2/\hat{K}_2$ as a factor of $\hat{L}_2$ to estimate S and N, respectively. By these formulae, the PBL method gives substantially biased estimates of S and N in the presence of selection. The extent of biased estimation of S and N by these formulae is larger than that by new method 1. For example, under the simulation scheme of pseudogene mutation and strong selection, S is overestimated by 28% and N is underestimated by 10%. These values are much larger than the corresponding values (11% and 4%) for new method 1. This is because the PBL method divides nucleotide sites into nondegenerate, twofold degenerate, and fourfold degenerate sites and formulae (29) and (30) use $\hat{A}_2/\hat{K}_2$ and $\hat{B}_2/\hat{K}_2$ to estimate S and N, respectively. The values of $\hat{A}_2/\hat{K}_2$ and $\hat{B}_2/\hat{K}_2$ are strongly affected by selection. On the other hand, new method 1 uses nucleotide data at the third position of codons to estimate the α/β ratio. Since these nucleotide data contain not only twofold but also fourfold degenerate sites, new method 1 is not so sensitive to selection as formulae (29) and (30). Under the simulation scheme of mitochondrial gene mutation and no selection, formulae (29) and (30) and formulae (27) and (28) give essentially the same estimates of S and N. Under the simulation scheme of influenza virus gene mutation and no selection or pseudogene mutation and no selection, estimates of S and N obtained by formulae (29) and (30) are better than those obtained by formulae (27) and (28). However, standard deviations of estimates obtained by formulae (29) and (30) are larger than those obtained by new methods 1 and 2.

Under the mitochondrial gene mutation scheme, estimates of S by the PBL method [formula (27) or (29)] and new methods 1 and 2 decreased as t increased. (Estimates of N increased as t increased.) For example, under the simulation scheme of strong selection, $L = 1{,}000$, and $t = 100$, estimates of S were 926.8, 1,021.6, 947.7, and 927.2 by formulae (27), and (29), and new methods 1 and 2, respectively. All of these values are smaller than the corresponding values in Table 9 ($t = 10$). This is because transitional substitutions detectable by Kimura's two-parameter formula rapidly reach saturation level when there are strong transition/transversion and nucleotide-frequency biases; $\hat{\alpha}/\hat{\beta}$ decreases for new methods 1 and 2, and $\hat{A}_4/\hat{K}_4$ for formula (27) and $\hat{A}_2/\hat{K}_2$ for formula (29) decrease.

Estimates of d_S and d_N

Table 10 shows estimates of d_S and d_N obtained by the PBL method and the new methods under the simulation scheme of influenza virus gene mutation, no selection, and $L = 1{,}000$. It is clear that new methods 1 and 2 give good estimates of d_S and d_N and that linear relationships between t and d_S or d_N are observed. Estimates of d_S and d_N by the PBL method are also good, although d_N is underestimated when $E(d_N) > 0.8$. These results indicate that under the simulation scheme of influenza virus gene mutation, the PBL method and new methods 1 and 2 have a necessary property; without selection, estimates of d_N should be nearly equal to those of d_S. The number of inapplicable cases is large for the PBL method. When $t = 100$, only 19% of replications is applicable. For new methods 1 and 2, no inapplicable cases were observed. This suggests that the sampling errors are smaller by the new methods than by the PBL method. Since in the PBL method nucleotide sites are divided into nondegenerate, twofold degenerate, and fourfold degenerate sites, as mentioned earlier, the sampling errors are large.

Table 11 shows estimates of d_S and d_N obtained by the PBL method and the new methods under the simulation scheme of influenza virus gene mutation, moderate selection, and $L = 1{,}000$. Estimates of d_S by new methods 1 and 2 are good, although $\hat{d}_S$ by new method 1 is slightly (7%) smaller than $E(d_S)$ for $t = 100$. The PBL method underestimates d_S slightly (5–6%) for all cases. Estimates of d_N by the PBL method and new methods 1 and 2 are smaller than $E(d_N)$ for almost all cases. This is probably due to varying substitution rates among nonsynonymous sites, because in the simulation scheme of moderate selection, the fixation probabilities of nonsynonymous changes vary among amino acid pairs interchanged. However, clear linear relationships between t and d_S or d_N were observed. Again, the number of inapplicable cases is large for the PBL method. When $t = 100$, more than one-third of replications are inapplicable. No inapplicable cases were observed for new methods 1 and 2.

Table 12 shows estimates of d_S and d_N obtained by the PBL method and the new methods under the simulation scheme of influenza virus gene mutation, strong selection, and $L = 1{,}000$. New methods 1 and 2 estimate d_S and d_N accurately for all cases. Errors of d_S and d_N are less than 5% for almost all cases. Thus, linearities of d_S and d_N against t are clearly observed. The PBL method underestimates d_S for all cases. The extent of the underestimation is about 10%. On the other hand, estimates of d_N are good. Inapplicable cases were observed for the PBL method. When $t = 100$, more than 10% of replica-

Table 10. Means and standard deviations of $\hat{d}_s$ and $\hat{d}_N$ obtained by the PBL method and new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of influenza virus gene mutation, no selection, and $L = 1{,}000$[a]

	Expectation	PBL	NEW1	NEW2
		$t = 10$		
n	—	100	100	100
d_S	0.100	0.103 ± 0.013	0.106 ± 0.012	0.108 ± 0.014
d_N	0.100	0.100 ± 0.008	0.100 ± 0.008	0.099 ± 0.008
d_N/d_S	1.000	0.978	0.941	0.914
		$t = 20$		
n	—	94	100	100
d_S	0.201	0.203 ± 0.017	0.207 ± 0.016	0.212 ± 0.017
d_N	0.201	0.197 ± 0.011	0.198 ± 0.011	0.197 ± 0.011
d_N/d_S	1.000	0.973	0.958	0.927
		$t = 30$		
n	—	82	100	100
d_S	0.301	0.304 ± 0.025	0.307 ± 0.025	0.315 ± 0.026
d_N	0.301	0.290 ± 0.018	0.294 ± 0.017	0.291 ± 0.016
d_N/d_S	1.000	0.954	0.957	0.924
		$t = 40$		
n	—	73	100	100
d_S	0.402	0.404 ± 0.028	0.403 ± 0.028	0.414 ± 0.031
d_N	0.402	0.386 ± 0.022	0.391 ± 0.022	0.387 ± 0.022
d_N/d_S	1.000	0.957	0.969	0.935
		$t = 50$		
n	—	58	100	100
d_S	0.502	0.507 ± 0.040	0.506 ± 0.037	0.520 ± 0.040
d_N	0.502	0.472 ± 0.025	0.483 ± 0.026	0.478 ± 0.025
d_N/d_S	1.000	0.931	0.955	0.919
		$t = 60$		
n	—	47	100	100
d_S	0.602	0.614 ± 0.045	0.604 ± 0.044	0.622 ± 0.046
d_N	0.602	0.560 ± 0.027	0.579 ± 0.029	0.573 ± 0.029
d_N/d_S	1.000	0.911	0.959	0.921
		$t = 70$		
n	—	36	100	100
d_S	0.703	0.720 ± 0.064	0.687 ± 0.049	0.707 ± 0.054
d_N	0.703	0.641 ± 0.030	0.672 ± 0.040	0.665 ± 0.039
d_N/d_S	1.000	0.890	0.978	0.941
		$t = 80$		
n	—	35	100	100
d_S	0.803	0.806 ± 0.052	0.767 ± 0.051	0.791 ± 0.054
d_N	0.803	0.740 ± 0.033	0.774 ± 0.038	0.765 ± 0.037
d_N/d_S	1.000	0.919	1.009	0.968
		$t = 90$		
n	—	25	100	100
d_S	0.904	0.945 ± 0.096	0.864 ± 0.069	0.895 ± 0.079
d_N	0.904	0.820 ± 0.044	0.861 ± 0.053	0.849 ± 0.052
d_N/d_S	1.000	0.868	0.996	0.949
		$t = 100$		
n	—	19	100	100
d_S	1.004	1.050 ± 0.103	0.943 ± 0.064	0.978 ± 0.072
d_N	1.004	0.875 ± 0.050	0.954 ± 0.061	0.942 ± 0.060
d_N/d_S	1.000	0.833	1.012	0.963

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases. n = number of applicable cases

Table 11. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by the PBL method and new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of influenza virus gene mutation, moderate selection, and $L = 1{,}000$[a]

	Expectation	PBL	NEW1	NEW2
		$t = 10$		
n	—	100	100	100
d_S	0.101	0.097 ± 0.012	0.100 ± 0011	0.104 ± 0.013
d_N	0.053	0.055 ± 0.006	0.053 ± 0.005	0.052 ± 0.005
d_N/d_S	0.529	0.567	0.535	0.504
		$t = 20$		
n	—	100	100	100
d_S	0.201	0.194 ± 0.018	0.200 ± 0.017	0.209 ± 0.019
d_N	0.106	0.106 ± 0.008	0.103 ± 0.007	0.102 ± 0.007
d_N/d_S	0.529	0.546	0.518	0.487
		$t = 30$		
n	—	100	100	100
d_S	0.302	0.287 ± 0.023	0.295 ± 0.023	0.311 ± 0.025
d_N	0.160	0.159 ± 0.010	0.156 ± 0.010	0.153 ± 0.010
d_N/d_S	0.529	0.552	0.527	0.491
		$t = 40$		
n	—	99	100	100
d_S	0.403	0.380 ± 0.031	0.391 ± 0.032	0.411 ± 0.036
d_N	0.213	0.206 ± 0.013	0.203 ± 0.012	0.200 ± 0.012
d_N/d_S	0.529	0.541	0.520	0.486
		$t = 50$		
n	—	92	100	100
d_S	0.503	0.473 ± 0.033	0.483 ± 0.034	0.512 ± 0.038
d_N	0.266	0.254 ± 0.014	0.252 ± 0.014	0.247 ± 0.014
d_N/d_S	0.529	0.536	0.521	0.483
		$t = 60$		
n	—	90	100	100
d_S	0.604	0.568 ± 0.037	0.581 ± 0.038	0.617 ± 0.043
d_N	0.319	0.298 ± 0.015	0.298 ± 0.015	0.293 ± 0.014
d_N/d_S	0.529	0.526	0.513	0.475
		$t = 70$		
n	—	84	100	100
d_S	0.704	0.667 ± 0.046	0.675 ± 0.046	0.721 ± 0.052
d_N	0.372	0.344 ± 0.017	0.346 ± 0.016	0.340 ± 0.016
d_N/d_S	0.529	0.516	0.513	0.471
		$t = 80$		
n	—	82	100	100
d_S	0.805	0.767 ± 0.065	0.779 ± 0.064	0.834 ± 0.074
d_N	0.426	0.386 ± 0.019	0.388 ± 0.019	0.382 ± 0.018
d_N/d_S	0.529	0.503	0.498	0.458
		$t = 90$		
n	—	74	100	100
d_S	0.906	0.854 ± 0.070	0.855 ± 0.066	0.917 ± 0.074
d_N	0.479	0.427 ± 0.026	0.432 ± 0.024	0.425 ± 0.023
d_N/d_S	0.529	0.500	0.506	0.463
		$t = 100$		
n	—	62	100	100
d_S	1.006	0.954 ± 0.092	0.933 ± 0.073	1.006 ± 0.086
d_N	0.532	0.464 ± 0.023	0.474 ± 0.023	0.466 ± 0.022
d_N/d_S	0.529	0.486	0.508	0.463

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases. n = number of applicable cases

Table 12. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by the PBL method and new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of influenza virus gene mutation, strong selection, and $L = 1{,}000$[a]

	Expectation	PBL	NEW1	NEW2
		$t = 10$		
n	—	100	100	100
d_S	0.100	0.093 ± 0.010	0.095 ± 0.010	0.101 ± 0.011
d_N	0.020	0.020 ± 0.003	0.020 ± 0.003	0.020 ± 0.003
d_N/d_S	0.200	0.217	0.213	0.196
		$t = 20$		
n	—	100	100	100
d_S	0.201	0.189 ± 0.015	0.193 ± 0.015	0.206 ± 0.016
d_N	0.040	0.040 ± 0.004	0.040 ± 0.004	0.039 ± 0.004
d_N/d_S	0.200	0.211	0.207	0.190
		$t = 30$		
n	—	100	100	100
d_S	0.301	0.285 ± 0.024	0.295 ± 0.023	0.315 ± 0.026
d_N	0.060	0.061 ± 0.007	0.062 ± 0.007	0.061 ± 0.007
d_N/d_S	0.200	0.215	0.210	0.192
		$t = 40$		
n	—	99	100	100
d_S	0.402	0.381 ± 0.032	0.396 ± 0.034	0.425 ± 0.038
d_N	0.080	0.082 ± 0.006	0.083 ± 0.006	0.082 ± 0.006
d_N/d_S	0.200	0.214	0.211	0.192
		$t = 50$		
n	—	96	100	100
d_S	0.502	0.467 ± 0.032	0.488 ± 0.034	0.525 ± 0.039
d_N	0.101	0.101 ± 0.007	0.103 ± 0.008	0.101 ± 0.007
d_N/d_S	0.200	0.216	0.211	0.192
		$t = 60$		
n	—	98	100	100
d_S	0.602	0.552 ± 0.038	0.579 ± 0.040	0.625 ± 0.045
d_N	0.121	0.122 ± 0.009	0.125 ± 0.008	0.123 ± 0.008
d_N/d_S	0.200	0.222	0.217	0.197
		$t = 70$		
n	—	94	100	100
d_S	0.703	0.641 ± 0.045	0.678 ± 0.052	0.736 ± 0.060
d_N	0.141	0.141 ± 0.010	0.145 ± 0.010	0.143 ± 0.010
d_N/d_S	0.200	0.220	0.215	0.194
		$t = 80$		
n	—	95	100	100
d_S	0.803	0.729 ± 0.052	0.775 ± 0.061	0.846 ± 0.069
d_N	0.161	0.162 ± 0.009	0.167 ± 0.009	0.164 ± 0.009
d_N/d_S	0.200	0.223	0.216	0.194
		$t = 90$		
n	—	93	100	100
d_S	0.904	0.819 ± 0.058	0.871 ± 0.073	0.951 ± 0.084
d_N	0.181	0.182 ± 0.012	0.189 ± 0.012	0.186 ± 0.012
d_N/d_S	0.200	0.223	0.217	0.195
		$t = 100$		
n	—	88	100	100
d_S	1.004	0.892 ± 0.066	0.952 ± 0.081	1.050 ± 0.094
d_N	0.201	0.205 ± 0.012	0.212 ± 0.013	0.208 ± 0.013
d_N/d_S	0.200	0.230	0.223	0.199

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases. n = number of applicable cases

Table 13. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by the PBL method and new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of pseudogene mutation, no selection, and $L = 1{,}000$[a]

	Expectation	PBL	NEW1	NEW2
		$t = 10$		
n	—	100	100	100
d_S	0.102	0.114 ± 0.016	0.102 ± 0.012	0.107 ± 0.013
d_N	0.105	0.101 ± 0.007	0.105 ± 0.007	0.103 ± 0.007
d_N/d_S	1.030	0.890	1.021	0.965
		$t = 20$		
n	—	99	100	100
d_S	0.204	0.227 ± 0.022	0.199 ± 0.017	0.207 ± 0.019
d_N	0.210	0.200 ± 0.013	0.209 ± 0.013	0.206 ± 0.013
d_N/d_S	1,030	0.882	1.049	0.996
		$t = 30$		
n	—	93	100	100
d_S	0.306	0.344 ± 0.029	0.290 ± 0.021	0.300 ± 0.022
d_N	0.316	0.295 ± 0.015	0.311 ± 0.015	0.307 ± 0.015
d_N/d_S	1.030	0.858	1.070	1.024
		$t = 40$		
n	—	82	100	100
d_S	0.409	0.475 ± 0.039	0.392 ± 0.027	0.403 ± 0.029
d_N	0.421	0.384 ± 0.018	0.408 ± 0.018	0.404 ± 0.018
d_N/d_S	1.030	0.809	1.041	1.003
		$t = 50$		
n	—	76	100	100
d_S	0.511	0.595 ± 0.040	0.479 ± 0.028	0.491 ± 0.030
d_N	0.526	0.476 ± 0.021	0.508 ± 0.021	0.503 ± 0.021
d_N/d_S	1.030	0.799	1.060	1.025
		$t = 60$		
n	—	63	100	100
d_S	0.613	0.726 ± 0.053	0.569 ± 0.037	0.580 ± 0.040
d_N	0.631	0.558 ± 0.024	0.603 ± 0.026	0.600 ± 0.026
d_N/d_S	1.030	0.769	1.060	1.034
		$t = 70$		
n	—	64	100	100
d_S	0.715	0.859 ± 0.073	0.647 ± 0.041	0.656 ± 0.046
d_N	0.737	0.646 ± 0.029	0.703 ± 0.031	0.700 ± 0.031
d_N/d_S	1.030	0.752	1.086	1.067
		$t = 80$		
n	—	53	100	100
d_S	0.817	0.997 ± 0.088	0.727 ± 0.046	0.734 ± 0.047
d_N	0.842	0.725 ± 0.029	0.794 ± 0.030	0.791 ± 0.029
d_N/d_S	1.030	0.727	1.092	1.077
		$t = 90$		
n	—	44	100	100
d_S	0.919	1.131 ± 0.113	0.801 ± 0.054	0.804 ± 0.055
d_N	0.947	0.798 ± 0.038	0.884 ± 0.038	0.883 ± 0.038
d_N/d_S	1.030	0.706	1.104	1.098
		$t = 100$		
n	—	47	100	100
d_S	1.022	1.311 ± 0.169	0.872 ± 0.062	0.864 ± 0.058
d_N	1.052	0.866 ± 0.035	0.970 ± 0.042	0.973 ± 0.042
d_N/d_S	1.030	0.660	1.112	1.126

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases. n = number of applicable cases

Table 14. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by the PBL method and new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of pseudogene mutation, moderate selection, and $L = 1{,}000$[a]

	Expectation	PBL	NEW1[b]	NEW2
		$t = 10$		
n[c]	—	99	100	100
d_S	0.101	0.112 ± 0.015	0.098 ± 0.011 (0.106 ± 0.013)	0.106 ± 0.013
d_N	0.054	0.052 ± 0.005	0.054 ± 0.005 (0.053 ± 0.005)	0.053 ± 0.005
d_N/d_S	0.534	0.461	0.549 (0.498)	0.497
		$t = 20$		
n	—	100	100	100
d_S	0.202	0.218 ± 0.023	0.190 ± 0.018 (0.203 ± 0.020)	0.203 ± 0.020
d_N	0.108	0.098 ± 0.008	0.104 ± 0.008 (0.102 ± 0.008)	0.102 ± 0.008
d_N/d_S	0.534	0.453	0.548 (0.501)	0.501
		$t = 30$		
n	—	99	100	100
d_S	0.303	0.325 ± 0.028	0.279 ± 0.020 (0.299 ± 0.023)	0.299 ± 0.023
d_N	0.162	0.149 ± 0.011	0.158 ± 0.011 (0.154 ± 0.011)	0.154 ± 0.011
d_N/d_S	0.534	0.457	0.566 (0.517)	0.517
		$t = 40$		
n	—	96	100	100
d_S	0.404	0.431 ± 0.039	0.364 ± 0.029 (0.388 ± 0.032)	0.388 ± 0.033
d_N	0.215	0.193 ± 0.011	0.206 ± 0.011 (0.202 ± 0.011)	0.202 ± 0.011
d_N/d_S	0.534	0.448	0.568 (0.522)	0.522
		$t = 50$		
n	—	93	100	100
d_S	0.505	0.549 ± 0.042	0.449 ± 0.030 (0.478 ± 0.036)	0.479 ± 0.034
d_N	0.269	0.237 ± 0.014	0.255 ± 0.014 (0.250 ± 0.014)	0.250 ± 0.014
d_N/d_S	0.534	0.431	0.568 (0.523)	0.522
		$t = 60$		
n	—	97	100	100
d_S	0.606	0.651 ± 0.047	0.523 ± 0.032 (0.558 ± 0.039)	0.558 ± 0.037
d_N	0.323	0.281 ± 0.014	0.304 ± 0.014 (0.298 ± 0.014)	0.298 ± 0.014
d_N/d_S	0.534	0.431	0.580 (0.534)	0.534
		$t = 70$		
n	—	95	100	100
d_S	0.707	0.764 ± 0.055	0.603 ± 0.038 (0.641 ± 0.049)	0.641 ± 0.042
d_N	0.377	0.326 ± 0.015	0.355 ± 0.016 (0.349 ± 0.015)	0.349 ± 0.015
d_N/d_S	0.534	0.426	0.590 (0.545)	0.545

Table 14. Continued

	Expectation	PBL	NEW1[b]	NEW2
		$t = 80$		
n	—	89	100	100
d_S	0.808	0.863 ± 0.064	0.671 ± 0.041 (0.708 ± 0.059)	0.710 ± 0.047
d_N	0.431	0.359 ± 0.015	0.395 ± 0.016 (0.390 ± 0.016)	0.389 ± 0.016
d_N/d_S	0.534	0.416	0.589 (0.551)	0.548
		$t = 90$		
n	—	82	100	100
d_S	0.909	0.971 ± 0.073	0.735 ± 0.044 (0.771 ± 0.058)	0.776 ± 0.050
d_N	0.485	0.400 ± 0.018	0.440 ± 0.019 (0.434 ± 0.018)	0.434 ± 0.018
d_N/d_S	0.534	0.412	0.598 (0.564)	0.559
		$t = 100$		
n	—	74	100	100
d_S	1.010	1.087 ± 0.086	0.801 ± 0.055 (0.832 ± 0.073)	0.842 ± 0.058
d_N	0.539	0.434 ± 0.021	0.481 ± 0.022 (0.477 ± 0.021)	0.475 ± 0.021
d_N/d_S	0.534	0.399	0.601 (0.573)	0.565

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases

[b] Values in parentheses were estimated when the $\hat{\alpha}/\hat{\beta}$ ratio was given. The ratio was estimated from the geometric mean of the estimated transition/transversion ratio at synonymous sites

[c] Number of applicable cases

tions are inapplicable. No inapplicable cases were observed for new methods 1 and 2.

Table 13 shows estimates of d_S and d_N obtained by the PBL method and the new methods under the simulation scheme of pseudogene mutation, no selection, and L = 1,000. It is clear that the PBL method overestimates d_S and underestimates d_N for all cases. The extents of the biased estimation are 11–28% for d_S and 4–18% for d_N. The ratio of the mean of d_N to that of d_S is substantially biased (0.660) when $t = 100$. Thus, under the simulation scheme of pseudogene mutation, the PBL method is not suitable for a statistical test for the neutral theory of molecular evolution; this method is favorable for the neutral theory. On the other hand, new methods 1 and 2 give good estimates of d_S and d_N unless the divergence of nucleotide sequences is large. Then $t = 70$, new methods 1 and 2 underestimate d_S by 10% and 8%, respectively. The extents of underestimation of d_N are not so large (5% for new methods 1 and 2). These results indicate that unless the divergence of nucleotide sequences is large, say, $E(d_S) > 0.7$, new methods 1 and 2 have a necessary property of methods for estimating d_S and d_N under the simulation scheme of pseudogene mutation. The number of inapplicable cases is large for the PBL method. When $t = 100$, more than half of replications are inapplicable. No inapplicable cases were observed for new methods 1 and 2.

Table 14 shows estimates of d_S and d_N obtained by the PBL method and the new methods under the simulation scheme of pseudogene mutation, moderate selection, and L = 1,000. We can see that the PBL method overestimates d_S and underestimates d_N for all cases. The extents of the biased estimation are 7–11% for d_S and 4–19% for d_N. Even when $t < 70$, estimates of d_N are 4–13% smaller than $E(d_N)$. New method 2 gives underestimates of d_S and d_N for large values of t. However, when $t < 70$, the extents of the biased estimation are small (less than 8% for d_S and d_N). When $t > 70$, the extents of underestimation of d_S and d_N are 12–18% and 10–12%, respectively. Underestimation of d_N by the PBL method and new method 2 is probably due to varying substitution rates among nonsynonymous sites. New method 1 gives underestimates of d_S even when t is not so large. For example, when $t = 40$, $\hat{d}_S$ is 90% of $E(d_S)$. This is because under this simulation scheme, new method 1 overestimates S, as seen in Table 9, and thus underestimates d_S. On the other hand, although N is underestimated, estimates of d_N are smaller than $E(d_N)$ for all cases but t = 10. The extent of underestimation of d_N becomes large as t increases. This is also probably due to varying substi-

Table 15. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by the PBL method and new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of pseudogene mutation, strong selection, and $L = 1{,}000$[a]

	Expectation	PBL	NEW1[b]	NEW2
		$t = 10$		
n[c]	—	100	100	100
d_S	0.101	0.102 ± 0.014	0.088 ± 0.011 (0.099 ± 0.012)	0.099 ± 0.013
d_N	0.021	0.021 ± 0.003	0.022 ± 0.003 (0.021 ± 0.003)	0.021 ± 0.003
d_N/d_S	0.207	0.201	0.248 (0.212)	0.212
		$t = 20$		
n	—	100	100	100
d_S	0.202	0.200 ± 0.019	0.172 ± 0.014 (0.194 ± 0.016)	0.194 ± 0.017
d_N	0.042	0.041 ± 0.005	0.044 ± 0.005 (0.042 ± 0.005)	0.042 ± 0.005
d_N/d_S	0.207	0.206	0.256 (0.219)	0.219
		$t = 30$		
n	—	99	100	100
d_S	0.304	0.314 ± 0.028	0.266 ± 0.021 (0.302 ± 0.027)	0.301 ± 0.025
d_N	0.063	0.061 ± 0.005	0.066 ± 0.006 (0.063 ± 0.005)	0.063 ± 0.005
d_N/d_S	0.207	0.195	0.248 (0.210)	0.211
		$t = 40$		
n	—	99	100	100
d_S	0.405	0.408 ± 0.031	0.346 ± 0.022 (0.394 ± 0.028)	0.392 ± 0.027
d_N	0.084	0.082 ± 0.006	0.089 ± 0.007 (0.086 ± 0.006)	0.086 ± 0.006
d_N/d_S	0.207	0.201	0.257 (0.217)	0.219
		$t = 50$		
n	—	97	100	100
d_S	0.506	0.519 ± 0.041	0.436 ± 0.030 (0.498 ± 0.040)	0.494 ± 0.037
d_N	0.105	0.102 ± 0.008	0.111 ± 0.008 (0.107 ± 0.008)	0.108 ± 0.008
d_N/d_S	0.207	0.196	0.255 (0.215)	0.218
		$t = 60$		
n	—	96	100	100
d_S	0.607	0.614 ± 0.042	0.512 ± 0.029 (0.587 ± 0.042)	0.582 ± 0.034
d_N	0.126	0.123 ± 0.008	0.135 ± 0.008 (0.130 ± 0.008)	0.131 ± 0.008
d_N/d_S	0.207	0.200	0.263 (0.222)	0.224
		$t = 70$		
n	—	96	100	100
d_S	0.708	0.708 ± 0.058	0.582 ± 0.042 (0.667 ± 0.054)	0.661 ± 0.049
d_N	0.147	0.141 ± 0.009	0.156 ± 0.009 (0.151 ± 0.009)	0.151 ± 0.009
d_N/d_S	0.207	0.200	0.268 (0.226)	0.229

Table 15. Continued

	Expectation	PBL	NEW1[b]	NEW2
		$t = 80$		
n	—	91	100	100
d_S	0.810	0.791 ± 0.059	0.649 ± 0.043 (0.740 ± 0.064)	0.732 ± 0.052
d_N	0.168	0.165 ± 0.011	0.182 ± 0.012 (0.177 ± 0.011)	0.177 ± 0.011
d_N/d_S	0.207	0.208	0.280 (0.239)	0.242
		$t = 90$		
n	—	96	100	100
d_S	0.911	0.880 ± 0.070	0.715 ± 0.050 (0.816 ± 0.075)	0.805 ± 0.060
d_N	0.189	0.186 ± 0.011	0.206 ± 0.012 (0.200 ± 0.011)	0.201 ± 0.011
d_N/d_S	0.207	0.212	0.288 (0.246)	0.249
		$t = 100$		
n	—	95	100	100
d_S	1.012	0.983 ± 0.080	0.790 ± 0.049 (0.894 ± 0.076)	0.884 ± 0.058
d_N	0.210	0.206 ± 0.011	0.230 ± 0.012 (0.224 ± 0.012)	0.224 ± 0.012
d_N/d_S	0.207	0.210	0.291 (0.250)	0.254

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases
[b] Values in parentheses were estimated when the $\hat{\alpha}/\hat{\beta}$ ratio was given. The ratio was estimated from the geometric mean of the estimated transition/transversion ratio at synonymous sites
[c] Number of applicable cases

tution rates among nonsynonymous sites. Inapplicable cases were observed only for the PBL method. When t = 100, about one-fourth of replications are inapplicable.

Table 15 shows estimates of d_S and d_N obtained by the PBL method and the new methods under the simulation scheme of pseudogene mutation, strong selection, and L = 1,000. The PBL method gives good estimates of d_S and d_N for all cases. New method 2 estimates d_S and d_N accurately when $t < 70$. For large values of t, underestimation of d_S is apparent. However, estimates of d_N are not so far from $E(d_N)$. Underestimation of d_S by new method 1 is substantial. Even when $t = 10$, the extent of underestimation of d_S is 13%. This is due to overestimation of S caused by inflation of the transition/transversion ratio at the third position of codons, as seen earlier. On the other hand, this method overestimates d_N. Inapplicable cases were observed for the PBL method but not for new methods 1 and 2.

Underestimation of d_S by new methods 1 and 2 for large values of t results from the nucleotide-frequency bias under the pseudogene mutation scheme. These methods use Kimura's two-parameter method to correct multiple substitutions. Kimura's two-parameter method does not take into account the unequal equilibrium frequencies of the four nucleotides.

Under the simulation scheme of mitochondrial gene mutation, when the divergence of nucleotide sequences are not small, say, $E(d_S) > 0.2$, the PBL method and new methods 1 and 2 underestimated d_S. For example, under the simulation scheme of strong selection, L = 1,000, and t = 50 [$E(d_S)$ = 0.500], estimates of d_S by the PBL method and new methods 1 and 2 were 0.389, 0.388, and 0.399, respectively. This underestimation of d_S for large values of t is caused by the strong nucleotide-frequency bias under the mitochondrial gene mutation scheme. The PBL method and new methods 1 and 2 use Kimura's two-parameter formula to correct multiple substitutions. This formula underestimates the number of nucleotide substitutions when there is a strong nucleotide-frequency bias at equilibrium. On the other hand, underestimation of d_N was not so serious when the divergence of nucleotide sequences was small, say, $E(d_N) < 0.2$. Again, inapplicable cases were observed only for the PBL method.

These results show that under the simulation scheme of influenza virus gene mutation, new methods 1 and 2 give good estimates of d_S and d_N. Under the simulation scheme of pseudogene, estimates obtained by new method 2 are good when $t < 70$. However, under this simulation scheme, new method 1 underestimates d_S in the presence of negative selection even when t is not large. The PBL method gives good estimates of d_S and d_N under the simulation scheme of influenza virus gene

mutation and no selection or pseudogene mutation and strong selection. However, under the simulation scheme of influenza virus gene mutation and moderate selection or strong selection, the PBL method underestimates d_S. Under the simulation scheme of pseudogene mutation and no selection or moderate selection, this method overestimates d_S. Under the simulation scheme of mitochondrial gene mutation, the PBL method and new methods 1 and 2 underestimate d_S when t is not small. Inapplicable cases were observed only for the PBL method. The total number of inapplicable cases for this method was 1,084 (12% of replications). This indicates that the sampling errors of d_S and d_N are larger by the PBL method than by new methods 1 and 2.

Similar results for L = 1,000 were obtained for L = 290. However, stochastic fluctuations due to the small number of codons compared are larger for L = 290 than for L = 1,000. Thus, the standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were larger for L = 290 than for L = 1,000. Furthermore, inapplicable cases were observed not only for the PBL method but also for new methods 1 and 2. For new methods 1 and 2, the numbers of inapplicable cases were as follows: under the simulation scheme of influenza virus gene mutation and no selection, 1 for both methods when t = 10; under the simulation scheme of influenza virus gene mutation and moderate selection, 2 for both when t = 10; under the simulation scheme of influenza virus gene mutation and strong selection, 2 for new method 1 and 3 for new method 2 when t = 10 and 1 for new method 2 when t = 100. The total numbers of inapplicable cases for new methods 1 and 2 were 5 (0.06% of replications) and 6 (0.07% of replications). All of these cases but one were observed when t = 10. Thus, the causes for these inapplicable cases are that when the divergence of nucleotide sequences compared is small, new methods 1 and 2 cannot estimate the numbers of synonymous and nonsynonymous sites; if the divergence of nucleotide sequences compared is small, new methods 1 and 2 cannot estimate the α/β ratio. On the other hand, for the PBL method, inapplicable cases were observed more frequently for large values of t than for small ones, as seen for L = 1,000. This is because arguments of logarithms in estimation formulae become negative more frequently for large values of t than for small ones. The total number of inapplicable cases was 419 (4.7% of replications). This value is smaller than that (1,084) for L = 1,000. This result is counterintuitive. Since stochastic fluctuations due to the small number of codons compared are larger for L = 290 than for L = 1,000, it is expected that the number of inapplicable cases is larger for L = 290 than for L = 1,000. This discrepancy between intuition and the above result may lie in a characteristic of the PBL method. However, it is also possible that this discrepancy is caused by a bug in the program that was used in this study. Anyway, in the case of L = 290, the number of inapplicable cases is much (about 85 times) larger for the PBL method than for new methods 1 and 2.

Effect of the Small Number of Codons Compared

Table 16 shows estimates of d_S and d_N obtained by the PBL method and the new methods under the simulation scheme of influenza virus gene mutation, no selection, and L = 50. The PBL method gives good estimates of d_S and d_N for many cases, as seen for L = 1,000. However, when t = 50 or 60, overestimation of d_S is apparent. As t increases, the number of inapplicable cases increases. On the other hand, new methods 1 and 2 give overestimates of d_S when t is small. In particular, overestimation of d_S by new method 2 is substantial when t = 10 or 20. Furthermore, it is clear that when t was small, inapplicable cases occurred frequently. These two phenomena (overestimation of d_S and inapplicable cases) are highly correlated. The α/β ratio is estimated from the transition/transversion ratio at the third position of codons (new method 1) or at synonymous sites (new method 2). Most nucleotide changes at the third position of codons are synonymous. All nucleotide changes at synonymous sites are synonymous. Thus, estimation of the α/β ratio (applicable or inapplicable) is highly correlated with estimation of d_S. As mentioned earlier, when the divergence of nucleotide sequences compared is small, new methods 1 and 2 cannot estimate the numbers of synonymous and nonsynonymous sites. This is the case particularly when nucleotide sequences compared are short. Thus, inapplicable cases occurred more frequently for L = 50 than for L = 1,000 or 290 when t is small. It is probable that in inapplicable cases, the proportion of $d_S < E(d_S)$ is large and that in applicable cases, the proportion of $d_S > E(d_S)$ is large. Thus, when t is small, new methods 1 and 2 overestimate d_S. On the other hand, estimation of d_N is not correlated with inapplicable cases (estimation of the α/β ratio). Thus, it is probable the proportion of $d_N > E(d_N)$ is the same for both applicable and inapplicable cases. Estimates of d_N are, therefore, not biased even when t is small. When t = 50 or 60, overestimation of d_S for new methods 1 and 2 is also apparent. Under the other simulation schemes, similar results were obtained.

The total number of inapplicable cases for the PBL method was 545 (6.1% of replications). This value is larger than that for L = 290 but much smaller than that for L = 1,000. The numbers of inapplicable cases for new methods 1 and 2 were 790 (8.8% of replications) and 1,127 (12.5% of replications), respectively. Most of these inapplicable cases occurred when $t < 40$. The numbers of inapplicable cases for $t < 40$ were 583 and 809 for new methods 1 and 2, respectively. These values are 73.8% and 71.8% of the total numbers of inapplicable cases for new methods 1 and 2, respectively. As mentioned earlier, this is because new methods 1 and 2 cannot estimate the numbers of synonymous and nonsynonymous sites when the divergence of short nucleotide sequences is small.

When we analyze short sequences or small parts of

Table 16. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by the PBL method and new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of influenza virus gene mutation, no selection, and $L = 50$[a]

	Expectation	PBL	NEW1	NEW2
		$t = 10$		
n	—	100	69	54
d_S	0.100	0.108 ± 0.053	0.112 ± 0.044	0.131 ± 0.045
d_N	0.100	0.102 ± 0.030	0.102 ± 0.030	0.101 ± 0.030
d_N/d_S	1.000	0.942	0.913	0.765
		$t = 20$		
n	—	100	86	59
d_S	0.201	0.205 ± 0.091	0.220 ± 0.090	0.242 ± 0.096
d_N	0.201	0.209 ± 0.058	0.208 ± 0.053	0.213 ± 0.056
d_N/d_S	1.000	1.019	0.946	0.878
		$t = 30$		
n	—	100	96	80
d_S	0.301	0.306 ± 0.126	0.303 ± 0.114	0.323 ± 0.118
d_N	0.301	0.318 ± 0.083	0.320 ± 0.080	0.320 ± 0.081
d_N/d_S	1.000	1.038	1.054	0.990
		$t = 40$		
n	—	98	97	91
d_S	0.402	0.424 ± 0.140	0.415 ± 0.129	0.433 ± 0.142
d_N	0.402	0.395 ± 0.090	0.404 ± 0.085	0.406 ± 0.089
d_N/d_S	1.000	0.930	0.972	0.938
		$t = 50$		
n	—	98	99	98
d_S	0.502	0.542 ± 0.190	0.517 ± 0.167	0.536 ± 0.182
d_N	0.502	0.477 ± 0.113	0.485 ± 0.111	0.482 ± 0.112
d_N/d_S	1.000	0.880	0.938	0.899
		$t = 60$		
n	—	89	98	96
d_S	0.602	0.675 ± 0.218	0.658 ± 0.232	0.687 ± 0.262
d_N	0.602	0.610 ± 0.145	0.619 ± 0.143	0.612 ± 0.145
d_N/d_S	1.000	0.903	0.941	0.892
		$t = 70$		
n	—	86	99	98
d_S	0.703	0.742 ± 0.246	0.754 ± 0.320	0.767 ± 0.304
d_N	0.703	0.679 ± 0.176	0.711 ± 0.186	0.708 ± 0.186
d_N/d_S	1.000	0.915	0.943	0.923
		$t = 80$		
n	—	80	96	96
d_S	0.803	0.855 ± 0.322	0.782 ± 0.273	0.830 ± 0.313
d_N	0.803	0.804 ± 0.270	0.817 ± 0.252	0.812 ± 0.265
d_N/d_S	1.000	0.940	1.045	0.978
		$t = 90$		
n	—	76	96	94
d_S	0.904	0.917 ± 0.282	0.834 ± 0.314	0.834 ± 0.249
d_N	0.904	0.901 ± 0.239	0.916 ± 0.243	0.910 ± 0.237
d_N/d_S	1.000	0.982	1.098	1.091
		$t = 100$		
n	—	61	89	92
d_S	1.004	0.999 ± 0.343	0.944 ± 0.328	1.002 ± 0.359
d_N	1.004	0.960 ± 0.285	1.019 ± 0.456	0.972 ± 0.356
d_N/d_S	1.000	0.961	1.079	0.971

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases. n = number of applicable cases

sequences, other longer sequences or the remaining larger parts of the sequences are often available. For example, when Hughes and Nei (1988) analyzed antigen recognition sites of MHC class I genes, the remaining parts of the genes consisted of 216 or 217 codons. Thus, I examined such a case where a particular region of genes is small but the remaining larger region of the genes is available.

Table 17 shows estimates of d_S and d_N obtained by the new methods under the simulation scheme of influenza virus gene mutation, no selection, and $L = 50$, when the $\hat{\alpha}/\hat{\beta}$ ratios were given. The α/β ratios were estimated by new methods 1 and 2 under the corresponding simulation scheme for $L = 290$. The geometric means of $\hat{\alpha}/\hat{\beta}$ for 100 pairs of nucleotide sequences were used to estimate d_S and d_N by new methods 1 and 2. The geometric mean of $\hat{\alpha}/\hat{\beta}$ rapidly converged to a value as the number of pairs of nucleotide sequences used increased. For example, when $t = 10$, the geometric means for new method 1 were 7.849, 8.421, 8.767, and 8.776 for n = 10, 20, 30, and 100, where n is the number of pairs of nucleotide sequences. The geometric means for new method 2 were 5.875, 6.439, 6.865, and 7.048 for n = 10, 20, 30, and 100. Note that these values are close to the true value of α/β (7.72).

It is clear that inapplicable cases are much less frequent in Table 17 than in Table 16. In particular, no inapplicable cases are observed for $t < 60$ in Table 17. As a result, estimates of d_S are closer to $E(d_S)$ in Table 17 than in Table 16, when t is small. When $t > 30$, estimates of d_S are essentially the same for Tables 16 and 17. However, new method 1 for t = 70 in Table 17 is an exceptional case. When t = 70, new method 1 overestimates d_S substantially in Table 17. In replication 85, synonymous substitutions occurred much more than the expectation. In this replication, the total number of synonymous differences was estimated to be 24.3 and $\hat{d}_S$ was computed to be 3.394. Thus, the mean and standard deviations of $\hat{d}_S$ were inflated for new method 1. On the other hand, replication 85 was not included for new method 1 in Table 16 nor for new method 2 in Tables 16 and 17, because these methods were not applicable to replication 85. Estimates of d_N are essentially the same for Tables 16 and 17, although estimates of d_N are slightly better in Table 16 than in Table 17 when $t = 90$ and 100. When the $\hat{\alpha}/\hat{\beta}$ ratio is given, the accuracies of new methods 1 and 2 are almost the same for that of the PBL method. When $t = 50$ or 60, overestimation of d_S is again observed in Table 17. Under the other simulation scheme, if the $\hat{\alpha}/\hat{\beta}$ ratio was given, inapplicable cases did not occur at all when $t < 50$. As a result, the accuracies of new methods 1 and 2 were almost the same for that of the PBL method.

The total numbers of inapplicable cases were 130 (1.4% of replications) and 190 (2.1% of replications) for new methods 1 and 2, respectively, when the $\hat{\alpha}/\hat{\beta}$ ratios

Table 17. Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by new methods 1 (NEW1) and 2 (NEW2) under the simulation scheme of influenza virus gene mutation, no selection, and $L = 50$, where the $\hat{\alpha}/\hat{\beta}$ ratios were given[a]

	Expectation	NEW1	NEW2
		$t = 10$	
n	—	100	100
d_S	0.100	0.105 ± 0.049	0.107 ± 0.051
d_N	0.100	0.101 ± 0.030	0.100 ± 0.030
d_N/d_S	1.000	0.960	0.933
		$t = 20$	
n	—	100	100
d_S	0.201	0.211 ± 0.099	0.215 ± 0.101
d_N	0.201	0.207 ± 0.057	0.206 ± 0.057
d_N/d_S	1.000	0.981	0.956
		$t = 30$	
n	—	100	100
d_S	0.301	0.297 ± 0.122	0.305 ± 0.126
d_N	0.301	0.319 ± 0.083	0.316 ± 0.082
d_N/d_S	1.000	1.074	1.037
		$t = 40$	
n	—	100	100
d_S	0.402	0.416 ± 0.141	0.426 ± 0.147
d_N	0.402	0.401 ± 0.088	0.397 ± 0.087
d_N/d_S	1.000	0.964	0.932
		$t = 50$	
n	—	100	100
d_S	0.502	0.518 ± 0.186	0.531 ± 0.194
d_N	0.502	0.484 ± 0.109	0.480 ± 0.107
d_N/d_S	1.000	0.934	0.903
		$t = 60$	
n	—	98	97
d_S	0.602	0.684 ± 0.346	0.689 ± 0.286
d_N	0.602	0.621 ± 0.146	0.615 ± 0.143
d_N/d_S	1.000	0.909	0.892
		$t = 70$	
n	—	99	94
d_S	0.703	0.849 ± 0.559	0.783 ± 0.377
d_N	0.703	0.710 ± 0.181	0.708 ± 0.179
d_N/d_S	1.000	0.836	0.904
		$t = 80$	
n	—	96	96
d_S	0.803	0.799 ± 0.348	0.848 ± 0.439
d_N	0.803	0.830 ± 0.271	0.817 ± 0.255
d_N/d_S	1.000	1.039	0.963
		$t = 90$	
n	—	94	94
d_S	0.904	0.829 ± 0.326	0.867 ± 0.398
d_N	0.904	0.930 ± 0.249	0.951 ± 0.319
d_N/d_S	1.000	1.121	1.098
		$t = 100$	
n	—	87	84
d_S	1.004	0.951 ± 0.430	0.937 ± 0.360
d_N	1.004	0.977 ± 0.321	0.964 ± 0.303
d_N/d_S	1.000	1.027	1.029

[a] Means and standard deviations of $\hat{d}_S$ and $\hat{d}_N$ were calculated by excluding inapplicable cases. n = number of applicable cases

were given. These values are much smaller than not only those in the case where the $\hat{\alpha}/\hat{\beta}$ ratio was not given but also than those for the PBL method.

Tajima (1993) has shown that when the number of sites compared is small, ordinary algorithms for estimating the number of substitutions per site (e.g., Jukes and Cantor's method, Kimura's two-parameter method) give biased estimates. When the expected number of substitutions per site is 0.5, the number of substitutions between short sequences tends to be overestimated by ordinary algorithms. For example, he showed that when the number of sites compared was 20, the numbers of nucleotide substitution per site were estimated to be 0.5342 and 0.5510 by Jukes and Cantor's method and Kimura's two-parameter method, respectively. Tajima's results probably apply to estimation of d_S and d_N. Actually, under any simulation schemes of mutation and selection, by not only the PBL method and the new methods but also the NG, MY, and LWL methods, estimates of d_S and d_N for $L = 50$ were larger than those for $L = 1{,}000$ or 290 when $t = 50$, 60, or 70. For example, under the simulation scheme of influenza virus gene mutation, no selection, $L = 50$, and $t = 60$, estimates of d_S by the NG, MY, and LWL methods were 0.893, 1.049, and 0.986, respectively. Under this simulation scheme, estimates of d_N were 0.481, 0.469, and 0.529 by the NG, MY, and LWL methods, respectively. All of these estimates of d_S and d_N are larger than the corresponding values for $L = 1{,}000$ (Table 10). However, the extents of overestimation are smaller for d_N than for d_S. This tendency was observed under all simulation schemes of mutation and selection. This is because the number of nonsynonymous sites is larger than that of synonymous sites. If we apply Tajima's algorithm to formulae for estimating d_S and d_N, biases observed for $t = 50$, 60, and 70 may be eliminated.

Standard Errors of $\hat{d}_S$ and $\hat{d}_N$

To examine the accuracies of the approximate formulae for $V(\hat{d}_S)$ and $V(\hat{d}_N)$, I conducted computer simulations. Kimura's two-parameter model ($\alpha/\beta = 10$) was used as a mutation matrix. The number of replications was 100 for each set of simulation schemes. Table 18 shows the results for $L = 290$ codons. In this table standard errors of $\hat{d}_S$ and $\hat{d}_N$ given by the approximate formulae are close to observed standard deviations of $\hat{d}_S$ and $\hat{d}_N$. In the case of $L = 50$ codons, the same results were obtained. Thus, the approximate formulae give good estimates of $V(\hat{d}_S)$ and $V(\hat{d}_N)$.

These results indicate that although new methods 1 and 2 estimate not only the numbers of synonymous and nonsynonymous differences but also the α/β ratio, the sampling errors of $\hat{d}_S$ and $\hat{d}_N$ by these methods are not so large as expected. This is because there is an interaction between estimation of differences and estimation of the

Table 18. Standard errors and observed standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by new methods 1 (NEW1) and 2 (NEW2)

	NEW1		NEW2	
	Observed[a]	Equation[b]	Observed[c]	Equation
	No selection			
	$t = 10$			
d_S	0.021	0.021	0.022	0.021
d_N	0.015	0.014	0.015	0.014
	$t = 30$			
d_S	0.040	0.044	0.042	0.045
d_N	0.027	0.029	0.028	0.029
	$t = 50$			
d_S	0.061	0.075	0.064	0.075
d_N	0.047	0.047	0.047	0.047
	$t = 100$			
d_S	0.155	0.220	0.171	0.239
d_N	0.120	0.126	0.119	0.125
	Moderate selection			
	$t = 10$			
d_S	0.020	0.021	0.021	0.021
d_N	0.009	0.009	0.009	0.009
	$t = 30$			
d_S	0.039	0.045	0.041	0.046
d_N	0.017	0.018	0.017	0.018
	$t = 50$			
d_S	0.066	0.074	0.068	0.077
d_N	0.027	0.025	0.027	0.025
	$t = 100$			
d_S	0.181	0.236	0.199	0.275
d_N	0.041	0.041	0.041	0.041
	Strong selection			
	$t = 10$			
d_S	0.022	0.021	0.023	0.021
d_N	0.006	0.006	0.006	0.006
	$t = 30$			
d_S	0.046	0.044	0.048	0.045
d_N	0.011	0.011	0.011	0.011
	$t = 50$			
d_S	0.046	0.044	0.048	0.045
d_N	0.011	0.011	0.011	0.011
	$t = 100$			
d_S	0.233	0.307	0.244	0.336
d_N	0.020	0.022	0.020	0.022

[a] Observed standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by new method 1 (NEW1)

[b] $\sqrt{\sum_{i=1}^{n} V(\hat{d}_i)/n}$, where n is the number of applicable cases and $V(\hat{d}_i)$ is the variance of $\hat{d}_S$ or $\hat{d}_N$ estimated by the approximate formulae in the i-th replication

[c] Observed standard deviations of $\hat{d}_S$ and $\hat{d}_N$ obtained by new method 2 (NEW2)

α/β ratio. For example, suppose that synonymous transitional substitutions happened to occur more frequently than their expectation. In this case, the number of synonymous transitional differences (S_{Ts}) becomes larger than the expectation. However, this does not necessarily lead to an increase in the proportion of synonymous transitional differences (P_S). This is because in the present case, the $\hat{\alpha}/\hat{\beta}$ ratio also becomes larger than its expectation. As a result, the number of synonymous sites (S) is estimated to be larger than its expectation. Thus, the increase of P_S ($=S_{Ts}/S$) does not become so large, although P_S probably becomes larger than its expectation. On the other hand, since the number of synonymous sites becomes larger than its expectation, the proportion of synonymous transversional differences [Q_S ($= S_{Tv}/S$)] becomes smaller than its expectation unless synonymous transversional substitutions also happened to occur more frequently than their expectation. Thus, the increase of $\hat{d}_S$ does not become so large as expected for Kimura's ordinary two-parameter method. This explanation applies to not only new method 2 but also new method 1, because most of synonymous substitutions occur at the third position of codons. However, the sampling errors of $\hat{d}_S$ and $\hat{d}_N$ by new method 2 are larger than those by new method 1 because estimation errors of the codon frequencies are also involved in new method 2.

Examination of $\hat{d}_S$ and $\hat{d}_N$

When we apply a method to analysis of nucleotide sequences, it is of particular importance that from observable quantities we can examine whether an estimate obtained by the method is good or not. Most of synonymous substitutions occur at the third position of codons. Thus, I estimated the number of nucleotide substitutions per site at the third position of codons (d_3) and examined relationships of $\hat{d}_S$ and $\hat{d}_3$.

Table 19 shows estimates of d_3 obtained by Jukes and Cantor's (JC) method, Kimura's two-parameter (K2P) method, Kimura's three-substitution-type (K3ST) method (1981), Takahata and Kimura's (TK) method (1981), Tajima and Nei's (TN) method (1984), Gojobori et al.'s (GIN) method (1982a), Tamura's three-parameter (T3P) method (1992), and Tamura and Nei's (TmN) method (1993) under the simulation scheme of pseudogene mutation, no selection, and L = 1,000. The TK, TN, GIN, T3P, and TmN methods consider deviation of the equilibrium nucleotide frequencies from equality (0.25), whereas the JC, K2P, and K3ST methods do not. Thus, it is expected that when the divergence of nucleotide sequences is large, the JC, K2P, and K3ST methods give underestimates of d_3 under the simulation scheme of pseudogene mutation. As expected, the difference between estimates of d_3 obtained by the JC, K2P, and K3ST methods and the TK, TN, GIN, T3P, and TmN methods becomes larger as t increases. However, since the model on which the T3P method is based is too specific, the method is probably not suitable for the present simulation scheme. When t = 70, $\hat{d}_3$ obtained by the K2P method is 93%, 94%, 93%, and 94% of $\hat{d}_3$ obtained by the TK, TN, GIN, and TmN methods, respectively. For the same value of t, estimates of d_S obtained by new methods 1 and 2 are 90% and 92% of E(d_S), respectively. For the other values of t, the extent of underestimation of d_S by new methods 1 and 2 agrees with the ratios of $\hat{d}_3$ obtained by the K2P method to $\hat{d}_3$ obtained by the TK, TN, GIN, and TmN methods. This relationship was also observed under the other selection schemes. For example, under the simulation scheme of moderate selection and t = 70, $\hat{d}_S$ obtained by new method 2 was 91% of E(d_S) and the ratios of $\hat{d}_3$ obtained by the K2P method to $\hat{d}_3$ obtained by the TK, TN, GIN, and TmN methods were 94%, 94%, 93%, and 94%, respectively. Under the simulation scheme of strong selection and t = 70, $\hat{d}_S$ obtained by new method 2 was 93% of E(d_S) and the ratios of $\hat{d}_3$ obtained by the K2P method to $\hat{d}_3$ obtained by the TK, TN, GIN, and TmN methods were 95%, 93%, 93%, and 94%, respectively. In the cases of the moderate selection and strong selection schemes, I did not consider $\hat{d}_S$ obtained by new method 1, because underestimation of d_S by this method is not mainly caused by the unequal frequencies of the four nucleotides. As seen earlier, the major cause for underestimation of d_S by this method is overestimation of S due to inflation of the α_3/β_3 ratio.

Under the simulation scheme of influenza virus gene mutation, serious underestimation of d_S by new methods 1 and 2 was not observed (Tables 10–12). As expected, estimates of d_3 obtained by the K2P method were almost the same as those obtained by the TK, TN, T3P, and TmN methods. However, estimates of d_3 obtained by the GIN method were much lower than those obtained by the K2P, TK, TN, T3P, and TmN methods, and were between those obtained by these methods and the JC method. For example, under the simulation scheme of no selection, L = 1,000, and t = 100 [E(d_3) = 1.022], estimates of d_3 obtained by the K2P, TK, TN, T3P, TmN, GIN, and JC methods were 0.991, 0.999, 1.030, 1.001, 1.037, 0.914, and 0.824, respectively. This tendency was observed for all cases under the simulation scheme of influenza virus gene mutation. The model on which the GIN method is based was proposed by Kimura (1981) to take into account only the unequal frequencies of the four nucleotides. Under the influenza virus gene mutation scheme, a transition/transversion bias is strong but a nucleotide-frequency bias is not strong, as seen in Tables 3 and 4. This is why the GIN method gives underestimates of d_3 under the simulation scheme of influenza virus gene mutation.

Under the simulation scheme of mitochondrial gene mutation, when the divergence of nucleotide sequences

Table 19. Means and standard deviations of $\hat{d}_3$ obtained by the JC, K2P, K3ST, TK, TN, GIN, T3P, and TmN methods under the simulation scheme of pseudogene mutation, no selection, and $L = 1{,}000$

	Expectation	JC	K2P	K3ST	TK	TN	GIN	T3P	TmN
					$t = 10$				
n[a]	—	100	100	100	100	100	100	100	100
d_3	0.103	0.104 ± 0.012	0.105 ± 0.012	0.105 ± 0.012	0.106 ± 0.013	0.106 ± 0.012	0.106 ± 0.012	0.105 ± 0.012	0.105 ± 0.012
					$t = 20$				
n	—	100	100	100	100	100	100	100	100
d_3	0.207	0.203 ± 0.015	0.205 ± 0.016	0.205 ± 0.016	0.210 ± 0.016	0.210 ± 0.016	0.210 ± 0.016	0.207 ± 0.016	0.208 ± 0.016
					$t = 30$				
n	—	100	100	100	100	100	100	100	100
d_3	0.310	0.298 ± 0.018	0.301 ± 0.019	0.301 ± 0.019	0.310 ± 0.020	0.310 ± 0.019	0.310 ± 0.019	0.305 ± 0.019	0.307 ± 0.019
					$t = 40$				
n	—	100	100	100	100	100	100	100	100
d_3	0.414	0.399 ± 0.026	0.404 ± 0.027	0.405 ± 0.027	0.421 ± 0.029	0.421 ± 0.029	0.423 ± 0.030	0.413 ± 0.028	0.417 ± 0.028
					$t = 50$				
n	—	100	100	100	100	100	100	100	100
d_3	0.517	0.491 ± 0.028	0.499 ± 0.029	0.499 ± 0.029	0.523 ± 0.035	0.523 ± 0.032	0.525 ± 0.034	0.512 ± 0.031	0.519 ± 0.032
					$t = 60$				
n	—	100	100	100	100	100	100	100	100
d_3	0.621	0.578 ± 0.033	0.589 ± 0.035	0.589 ± 0.035	0.624 ± 0.042	0.623 ± 0.039	0.628 ± 0.041	0.608 ± 0.038	0.619 ± 0.040
					$t = 70$				
n	—	100	100	100	100	100	100	100	100
d_3	0.724	0.663 ± 0.036	0.676 ± 0.039	0.676 ± 0.039	0.723 ± 0.047	0.720 ± 0.044	0.730 ± 0.048	0.702 ± 0.043	0.718 ± 0.045
					$t = 80$				
n	—	100	100	100	100	100	100	100	100
d_3	0.828	0.750 ± 0.043	0.766 ± 0.045	0.767 ± 0.046	0.832 ± 0.058	0.822 ± 0.052	0.838 ± 0.059	0.802 ± 0.050	0.824 ± 0.056
					$t = 90$				
n	—	100	100	100	100	100	100	100	100
d_3	0.931	0.824 ± 0.049	0.842 ± 0.052	0.843 ± 0.053	0.925 ± 0.067	0.912 ± 0.060	0.939 ± 0.073	0.889 ± 0.060	0.918 ± 0.066
					$t = 100$				
n	—	100	100	100	100	100	100	100	100
d_3	0.1035	0.901 ± 0.056	0.920 ± 0.061	0.922 ± 0.061	1.024 ± 0.083	1.005 ± 0.071	1.043 ± 0.089	0.978 ± 0.071	1.019 ± 0.081

[a] Number of applicable cases

is small, new methods 1 and 2 give good estimates of d_S, as mentioned earlier. In such a case, the K2P method gave essentially the same value of $\hat{d}_3$ as the TK, TN, GIN, T3P, and TmN methods did. For example, under the simulation scheme of mitochondrial gene mutation, strong selection, $L = 1{,}000$, and $t = 10$ [$E(d_3) = 0.092$], estimates of d_3 by the K2P, TK, TN, GIN, T3P, and TmN methods were 0.090, 0.090, 0.092, 0.096, 0.090, and 0.094, respectively. On the other hand, new methods 1 and 2 underestimate d_S for large values of t. In such a case, the difference among estimates of d_3 obtained by the K2P, TK, TN, GIN, T3P, and TmN methods were also clear. When $t = 50$ [$E(d_3) = 0.460$], estimates of d_3 by the K2P, TK, TN, GIN, T3P, and TmN methods were 0.356, 0.360, 0.407, 0.467, 0.360, and 0.478, respectively. In this case, not only the K2P method but also the TK, TN, and T3P methods gave underestimates of d_3. Under the other selection schemes, underestimation of d_3 by the K2P, TK, TN, and T3P methods were observed for large values of t.

These results indicate that from the ratios of $\hat{d}_3$ obtained by the K2P method to $\hat{d}_3$ obtained by the TK, TN, GIN, and TmN methods, we can examine whether estimates of d_S obtained by the new methods are good or not. To make such an examination of $\hat{d}_S$, we have to know the properties of methods for estimating the number of nucleotide substitutions. Since some authors (e.g., Gojobori et al. 1982a, 1990) generalized results obtained under a particular simulation scheme, caution should be taken. Similarly, from the numbers of nucleotide substitutions at the 1st, 2nd, and 1st + 2nd positions of codons estimated by the K2P, TK, TN, GIN, and TmN methods, we can examine whether estimates of d_N obtained by the new methods are good or not.

In addition, I found that even if new methods 1 and 2 gave biased estimates of d_S, the estimated transition/

transversion ratio at synonymous sites was close to the true value. For example, under the simulation scheme of pseudogene mutation, strong selection, and $L = 1{,}000$, new method 1 underestimates d_S substantially, as seen in Table 15. However, under this simulation scheme, the geometric mean of the transition/transversion ratio estimated by new method 1 was computed to be 2.577 when $t = 10$. This value is close to the true value (Table 3). Probably the numbers of transitional and transversional substitutions were underestimated to the same extent.

These findings suggest that from the estimated transition/transversion ratio at synonymous sites and the estimated frequencies of twofold (+3-fold) and fourfold degenerate codons, we can estimate the α/β ratio, as we have evaluated the extent of overestimation of the number of synonymous sites. We neglect synonymous substitutions at the first position of codons because the proportion of synonymous substitutions at this position is much smaller than that of synonymous substitutions at the third position of codons. We again treat threefold degenerate codons as twofold degenerate codons, as in new method 2. At the third position of twofold (+3-fold) degenerate codons, only transitional substitutions are synonymous. At the third position of fourfold degenerate codons, both transitional and transversion substitutions are synonymous. Thus, from $\alpha_S/\beta_S \approx (q_2\alpha + q_3\alpha + q_4\alpha)/(q_4\beta)$, we have $\alpha/\beta \approx q_4/(q_2 + q_3 + q_4)\alpha_S/\beta_S$. Noting that α_S/β_S is twice the transition/transversion ratio at synonymous sites, we can estimate the α/β ratio from the estimated transition/transversion ratio at synonymous sites and the estimated frequencies of twofold (+3-fold) and fourfold degenerate codons. Unless the number of nucleotide sequences compared is small and the number of codons in the sequences is small, the estimates of q_2, q_3, and q_4 do not deviate much from the true values. Note also that the geometric mean of the estimated transition/transversion ratio rapidly converges to a value, as shown earlier.

By the above procedure, I evaluated estimates of S, N, d_S, and d_N under the simulation scheme of pseudogene mutation, strong selection or moderate selection, and $L = 1{,}000$. The results obtained are shown in parentheses in Tables 9, 14, and 15. As shown in Table 9, this procedure gives better estimates of S and N than new method 1. Furthermore, we can see from Tables 14 and 15 that this procedure gives much better estimates of d_S and d_N than new method 1.

This procedure is similar to estimation of the α/β ratio by new method 2. The major difference between this procedure and new method 2 is that in this procedure, the α/β ratio is estimated from the geometric mean of the estimated transition/transversion ratio at synonymous sites for all possible pairs of nucleotide sequences compared, whereas in new method 2, the α/β ratio is estimated from only a pair of nucleotide sequences compared. Since the frequencies of twofold (+3-fold) and fourfold degenerate codons deviate more in a pair of nucleotide sequences than in all available sequences, estimation errors of the α/β ratio are probably larger by new method 2 than by this procedure. Since this procedure is not restricted to new method 1, it is also applicable to new method 2.

On the other hand, it is difficult to make an examination of estimates of d_S obtained by the PBL method. This is because the properties of $\hat{d}_S$ obtained by this method are not so simple as those by new methods 1 and 2. Although the PBL method uses Kimura's two-parameter method to correct multiple substitutions, the properties of $\hat{d}_S$ obtained by the PBL method do not always agree with those of $\hat{d}_3$ obtained by the K2P method. For example, as seen earlier, under the simulation scheme of pseudogene mutation, no selection, $L = 1{,}000$, and $t = 70$, $\hat{d}_3$ obtained by the K2P method is 93%, 94%, 93%, and 94% of $\hat{d}_3$ obtained by the TK, TN, GIN, and TmN methods, respectively. On the other hand, $\hat{d}_S$ obtained by the PBL method is 120% of $E(d_S)$, as seen in Table 13. For all the other cases but the simulation scheme of influenza virus gene mutation and no selection, the properties of $\hat{d}_S$ obtained by the PBL method disagreed with those of $\hat{d}_3$ obtained by the K2P method. Furthermore, no clear-cut relationships between $\hat{d}_S$ and mutation scheme or selection scheme were observed. Under the simulation scheme of influenza virus gene mutation and no selection, $\hat{d}_S$ obtained by the PBL method agrees with $E(d_S)$ (Table 10). However, in the presence of negative selection, the PBL method underestimates d_S (Tables 11 and 12). Under the simulation scheme of pseudogene mutation and no selection or moderate selection, the PBL method overestimates d_S (Tables 13 and 14). However, under the simulation scheme of pseudogene mutation and strong selection, this method gives good estimates of d_S (Table 15). Thus, from observable quantities, we cannot examine whether estimates obtained by the PBL method are good or not.

Discussion

We have seen that even when the divergence of nucleotide sequences is small, commonly used methods such as the MY, LWL, and NG methods give overestimates of d_S and underestimates of d_N. Thus, these methods may not be able to reject the neutral mutation hypothesis when positive selection operates. This result calls for reexamination of some genes because evolutionary pictures of genes have often been discussed on the basis of results obtained by the MY, LWL, and NG methods. The extents of biased estimation of d_S and d_N depend on the simulation schemes of mutation and selection. Thus, we cannot correct biased estimates of d_S and d_N by certain methods, e.g., multiplying $\hat{d}_S$ and $\hat{d}_N$ by certain factors. Nevertheless, estimates of d_S and d_N obtained by the MY, LWL, and NG methods can be used as evolutionary distances for reconstruction of phylogenetic trees by dis-

tance matrix methods such as the UPGMA (Sneath and Sokal 1973) and the neighbor-joining method (Saitou and Nei 1987) because these estimates are roughly proportional to t.

The PBL method and new methods 1 and 2 give better estimates of d_S and d_N than the MY, LWL, and NG methods unless there are strong transition/transversion and nucleotide-frequency biases like mitochondrial genes. In addition, the PBL method [equations (27) and (28)] and new methods 1 and 2 estimate S and N more accurately than the MY, LWL, and NG methods. For analysis of actual nucleotide sequences, however, new methods 1 and 2 are preferable to the PBL method for the following reasons: (1) Inapplicable cases occur much more frequently for the PBL method than for new methods 1 and 2 unless the number of codons compared is small. This indicates that the sampling variances of estimates of d_S and d_N are larger by the PBL method than by new methods 1 and 2. When the number of codons compared is small, estimates of d_S and d_N by new methods 1 and 2 are as good as those by the PBL method if the $\hat{\alpha}/\hat{\beta}$ ratio is given. In this case, inapplicable cases occur less frequently for new methods 1 and 2 than for the PBL method. Recent studies (Schöniger and von Haeseler 1993; Tajima and Takezaki 1994) have shown that when phylogenetic trees are reconstructed by distance matrix methods, the proportion of obtaining the true topology is higher for the use of evolutionary distances with smaller sampling variances than for the use of evolutionary distances with larger sampling variances. Thus, it is probable that the true topology is obtained with a higher probability from estimates of d_S or d_N obtained by new methods 1 and 2 than from those obtained by the PBL method. (2) We can reduce estimation errors of the α/β ratio, using the geometric mean of the estimated transition/transversion ratio at synonymous sites and the frequencies of twofold (+threefold) and fourfold degenerate codons in all available nucleotide sequences compared. This procedure probably improves the accuracies of new methods 1 and 2 and reduces the sampling variances of estimates by these methods, particularly when the number of codons compared is small. Moreover, to obtain further better estimates of d_S and d_N, we can use an even better estimate of the α/β ratio obtained by a certain method. For example, we can estimate tha α/β ratio from noncoding regions such as introns and flanking regions of genes analyzed. The estimate of the α/β ratio is probably better than that obtained from coding regions because noncoding regions are larger than coding regions. On the other hand, we cannot improve the accuracy of the PBL method or reduce the sampling variances of estimates by this method, even if the number of nucleotide sequences compared increases or noncoding regions of genes analyzed are available. (3) From estimates of d_3 by the K2P, TK, TN, GIN, and TmN methods, we can examine whether estimates of d_S obtained by new methods 1 and 2 are good or not. On the other hand, we cannot make an examination of estimates of d_S obtained by the PBL method.

Recently, Kondo et al. (1993) proposed a method for estimating d_S, which is based on Hasegawa et al.'s (1985) model. Since the model considers different rates of transition and transversion and the frequencies of the four nucleotides, their method is probably applicable to the case where there are strong transition/transversion and nucleotide-frequency biases. The method is also applicable to the case where there are not such strong biases. However, the method is time-consuming because it requires computer simulations of 1,000 replications to estimate d_S. Furthermore, the method cannot estimate d_N, because we do not know the fixation probabilities of nonsynonymous changes for given nucleotide sequences, which are necessary in order to conduct computer simulations. On the other hand, unless there are strong transition/transversion and nucleotide-frequency biases, new methods 1 and 2 give reasonably good estimates of d_S and d_N. In addition, new methods 1 and 2 are time-saving. Thus, I recommend that unless there are strong transition/transversion and nucleotide-frequency biases, we use new methods 1 and 2 to estimate d_S and d_N. Only when there are strong transition/transversion and nucleotide-frequency biases do we have to use Kondo et al.'s time-consuming method to estimate d_S. In such a case, new methods 1 and 2 are recommended to estimate d_N if $E(d_N)$ is probably small—say, <0.2. In this case, the $\hat{\alpha}/\hat{\beta}$ ratio should be given to new methods 1 and 2.

We have seen that under the simulation scheme of moderate selection, all of the PBL method and new methods 1 and 2 underestimate d_N. As mentioned earlier, this is probably because the substitution rates vary among nonsynonymous sites. In the case where there is variation of the substitution rates among sites, several studies (Nei and Gojobori 1986; Jin and Nei 1990) showed that the gamma distance method gives better estimates than ordinary methods assuming the uniform substitution rate among sites. For new methods 1 and 2, we can estimate the gamma distance of d_N by Jin and Nei's (1990) equation (A4), replacing $\hat{P}$ and $\hat{Q}$ with $\hat{P}_N$ and $\hat{Q}_N$, respectively. Furthermore, we can obtain the sampling variance of $\hat{d}_N$ by Jin and Nei's equations (A5), (A6), and (A7), replacing n, $\hat{P}$, and $\hat{Q}$ with $\hat{N}$, $\hat{P}_N$, and $\hat{Q}_N$, respectively. However caution should be taken because their formula for the sampling variances of gamma distances neglects estimation errors of the gamma parameter a. Taking into account these estimation errors, the sampling variances of gamma distances are larger than those given by Jin and Nei's equations.

Nei and Gojobori (1986) found that the gamma distance of d_N ($a = 1$) by the NG method agreed with the "expectation" of d_N under the simulation scheme of pseudogene mutation and moderate selection. It is likely that for an appropriate value of a, the gamma distance method described above improves the goodness-of-fit of $\hat{d}_N$ by new methods 1 and 2 under the simulation scheme

of moderate selection. Actually, the gamma distance of d_N ($a = 2$) by the new methods improved the goodness-of-fit of $\hat{d}_N$ under the simulation scheme of moderate selection. For example, under the simulation scheme of influenza virus gene mutation, moderate selection, and $L = 1{,}000$, the means of $\hat{d}_N$ obtained by the gamma version of new method 2 were 0.054, 0.106, 0.163, 0.217, 0.274, 0.331, 0.392, 0.448, 0.508, and 0.568 for $t = 10$, 20, 30, 40, 50, 60, 70, 80, 90, and 100, respectively. In practice, it is difficult to estimate the gamma parameter a for nonsynonymous sites in actual nucleotide sequences. This is because unlike Tamura and Nei's (1993) analysis of the mitochondrial DNA control region, the number of nonsynonymous sites is usually small and furthermore restrictive for a set of given nucleotide sequences; we cannot make sequences longer to reduce estimation errors of the gamma parameter a for coding regions of genes analyzed.

Miyata et al. (1987a,b, 1990) have analyzed 35 autosome-linked genes between human and mouse or between human and rat and estimated the d_S values by the MY method. Miyata et al.'s (1990) Table 21-1 shows that the estimates of d_S for conservative-type and divergent-type genes are 0.55 and 0.78, respectively, on the average, and that the overall mean of $\hat{d}_S$ is 0.66. However, since the MY method overestimates d_S, the true values of d_S are probably lower to some extent than the above values obtained by Miyata et al. Actually, Ohta's (1993) Fig. 2 shows that for 17 single-copy nuclear genes, the means of $\hat{d}_S$ obtained by new method 1 are 0.59, 0.53, and 0.33, between rodentia and artiodactyla, between rodentia and primates, and between artiodactyla and primates, respectively. The figure also shows that if the PBL method is used, the corresponding values are 0.64, 0.57, and 0.34, respectively. [Appendix 1 in Wolfe and Sharp (1993) shows that most of the 17 genes analyzed by Ohta represent faster genes.] This result indicates that unless the nucleotide frequencies are strongly biased—say, SD > 0.11—the new methods can be applied in order to analyze most single-copy nuclear genes among mammals.

Finally, I would like to emphasize that even the PBL method and the new methods developed in this study depend on simple assumptions and give only approximate estimates of d_S and d_N. Thus, further studies on methods for estimating d_S and d_N are required. In particular, it is important to develop estimation methods which consider a nucleotide-frequency bias, because Kondo et al.'s (1993) method give only estimates of d_S but not estimates of d_N.

Computer Programs

Computer programs for estimating d_S, d_N, gamma distance of d_N, and their standard errors by new methods 1 and 2 are available on request (E-mail address: yina@ddbj.nig.ac.jp).

Acknowledgments. I thank Drs. T. Ohta and F. Tajima for their helpful suggestions and comments during the course of this study. I am also grateful to Dr. T. Ohta, who provided me with computing facilities. I thank an anonymous reviewer for his comments, which improved the presentation of this paper. Thanks are also due to Dr. S. Horai, whose data on $\hat{d}_S$ and $\hat{d}_N$ for primate mitochondrial genes led me to reevaluation of the MY, LWL, and NG methods.

References

Anderson S, Bankier AT, Barrell BG, De Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Shreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Easteal S (1985) Generation time and the rate of molecular evolution. Mol Biol Evol 2:450–453

Easteal S (1990) The pattern of mammalian evolution and the relative rate of molecular evolution. Genetics 124:165–173

Gojobori T, Ishii K, Nei M (1982a) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J Mol Evol 18:414–423

Gojobori T, Li W-H, Graur D (1982b) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Gojobori T, Moriyama EN, Kimura M (1990) Statistical methods for estimating sequence divergence. In: Doolittle RF (ed) Methods in enzymology 183: Molecular evolution: computer analysis of protein and nucleic acid sequences. Academic Press, New York, pp 531–550

Gojobori T (1983) Codon substitution in evolution and the "saturation" of synonymous changes. Genetics 105:1011–1027

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hayasaka K, Gojobori T, Horai S (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. Mol Biol Evol 5:626–644

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, Takahata N (1992) Man's place in hominoidea revealed by mitochondrial DNA genealogy. J Mol Evol 35:32–43

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. Nature 335:167–170

Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7:82–102

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Kikuno R, Hayashida H, Miyata T (1985) Rapid rate of rodent evolution. Proc Jpn Acad 61B:153–155

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

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

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kondo R. Horai S, Satta Y, Takahata N (1993) Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. J Mol Evol 36:517–531

Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Li W-H, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330–342

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li W-H, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93–96

Miyata T, Miyazawa S, Yasunaga T (1979) Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219–236

Miyata T, Yasunaga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol 16:23–36

Miyata T, Hayashida H, Kuma K, Yasunaga T (1987a) Male-driven molecular evolution demonstrated by different rates of silent substitutions between autosome- and sex chromosome-linked genes. Proc Jpn Acad 63B:327–331

Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T (1987b) Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harbor Symp Quanti Biol 52:863–867

Miyata T, Kuma K, Iwabe N, Hayashida H, Yasunaga T (1990) Different rates of evolution of autosome-, X chromosome- and Y chromosome-linked genes: hypothesis of male-driven molecular evolution. In: Takahata N, Crow JF (eds) Population biology of genes and molecules. Baifukan Press, Tokyo, pp 341–357

Moriyama EN (1987) Higher rates of nucleotide substitution in Drosophila than in mammals. Jpn J Genet 62:139–147

Moriyama EN, Gojobori T (1992) Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics 130:855–864

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286

Ohta T (1993) An examination of generation-time effect on molecular evolution. Proc Natl Acad Sci USA 90:10676–10680

Pamilo P, Bianchi NO (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281

Saitou N (1987) Patterns of nucleotide substitutions in influenza A virus genes. Jpn J Genet 62:439–443

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Schöniger M, von Haeseler A (1993) A simple method to improve the reliability of tree reconstructions. Mol Biol Evol 10:471–483

Sharp PM, Li W-H (1989) On the rate of DNA sequence evolution in *Drosophila*. J Mol Evol 28:398–402

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

Tajima F (1993) Unbiased estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 10:677–688

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269–285

Tajima F, Takezaki N (1994) Estimation of evolutionary distance for reconstructing molecular phylogenetic tree. Mol Biol Evol 11:278–286

Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 98:641–657

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

Tamura K, Nei M (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

Tanaka T, Nei M (1989) Positive Darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol Evol 6:447–459

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Wolfe KH, Sharp PM (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J Mol Evol 37:441–456

Wu C-I, Li W-H (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745