Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models

Ziheng Yang and Rasmus Nielsen†*

*Department of Biology, University College London, England; and †Department of Organismic and Evolutionary Biology, Harvard University

Approximate methods for estimating the numbers of synonymous and nonsynonymous substitutions between two DNA sequences involve three steps: counting of synonymous and nonsynonymous sites in the two sequences, counting of synonymous and nonsynonymous differences between the two sequences, and correcting for multiple substitutions at the same site. We examine complexities involved in those steps and propose a new approximate method that takes into account two major features of DNA sequence evolution: transition/transversion rate bias and base/codon frequency bias. We compare the new method with maximum likelihood, as well as several other approximate methods, by examining infinitely long sequences, performing computer simulations, and analyzing a real data set. The results suggest that when there are transition/transversion rate biases and base/codon frequency biases, previously described approximate methods for estimating the nonsynonymous/synonymous rate ratio may involve serious biases, and the bias can be both positive and negative. The new method is, in general, superior to earlier approximate methods and may be useful for analyzing large data sets, although maximum likelihood appears to always be the method of choice.

Introduction

Estimation of synonymous and nonsynonymous substitution rates is important in understanding the dynamics of molecular sequence evolution (Kimura 1983; Gillespie 1991; Ohta 1995). As synonymous (silent) mutations are largely invisible to natural selection (but see Akashi 1995), while nonsynonymous (amino- acid-replacing) mutations may be under strong selective pressure, comparison of the rates of fixation of those two types of mutations provides a powerful tool for understanding the mechanisms of DNA sequence evolution. For example, variable nonsynonymous/synonymous rate ratios among lineages may indicate adaptive evolution (Messier and Stewart 1997) or relaxed selective constraints along certain lineages (Crandall and Hillis 1997). Likewise, models of variable nonsynonymous/synonymous rate ratios among sites may provide important insights into functional constraints at different amino acid sites and may be used to detect sites under positive selection (Nielsen and Yang 1998).

The simplest problem in this regard is estimation of the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site between two sequences. In the past two decades, a number of intuitive methods have been suggested for this estimation. They involve ad hoc treatments that cannot be justified rigorously, and they will be referred to here as approximate methods. In common, they involve three steps. First, the numbers of synonymous (S) and nonsynonymous (N) sites in the sequences are counted. Second, the numbers of synonymous and nonsynonymous differences between the two sequences are counted. Third, a correction for multiple substitutions at the same site is applied to calculate the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site between the two sequences. Here, we use the notation of Nei and Gojobori (1986, referred to later as ''NG''), which appears to be the most commonly used approximate method; definitions of symbols are given in table 1. The reader is referred to Ina (1995, 1996) for a recent discussion of important concepts. While the above strategy appears simple, well-known features of DNA sequence evolution, such as unequal transition and transversion rates and unequal nucleotide or codon frequencies, make it a real challenge to count sites and differences correctly.

Miyata and Yasunaga (1980) and Perler et al. (1980) were the pioneers in this endeavor and developed the basic concepts (see also the simulation study of Gojobori 1983). The nucleotide substitution (mutation) model underlying the method of Miyata and Yasunaga (1980) and its simplified version (Nei and Gojobori 1986) is the JC69 model (Jukes and Cantor 1969). JC69 is also the underlying mutation model in the method of Li, Wu, and Luo (1985), although the method uses the two-parameter model of Kimura (1980) to correct for multiple hits at the same site. Here, we follow Ina (1995) and use the word ''mutation'' to refer to DNA-level processes before the operation of natural selection at the protein level, although synonymous mutations may also be under selective constraints (Akashi 1995). As transitions are more likely to be synonymous at third positions than are transversions, use of the JC69 model to count sites tends to underestimate the number of synonymous sites and overestimate the number of nonsynonymous sites. The method of Li, Wu, and Luo (1985) has been improved by Li (1993), Pamilo and Bianchi (1993), and Comeron (1995) to correct for this bias in counting sites. Ina (1995) appears to be the first to fully account for the transition/transversion bias in all steps of the estimation. We note that the underlying mutation (substitution) models used in all of the above methods are no more realistic than that of Kimura (1980). Moriyama and Powell (1997) made a useful attempt to cor-

Key words: synonymous rate, nonsynonymous rate, approximate methods, maximum likelihood, molecular evolution, adaptive evolution, positive selection.

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

Mol. Biol. Evol. 17(1):32–43. 2000

Table 1
Definitions of Major Symbols

Symbol	Definition
S	Number of synonymous sites in a sequence
N	Number of nonsynonymous sites in a sequence
L_c	Number of codons in the sequence ($=(S + N)/3$)
d_S	Number of synonymous substitutions per synonymous site
d_N	Number of nonsynonymous substitutions per nonsynonymous site
t	Time (distance), measured by the expected number of nucleotide substitutions per codon, ($=d_S \times 3S/(S + N) + d_N \times 3N/(S + N)$)
ω	Nonsynonymous/synonymous rate ratio ($=d_N/d_S$)
κ	Transition/transversion (mutation) rate ratio
π_j	Equilibrium frequency of codon j
ρ_S^1	Proportion of synonymous sites ($=S/(S + N)$)
ρ_N^1	Proportion of nonsynonymous sites ($=1 - \rho_S^1 = N/(S + N)$)
ρ_S^*	Proportion of synonymous substitutions ($=Sd_S/(Sd_S + Nd_N)$)
ρ_N^*	Proportion of nonsynonymous substitutions ($=1 - \rho_S^* = Nd_N/(Sd_S + Nd_N)$)

rect for biased base frequencies. Their modification, however, is limited to multiple-hit correction, and base frequency bias is not considered in the important steps of counting sites and differences.

A maximum-likelihood (ML) method for estimating d_S and d_N between two sequences was developed by Goldman and Yang (1994) based on an explicit model of codon substitution. The ML method does not involve ad hoc approximations. Furthermore, the ML method is flexible in that knowledge of the substitution process such as transition/transversion bias, codon usage biases, and even chemical differences between amino acids can easily be incorporated into the model.

Relying on insights gained through previous methods, particularly ML estimation, we propose in this paper an approximate method for estimating d_S and d_N that accounts for two major features of DNA sequence evolution: the transition/transversion bias and the base (codon) frequency bias. We examine the similarities and differences among the ML method, the approximate method of this paper, and two other approximate methods by a consistency analysis of infinitely long sequences and computer simulation of finite data. A real data set is also analyzed using different estimation methods.

Methods for Estimating Synonymous and Nonsynonymous Rates

Maximum-Likelihood Estimation

First, we describe the ML method of Goldman and Yang (1994) to introduce the notation and to provide justification for certain steps in our approximate method. An explicit model of codon substitution is required by the ML method. The model we consider in this paper is a simplified version of the model of Goldman and Yang (1994). The substitution rate from any sense codons i to j ($i \neq j$) is given by

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases} \quad (1)$$

Parameter κ is the (mutational) transition/transversion rate ratio, and $\omega = d_N/d_S$ is the nonsynonymous/synonymous rate ratio. The equilibrium codon frequencies (π_j) are calculated using the nucleotide frequencies at the three codon positions; that is, codon frequencies are proportional to the products of nucleotide frequencies at the three codon positions. This approach was found to produce results similar to those obtained using all codon frequencies as free parameters, although codon frequencies are often quite different from those expected from nucleotide frequencies at codon positions (e.g., Goldman and Yang 1994; Pedersen, Wiuf, and Christiansen 1998; Yang and Nielsen 1998). We note that different assumptions about codon frequencies can be easily incorporated into the ML method and the approximate method of this paper. Equation (1) is similar to the simulation model of Gojobori (1983) and the likelihood model of Muse and Gaut (1994), although those authors did not consider the transition/transversion rate bias or different base frequencies at the three codon positions.

The diagonal elements of the rate matrix, $Q = \{q_{ij}\}$, are determined by the mathematical requirement (e.g., Grimmett and Stirzaker 1992, p. 241) that the row sums are zero:

$$\sum_j q_{ij} = 0, \quad \text{for any } i. \quad (2)$$

Because time and rate are confounded, we multiply the rate matrix by a scaling factor so that the expected number of nucleotide substitutions per codon is one:

$$-\sum_i \pi_i q_{ii} = \sum_i \pi_i \sum_{j \neq i} q_{ij} = 1. \quad (3)$$

This scaling means that time t (or, equivalently, branch length or sequence distance) is measured by the expected number of (nucleotide) substitutions per codon.

The transition probability matrix is calculated as

$$P(t) = \{p_{ij}(t)\} = e^{Qt}, \quad (4)$$

where $p_{ij}(t)$ is the probability that codon i will become codon j after time t. The probability of observing a codon site with codons i and j in the two sequences is then

$$f_{ij}(t) = \pi_i p_{ij}(t). \quad (5)$$

The log-likelihood function is given by the multinomial probability

$$\ell(t, \kappa, \omega) = \sum_{i,j} n_{ij} \log\{f_{ij}(t)\} = \sum_{i,j} n_{ij} \log\{\pi_i p_{ij}(t)\}, \quad (6)$$

where n_{ij} is the number of sites occupied by codons i and j in the two sequences. The codon frequencies π_i

are estimated using the observed nucleotide frequencies at the three codon positions in the two sequences. Parameters t, κ, and ω are estimated by maximizing the likelihood function numerically, and the estimates are used to calculate d_S and d_N. Specifically, the proportions of synonymous and nonsynonymous substitutions are given as

$$\rho_S^* = \sum_{\substack{i \neq j \\ aa_i = aa_j}} \pi_i q_{ij} \tag{7}$$

and $\rho_N^* = 1 - \rho_S^*$, respectively. The summation is taken over all codon pairs i and j ($i \neq j$) that code for the same amino acid, and aa_i is the amino acid encoded by codon i. The numbers of synonymous and nonsynonymous substitutions per codon are then $t\rho_S^*$ and $t\rho_N^*$, respectively. The proportions of synonymous and nonsynonymous sites are defined as the proportions of synonymous and nonsynonymous ''mutations'' before the operation of natural selection at the amino acid level (Goldman and Yang 1994; Ina 1995). These can be calculated in a manner similar to equation (7) as ρ_S^1 and ρ_N^1 (equivalent to ρ_S^∞ and ρ_N^∞ in Goldman and Yang 1994), using the ML estimate of κ but with $\omega = 1$ fixed. The numbers of synonymous and nonsynonymous sites per codon are $3\rho_S^1$ and $3\rho_N^1$, respectively. The numbers of synonymous and nonsynonymous substitutions per site are then $d_S = t\rho_S^*/(3\rho_S^1)$ and $d_N = t\rho_N^*/(3\rho_N^1)$, respectively (see table 1).

A New Approximate Method

We suggest an approximate method for estimating d_S and d_N using the strategy adopted by previous authors: counting sites, counting differences, and correcting for multiple hits. In all three steps, we take into account the transition/transversion rate bias and the base (codon) frequency bias. Approximately, our method is based on the HKY85 nucleotide mutation (substitution) model (Hasegawa, Kishino, and Yano 1985). Although this is not the most general model available, it accounts for two most important features of the mutation process, that is, the transition/transversion bias and unequal base frequencies. Previous results (e.g., Yang 1994*a*) suggest that adding further complication is often unnecessary.

Estimating the Transition/Transversion Rate Ratio (κ)

We use the fourfold-degenerate sites at the third codon positions and the nondegenerate sites to estimate κ. Mutations at the fourfold-degenerate sites do not change the amino acid, and thus the transition/transversion rate bias at those sites should reflect the mutational bias. Mutations at nondegenerate sites all lead to amino acid changes and can also be used to estimate κ (see eq. 1). Here, we assume that different nonsynonymous substitutions have the same rate irrespective of the pair of amino acids involved, although the assumption is unrealistic (Yang, Nielsen, and Hasegawa 1998). We calculate an average of κ, weighted by the numbers of nucleotide sites in the two site classes. Since no simple formula is available for estimating κ under the HKY85 model, we use the formula for the F84 model (Yang 1994*b*) instead, relying on the similarity of the two models. We calculate

$$\begin{aligned} A = {} & \{2(\pi_T\pi_C + \pi_A\pi_G) + 2(\pi_T\pi_C\pi_R/\pi_Y + \pi_A\pi_G\pi_Y/\pi_R) \\ & \times [1 - V/(2\pi_Y\pi_R)] - T\} \\ & \div [2(\pi_T\pi_C/\pi_Y + \pi_A\pi_G/\pi_R)], \\ B = {} & 1 - \frac{V}{2\pi_Y\pi_R}, \end{aligned} \tag{8}$$

where T and V are proportions of transitional and transversional differences, respectively, and $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. Then,

$$a = -\log\{A\}, \qquad b = -\log\{B\}, \tag{9}$$

and

$$\begin{aligned} \kappa_{F84} = {} & a/b - 1, \\ t = {} & [4\pi_T\pi_C(1 + \kappa_{F84}/\pi_Y) + 4\pi_A\pi_G(1 + \kappa_{F84}/\pi_R) \\ & + 4\pi_Y\pi_R] \times b \end{aligned} \tag{10}$$

(Yang 1994*b*). The estimated κ_{F84} is then transformed to κ (i.e., κ_{HKY85}) using the following formula (see Goldman 1993)

$$\kappa_{HKY85} = 1 + \frac{(\pi_T\pi_C/\pi_Y + \pi_A\pi_G/\pi_R)\kappa_{F84}}{\pi_T\pi_C + \pi_A\pi_G}. \tag{11}$$

For data of multiple sequences, we suggest estimating a common κ by averaging estimates from all pairwise comparisons and using the combined estimate of κ in the calculation of pairwise d_N and d_S rates.

Counting Synonymous and Nonsynonymous Sites

Ina's (1995) table 1 for counting synonymous and nonsynonymous sites in each codon is correct for mutation models more general than that of Kimura (1980), although Ina's table involves minor errors for codons that can change to stop codons in one step. In general, synonymous and nonsynonymous sites can be counted as in the ML method discussed above for any codon-substitution model (Goldman and Yang 1994). We count sites using codons in the two compared sequences, rather than the equilibrium codon frequencies expected from the model (see discussion in Yang and Nielsen 1998). As there should be about 4% loss of sites due to mutations to stop codons, this scaling means that we are slightly underestimating d_S and d_N, although the ω ratio is not affected (Yang and Nielsen 1998). The numbers of sites (S and N) are scaled so that $S + N = 3L_c$, where L_c is the number of codons. Nucleotide frequencies at synonymous and nonsynonymous sites are recorded and used later for multiple-hit corrections.

Counting Synonymous and Nonsynonymous Differences

Observed nucleotide differences between the two sequences are classified into four categories: synonymous transitions, synonymous transversions, nonsynonymous transitions, and nonsynonymous transversions. When the two compared codons differ at one position,

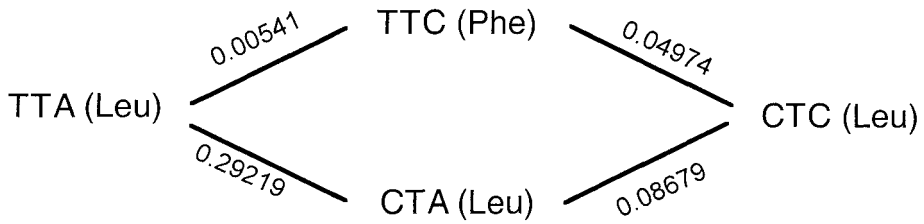

FIG. 1.—Two parsimonious pathways between codons TTA and CTC. Probabilities are calculated using parameter estimates for the human and orangutan mitochondrial genes.

the classification is obvious. When they differ at two or three positions, there will be two or six parsimonious pathways along which one codon could change into the other, and all of them should be considered. Since different pathways may involve different numbers of synonymous and nonsynonymous changes, they should be weighted differently. Miyata and Yasunaga (1980) and Li, Wu, and Luo (1985) made valuable attempts to weight pathways. They also pointed out that equal weighting of pathways, which is used in later methods such as those of Nei and Gojobori (1986) and Ina (1995), biases estimates of ω toward 1; that is, equal weighting tends to overestimate ω when $\omega < 1$ and to underestimate ω when $\omega > 1$.

The most appropriate weights are the relative probabilities of pathways, which we will use in the new approximate method. The probabilities depend on the parameters being estimated: the sequence divergence level (t), the transition/transversion rate ratio (κ), and the d_N/d_S ratio (ω). For given values of t, κ, and ω, it is easy to construct the rate matrix Q and calculate the transition probability matrix $P(t)$ (see eqs. 1 and 4). We use the Taylor expansion in this case for its fast speed.

$$P(t) = e^{Qt} = I + Qt + \frac{1}{2!}(Qt)^2 + \frac{1}{3!}(Qt)^3 + \cdots. \quad (12)$$

The number of terms used is determined by a preset accuracy level. The weight, that is, the probability for each pathway, is calculated as the product of the probabilities of all changes involved in the pathway. Pathways involving stop codons are given weight 0. If all pathways involve stop codons (for example, between AAG and TGG in the mammalian mitochondrial code), ad hoc decisions have to be made.

An example is given in figure 1 using a pair of codons in the mitochondrial genes of the human and the orangutan. The concatenated sequence of 12 genes on the H-strand has 3,331 codons (see below). Between the two sequences, 1,198 codons are different at one position, 151 are different at two positions, and 21 are different at all three positions. The 329th codon is TTA in the human and CTC in the orangutan. Transition probabilities for changes involved in each of the two pathways are given in figure 1, calculated using the estimates obtained by the new method ($t = 0.873$, $\kappa = 10.61$, and $\omega = 0.057$). The probability for the first path is then $p_{TTA,TTC}(t) \times p_{TTC,CTC}(t) = 0.00541 \times 0.04974 = 0.00027$, while that for the second path is $0.29219 \times 0.08679 = 0.02536$. Weights for the two pathways are thus 0.011 and 0.989, and there are 0.022 nonsynonymous and 1.978 synonymous differences between the two codons. Since the nonsynonymous rate is much lower than the synonymous rate ($\omega < 1$), the first path is much less likely. Equal weighting of pathways would give one synonymous and one nonsynonymous difference between the two codons. Other codon pairs may not be so extreme, as the different pathways may involve the same numbers of different types of changes.

Correcting for Multiple Hits Using Estimated Numbers of Sites and Differences

We use the distance formula (eq. 10) for the F84 model (Tateno, Takezaki, and Nei 1994; Yang 1994*b*) to correct for multiple substitutions at the same site to calculate d_S and d_N. For each of the synonymous and nonsynonymous site classes, the proportions of transitional and transversional differences are calculated separately to give T and V for use in equation (8). Different base frequencies are also used for the two site classes. The correction formula used here, as well as those used in all previous approximate methods, is ad hoc. The formula is derived from a Markov process of nucleotide substitution with four states, where each nucleotide can change into one of three other nucleotides. When the formula is used for synonymous (or nonsynonymous) sites only, this basic assumption of the Markov model is violated (Lewontin 1989). However, a proper treatment of the evolutionary process at synonymous and nonsynonymous sites appears to require the use of a codon substitution model, as in the likelihood method, and an analytical derivation of a correction formula based on such a model seems intractable.

The Algorithm

Our method for estimating d_S and d_N can be summarized in the following iterative algorithm.

1. Estimate κ from the fourfold-degenerate sites and the nondegenerate sites under the HKY85-F84 model using base (codon) frequencies from the real data. The estimated κ is used in later steps.
2. Count the numbers of synonymous and nonsynonymous sites (S and N, respectively) using the estimated κ and the observed base (codon) frequencies.
3. Choose starting values for t and ω (e.g., using estimates from the NG method).
4. Count the numbers of synonymous and nonsynonymous differences (both transitions and transversions)

using κ, the codon frequencies, and the current values of t and ω. The transition probability matrix $P(t)$ is calculated by equation (12) and used to weight pathways when the two codons differ at more than one position. This step generates the proportions of transitional (T) and transversional (V) differences for each of the synonymous and nonsynonymous site classes.

5. Correct for multiple hits to calculate d_S and d_N using counts of sites and differences and base frequencies at synonymous and nonsynonymous sites. This step updates t and ω: $t = d_S \times 3S/(S + N) + d_N \times 3N/(S + N)$, and $\omega = d_N/d_S$.
6. Repeat steps 4–5 until the algorithm converges.

In general, two or three rounds of iteration are sufficient. Some variations of the above algorithm are possible. For example, one may use an estimate of κ obtained externally. Furthermore, no iteration is needed if pathways are weighted equally when counting differences.

Comparison of Methods for Estimating d_S and d_N

We examine the performance of the following four methods for estimating d_S and d_N: ML (Goldman and Yang 1994), NG (Nei and Gojobori 1986), the method of Ina (1995; Method I), and the method of this paper. PAML (Yang 1999) was used for the ML and NG methods, and Ina's program was used for Ina's method. For error-checking, independent programs for the new method were written by the two authors.

Two approaches are taken to evaluate the methods. The first examines infinitely long sequences and may be termed a ''consistency analysis.'' Instead of the observed codons in the two sequences, the data consist of the expected frequencies (f_{ij}) of all 61 × 61 codon ''site patterns,'' calculated using equation (5) for given parameters t, κ, ω, and codon frequencies π_j. ML estimates are known to be statistically consistent when the model is correct (Stuart, Ord, and Arnold 1999, chapter 18). Since the d_N and d_S rates (and their ratio) are defined as functions of parameters t, κ, ω, and π_j, ML estimates of d_N and d_S will also be consistent. Approximate methods, including the method of this paper, involve ad hoc approximations and in general do not give the true values as estimates. They are statistically inconsistent. However, a good approximate method should not deviate too far from the truth with an infinite amount of data. The second approach we take is computer simulation. Finite data sets are generated by simulation and then analyzed by different methods to examine their biases and sampling variances.

We examine effects of the transition/transversion rate ratio (κ), base (codon) frequencies, and the selective pressure on the gene reflected in parameter ω. We initially fix $t = 1$ nucleotide substitution per codon, although the effect of sequence divergence is examined later. For a neutral gene ($\omega = 1$), this translates to 1/3 synonymous and 1/3 nonsynonymous substitutions per site. Three values of ω are considered: $\omega = 1$ (no selection), $\omega = 0.3$ (purifying selection) and $\omega = 3$ (positive selection). Estimates of ω (d_N/d_S) from real data vary widely from gene to gene, and $\omega = 0.3$ appears to represent moderate purifying selection (see, e.g., Ohta 1995; Li 1997; Yang and Nielsen 1998; Eyre-Walker and Keightley 1999). There are not many genes under positive selection, but estimates at about $\omega = 3$ are found in real data (e.g., Lee, Ota, and Vacquier 1995; Messier and Stewart 1997). Three sets of base frequencies at codon positions are used. The first set has equal base (codon) frequencies. The second set is from primate mitochondrial protein-coding genes and has very biased base frequencies. The third set is from HIV *env* genes (table 2). The universal genetic code is used in both the consistency analysis and the computer simulation.

Table 2
Base Frequencies at Codon Positions in Two Data Sets

	T	C	A	G
Primate mitochondrial protein-coding genes				
Position 1	0.205	0.283	0.308	0.204
Position 2	0.410	0.279	0.190	0.121
Position 3	0.151	0.433	0.371	0.045
HIV-1 *env* genes				
Position 1	0.190	0.162	0.367	0.281
Position 2	0.289	0.188	0.302	0.221
Position 3	0.249	0.169	0.368	0.214

Consistency Analysis Using Infinite Data

Consistency is the property that the estimate converges to the true value of the parameter as the amount of data approaches infinity. While consistency is a weak requirement, the approximate methods examined here are all inconsistent. It is nevertheless interesting to examine which steps of the approximate methods (i.e., counting sites, counting differences, and correcting for multiple hits) cause the bias. This is relatively easy since infinite data do not involve sampling errors, and estimates of sites (S and N) and rates (d_S and d_N) can be directly compared with the correct values.

Estimates of ω by different methods are plotted against the transition/transversion rate ratio κ for different values of the d_N/d_S ratio (ω). Results for the three sets of codon frequencies are shown in figure 2*A–I*.

Equal Codon Frequencies

We consider the NG method first. When base (codon) frequencies are equal and transition and transversion rates are equal ($\kappa = 1$), assumptions of the NG method are largely satisfied. In this case, NG indeed gives estimates close to the true values. Estimates of ω given by NG when $\kappa = 1$ are 1.001, 0.318, and 2.523 for the true $\omega = 1$, 0.3, and 3, respectively (fig. 2*A–C*). The method is biased toward 1 when the true $\omega \neq 1$ due to its use of equal weighting of pathways when counting sites. When there is transition bias ($\kappa > 1$), NG underestimates the ω ratio, and the bias is more serious when the transition bias is more extreme. This bias is mainly generated in the step of counting sites.

The case of $\kappa = 10$ is explored in table 3, which shows that NG substantially underestimates ω and gives

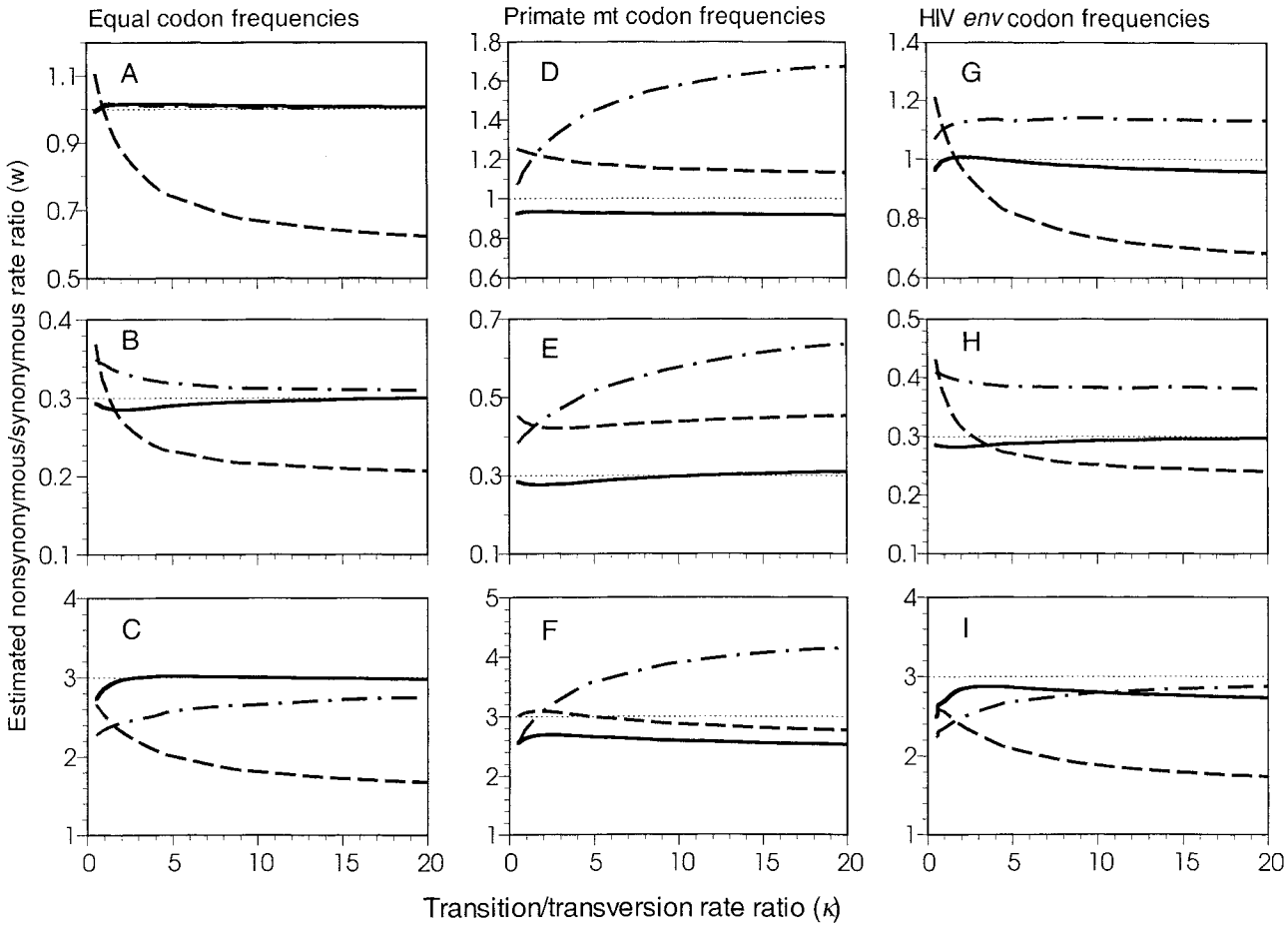

FIG. 2.—Estimates of ω (=d_N/d_S) as a function of κ for infinite data. The methods are ML (dotted lines, true values), NG (dashed lines), Ina's (1995) method (mixed lines), and the method of this paper (thick lines). Sequences of 2,000,000 codons were simulated for Ina's method, while for other methods, the data are the expected frequencies of the 61 × 61 site patterns. The universal genetic code is used. Three selection schemes are used: (*A*, *D*, and *G*) no selection (ω = 1), (*B*, *E*, and *H*) purifying selection (ω = 0.3), and (*C*, *F*, and *I*) positive selection (ω = 3). Three sets of base/codon frequencies are used: (*A–C*) equal frequencies, (*D–F*) frequencies from the primate mitochondrial protein-coding genes, and (*G–I*) frequencies from the HIV-1 *env* genes.

Table 3
Estimates of d_N, d_S, and ω in Infinite Data by Different Methods when κ = 10

True Values (ML)			NG			Ina (1995)			YN (this paper)		
Equal codon frequencies											
(*S*% = 32.5%)			(*S*% = 25.3%)			(*S*% = 33%)			(*S*% = 32.6%)		
0.333	0.333	1	0.277	0.413	0.669	0.346	0.343	1.010	0.332	0.329	1.009
0.190	0.633	0.3	0.166	0.771	0.216	0.202	0.646	0.312	0.190	0.642	0.295
0.426	0.142	3	0.340	0.188	1.812	0.434	0.158	2.749	0.423	0.141	3.001
Mitochondrial codon frequencies											
(*S*% = 23.7%)			(*S*% = 26.7%)			(*S*% = 31–32%)			(*S*% = 23.7%)		
0.333	0.333	1	0.291	0.253	1.151	0.348	0.221	1.575	0.291	0.316	0.922
0.215	0.715	0.3	0.204	0.463	0.439	0.237	0.407	0.582	0.198	0.665	0.298
0.396	0.132	3	0.330	0.115	2.878	0.397	0.101	3.925	0.336	0.129	2.592
HIV codon frequencies											
(*S*% = 28.5%)			(*S*% = 24.0%)			(*S*% = 30–32%)			(*S*% = 28.6%)		
0.333	0.333	1	0.272	0.370	0.735	0.339	0.298	1.139	0.312	0.320	0.974
0.200	0.668	0.3	0.176	0.699	0.252	0.213	0.555	0.385	0.193	0.659	0.293
0.411	0.137	3	0.321	0.170	1.884	0.399	0.143	2.795	0.378	0.135	2.799

NOTE.—The three values listed for each method are d_N, d_S, and ω. Estimates listed for maximum likelihood (ML) are the correct values. *S*% is the proportion of synonymous sites. Estimates of κ by the new method are 10.2–10.3, 10.4–10.9, and 10.4–10.6 for the three sets of codon frequencies, respectively. Corresponding estimates by Ina's method are 10.3–14.0, 4.7–6.0, and 5.5–9.5.

0.669, 0.216, and 1.812 for $\omega = 1$, 0.3, and 3, respectively. In this case, the proportion of synonymous sites (S%) should be 33%, but NG (assuming $\kappa = 1$) gives 26% (Yang and Nielsen 1998, fig. 3). Use of equal weighting (assuming $\omega = 1$) in counting differences by NG tends to bias the estimate of ω toward 1. However, compared with the bias in counting sites, the bias in counting differences is much less important because there may not be many pairs of codons that are different at two or three positions and because different pathways may involve the same numbers of synonymous and nonsynonymous changes. As mentioned above, NG underestimates S considerably (25.5/32.5 = 0.785) when $\kappa = 10$, and almost all of this underestimation is translated into the overestimation of d_S (0.3333/0.4134 = 0.806 for $\omega = 1$). For similar reasons, the underestimation of ω by the NG method is more serious for $\omega = 3$ than for $\omega = 0.3$ (fig. 2*B* and *C*). In the latter case, equal weighting assuming $\omega = 1$ in the NG method counterbalances the effect of ignoring κ in counting sites, while in the former case, the two biases are in the same direction (table 3).

Ina's (1995) method gives quite reliable estimates of ω for large values of κ (fig. 2*A–C*). For small values of κ, the method tends to overestimate ω when $\omega < 1$ and to underestimate ω when $\omega > 1$. This pattern appears to be due to the use of equal weighting of pathways in counting differences. The method of this paper underestimates ω slightly when the transition/transversion bias is weak (that is, when κ is close to 1) and when $\omega \neq 1$ (fig. 2*B–C*).

Primate Mitochondrial Codon Frequencies

Results obtained using base frequencies at the three codon positions from the primate mitochondrial genes (see table 2) are shown in figure 2*D–F*. Estimates of S, d_S, and d_N for $\kappa = 10$ are shown in table 3. The results are very different from those of figure 2*A–C* under equal codon frequencies. Except for small values of κ ($\kappa < 2$), Ina's method performs more poorly than NG. The two methods give very different counts of sites (S). While transition bias always leads to more synonymous sites, the effect of base frequency bias is more complicated. Extreme codon usage bias can cause the proportion of synonymous sites to range from 0% (e.g., when only codons TTC and TTA are present in the sequences) to 100% (e.g., when only codons CTT, CTC, CTA, and CTG are present). There tend to be more synonymous sites if the two most frequent nucleotides at third positions are both purines or both pyrimidines.

For the mitochondrial genes, S is much smaller than expected under equal base frequencies, causing NG to overestimate rather than underestimate S. For example, NG gives S% = 26.7%, which is higher than the correct value at either $\kappa = 1$ (23.1%) or $\kappa = 10$ (23.7%) (table 3). The overestimation of S caused by ignoring the base frequency bias more than compensates for the underestimation caused by ignoring the transition bias. As a result, NG overestimates ω when $\omega = 1$ (with estimates from 1.1 to 1.3; fig. 2*D*) or $\omega = 0.3$ (with estimates from 0.42 to 0.46; fig. 2*E*). When $\omega = 3$, equal weighting of pathways (assuming $\omega = 1$) in counting differences combined with the assumption of no transition bias ($\kappa = 1$) in counting sites cancels the effect of the base frequency bias in counting sites, such that NG produces a quite reliable estimate of ω (fig. 2*F*). Ina's (1995) method, by considering the transition bias alone and ignoring the base/codon frequency bias, substantially overestimates the proportion of synonymous sites and overestimates ω (table 3 and fig. 2*D–F*). Nevertheless, it should be noted that the observed pattern depends on the particular set of codon frequencies. For frequencies from other data sets, NG may be considerably worse than Ina's method.

The method of this paper is slightly better than NG for $\omega = 1$ although the two methods have opposite biases. When $\omega = 0.3$, the new method has little bias. When $\omega = 3$, the new method underestimates ω, with estimates from 2.5 to 2.6. Since the new method counts sites correctly (see table 3), the bias must be due to counting of differences and correction for multiple hits. Table 3 suggests that the new method underestimates both d_S and d_N, but the underestimation of d_N is more serious, leading to underestimation of the ω ratio. Apart from the case in which $\omega = 3$, the new method is better than both NG and Ina's method.

HIV env *Codon Frequencies*

Figure 2*G–I* shows estimates of ω when base/codon frequencies from the HIV envelope genes (see table 2) are used. Base frequencies in this gene are less biased than are those in the mitochondrial genes, and the effect of ignoring the base frequency bias is minor. For example, the correct proportion of synonymous sites at $\kappa = 1$ is 21.9%, while NG gives 24.0%, with very slight overestimation. Patterns in figure 2*G- -I* are quite similar to those for equal codon frequencies (fig. 2*A–C*). Exactly at $\kappa = 1$, NG gives the estimates 1.105, 0.371, and 2.554 when the true $\omega = 1$, 0.3, and 3, respectively. The estimates are biased toward 1, mainly due to the use of equal weighting in counting differences. When $\kappa = 10$, NG underestimates the proportion of synonymous sites (24.0% vs. the correct value, 28.5%). The bias is not as extreme as that in the case of equal codon frequencies, as unequal base/codon frequencies appear to counterbalance the effect of transition bias to some extent (tables 3 and 4).

Ina's (1995) method overestimates the ω ratio because it ignores the base frequency bias and thus overestimates the number of synonymous sites. The bias is not as extreme as it is for mitochondrial genes. The new method gives estimates very close to the true values for $\omega = 1$ and $\omega = 0.3$. When $\omega = 3$, the new method slightly underestimates the ratio, as in the case of mitochondrial codon frequencies.

Computer Simulations

The data consist of a pair of codon sequences and are simulated by sampling codon site patterns from the multinomial distribution specified by the site pattern probabilities f_{ij} (eq. 5). The sequence has $L_c = 100$ or 500 codons. Three values of κ are used: 1 (no bias), 2

Table 4
Average Estimates of ω in Simulated Replicates

	ω = 1				ω = 0.3				ω = 3			
	ML	NG	Ina (1995)	YN (this paper)	ML	NG	Ina (1995)	YN (this paper)	ML	NG	Ina (1995)	YN (this paper)
Equal codon frequencies												
κ = 1 . . . L_c = 100	1.107 ± 0.043	1.053	1.041	1.101	0.321 ± 0.010	0.331	0.347	0.316	3.579 ± 0.183	2.814	2.503	3.452
L_c = 500	1.004 ± 0.016	1.012	1.010	1.046	0.310 ± 0.005	0.322	0.344	0.302	3.104 ± 0.071	2.569	2.380	3.002
κ = 2 . . . L_c = 100	1.122 ± 0.039	0.909	1.030	1.069	0.309 ± 0.011	0.278	0.337	0.297	3.733 ± 0.247	2.569	2.568	3.470
L_c = 500	1.040 ± 0.017	0.886	1.015	1.034	0.299 ± 0.004	0.277	0.334	0.295	3.043 ± 0.064	2.303	2.456	3.005
κ = 20 . . L_c = 100	1.030 ± 0.032	0.669	1.014	1.097	0.284 ± 0.010	0.213	0.313	0.312	3.439 ± 0.155	1.845	2.907	3.385
L_c = 500	0.993 ± 0.016	0.635	1.011	1.019	0.296 ± 0.006	0.210	0.307	0.303	3.148 ± 0.062	1.697	2.773	2.986
Mitochondrial codon frequencies												
κ = 1 . . . L_c = 100	1.059 ± 0.034	1.312	1.205	1.016	0.310 ± 0.011	0.448	0.417	0.293	4.297 ± 0.381	3.414	3.056	3.317
L_c = 500	1.026 ± 0.019	1.260	1.175	0.967	0.310 ± 0.006	0.439	0.411	0.287	3.197 ± 0.070	3.132	2.848	2.739
κ = 2 . . . L_c = 100	1.122 ± 0.054	1.299	1.330	1.008	0.319 ± 0.010	0.440	0.456	0.291	3.683 ± 0.218	3.433	3.381	3.141
L_c = 500	1.062 ± 0.020	1.226	1.281	0.957	0.306 ± 0.005	0.427	0.446	0.282	3.213 ± 0.071	3.187	3.194	2.822
κ = 20 . . L_c = 100	1.057 ± 0.034	1.240	1.785	1.004	0.296 ± 0.013	0.460	0.636	0.322	3.708 ± 0.210	3.161	4.479	3.007
L_c = 500	1.010 ± 0.018	1.145	1.693	0.933	0.298 ± 0.006	0.462	0.646	0.313	3.150 ± 0.065	2.847	4.191	2.611
HIV codon frequencies												
κ = 1 . . . L_c = 100	1.136 ± 0.046	1.158	1.142	1.112	0.317 ± 0.010	0.398	0.400	0.321	5.263 ± 1.019	2.845	2.615	3.311
L_c = 500	1.035 ± 0.020	1.107	1.098	1.036	0.318 ± 0.006	0.373	0.410	0.299	3.302 ± 0.087	2.636	2.356	2.893
κ = 2 . . . L_c = 100	1.170 ± 0.051	1.016	1.152	1.098	0.294 ± 0.010	0.332	0.394	0.308	4.719 ± 0.772	2.634	2.709	3.417
L_c = 500	1.050 ± 0.018	0.993	1.138	1.052	0.305 ± 0.005	0.319	0.397	0.291	3.075 ± 0.071	2.448	2.525	3.003
κ = 20 . . L_c = 100	1.106 ± 0.039	0.732	1.167	1.051	0.305 ± 0.011	0.252	0.382	0.310	3.402 ± 0.157	1.911	3.056	3.130
L_c = 500	1.008 ± 0.018	0.681	1.139	0.952	0.293 ± 0.006	0.248	0.383	0.304	3.049 ± 0.057	1.778	2.907	2.782

NOTE.—The number of replicates is 100 for maximum likelihood (ML) and 500 for other methods. Standard errors are presented for ML only.

(small bias), and 20 (large bias). Most estimates of κ from nuclear genes are in the range (1.5, 5), so a value of 2 is typical. Estimates from mitochondrial genes vary considerably among data sets, from 2 or 3 to over 100.

The averages of the ω estimates among simulated replicates are listed in table 4 for the three sets of codon frequencies. Standard errors for the ML estimates are also presented, while those for other methods (not shown) are very small due to the use of many more replicates. Averages of d_S and d_N are calculated for all methods but not shown. We note that the simulation results are highly consistent with those found for infinite data, discussed above. For example, if a method gives estimates smaller than the true value in infinite data, it tends to have negative biases in finite samples as well.

ML estimates are known to be often biased in small samples. Table 4 shows that MLEs of ω are nearly unbiased when ω = 1 or 0.3 for all three sets of codon frequencies and for all values of κ. However, it is biased to larger values when ω = 3. Although the bias is small in large genes (with 500 codons), it can be quite large for small genes (L_c = 100), especially when κ is small.

The NG method has little bias if codon frequencies are equal and if there is no transition/transversion bias (κ = 1). When ω ≠ 1, the method tends to bias toward 1 due to its use of equal weighting in counting sites. The bias is nevertheless small. These results agree well with previous simulations by Ota and Nei (1994) and Muse (1996), who used similar simple models to examine the performance of NG. However, NG is biased in most other parameter combinations. The biases in general agree with findings of the consistency analysis (fig. 2). In particular, ignoring the transition/transversion bias leads to underestimates of ω and ignoring unequal base frequencies leads to overestimates of ω. For equal codon frequencies and no selection (ω = 1), NG gives severe underestimates of ω when κ is large. The effects of the transition/transversion rate bias and base frequency biases tend to cancel each other, such that NG has smaller biases than ML when ω = 3 for the mitochondrial codon frequencies. In almost all other cases, ML has smaller biases than NG.

Ina's (1995) method has small biases when base frequencies are equal. The method tends to overestimate ω when ω < 1 and to underestimate ω when ω > 1, probably due to its use of equal weighting in counting sites. This is the same pattern as that found in infinite data (fig. 2*A–C*). Ina's method considerably overestimates ω for all values of ω under the mitochondrial codon frequencies, probably because it overestimates the number of synonymous sites. For the HIV *env* codon frequencies, the method overestimates ω when ω ≤ 1 and underestimates ω when ω > 1, as noted for infinite data (fig. 2*G–I*).

The new method appears to have little bias over most of the parameter space examined (table 4). When ω ≤ 1, it is less biased than NG or Ina's (1995) method. When ω = 3, it tends to overestimate ω in small samples, like the ML method, but the bias seems smaller than that of ML. The new method appears to provide a close approximation of ML over the range of parameter values examined.

Since all methods are biased for at least some parameter combinations, the mean squared error (MSE) may be an appropriate criterion by which to compare methods. The MSE of a parameter estimator $\hat{\theta}$ is defined

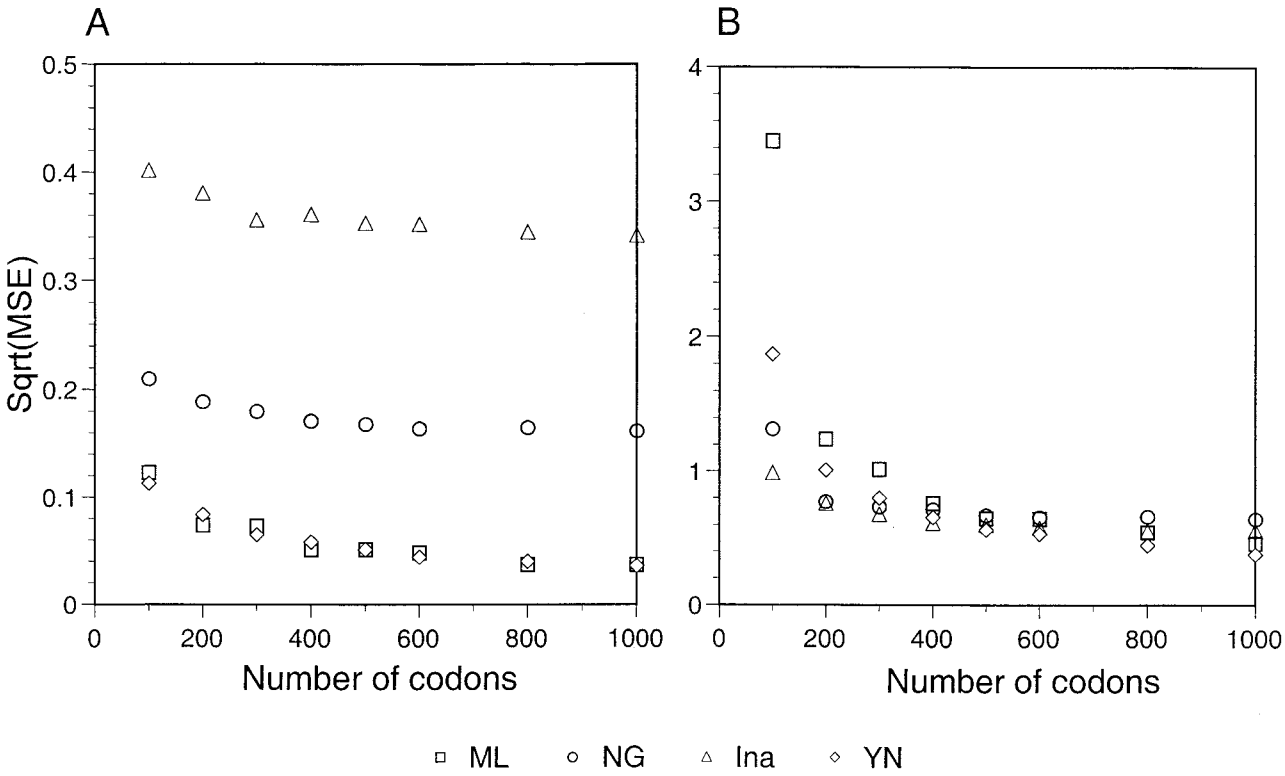

FIG. 3.—Square root of the mean square error, MSE($\hat{\omega}$), of the estimated ω ratio, as a function of the sequence length L_c, calculated by computer simulation. Parameter values are (*A*) $t = 1$, $\kappa = 20$, $\omega = 0.3$, with base/codon frequencies from the primate mitochondrial genes (table 2), and (*B*) $t = 1$, $\kappa = 2$, $\omega = 3$, with base/codon frequencies from the HIV-1 *env* genes (table 2). The universal genetic code is used.

as $\text{MSE}(\hat{\theta}) = \text{E}(\hat{\theta} - \theta)^2$, where θ is the true value. Since $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{E}(\hat{\theta}) - \theta]^2$, this measures both bias and variance. The square root of the MSE is plotted in figure 3 against the sequence length (number of codons). Two parameter combinations are considered. The first is for mitochondrial genes with a strong transition bias ($\kappa = 20$) and purifying selection ($\omega = 0.3$), and the second is for the HIV *env* gene with moderate transition bias ($\kappa = 2$) and positive selection ($\omega = 3$). In the first case (fig. 3*A*), ML and the method of this paper performed very similarly, and both have the smallest MSEs, while Ina's (1995) method has very large MSEs due to the positive bias of the method (fig. 2*E*). In the second case (fig. 3*B*), ML performed much worse than other methods for short genes ($L_c < 300$ codons) due to its large positive bias at $\omega = 3$, while for large genes ($L_c > 500$), ML and the new method are better than NG and Ina's method. In both cases, the new method lies between NG and ML.

We also performed a small-scale simulation to examine the effect of sequence divergence level (*t*). The results are shown in figure 4. We examine two sets of parameter values, with the sequence length fixed at $L_c = 500$. In the first case (fig. 4*A*), equal codon frequencies are used with $\kappa = 2$ and $\omega = 0.3$. The new method of this paper overestimates ω at small divergences but underestimates ω at large divergences. Other methods are insensitive to sequence divergence level. ML and the new method are less biased than NG and Ina's method. In the second case (fig. 4*B*), mitochondrial codon frequencies are used with $\kappa = 20$ and $\omega = 0.3$. In this case, ML and the new method have little bias over the whole range of the sequence divergence level. Note that the synonymous rate is quite high, with $d_S = 0.71$ at $t = 1$, and $d_S = 1.1$ at $t = 1.5$. NG and Ina's method involve positive biases, and the biases become more serious when the sequences are more divergent. Although average estimates of both d_N and d_S by NG increase with *t*, d_N increases at a faster rate, such that the ω ratio increases with the increase of *t*. Muse (1996) discussed the fact that at high sequence divergences, NG does not produce distance estimates linear with time.

Comparison of Human and Orangutan Mitochondrial Genes

The concatenated sequences of the 12 protein-coding genes on the H-strand of the mitochondrial genome from the human (*Homo sapiens*, GenBank accession number D38112) and the orangutan (*Pongo pygmaeus p.*, GenBank accession number D38115) are compared using different methods. The results are shown in table 5. We also included the method of Li (1993) in the comparison, implemented in X. Xia's DAMBE program (available at http://web.hku.hk/~xxia). ML is applied with different assumptions concerning the transition bias and the codon frequency bias. Note that estimates of the d_N/d_S ratio vary by up to threefold depending on the method/model used. The pattern is especially revealing for ML estimates under different models. For example,

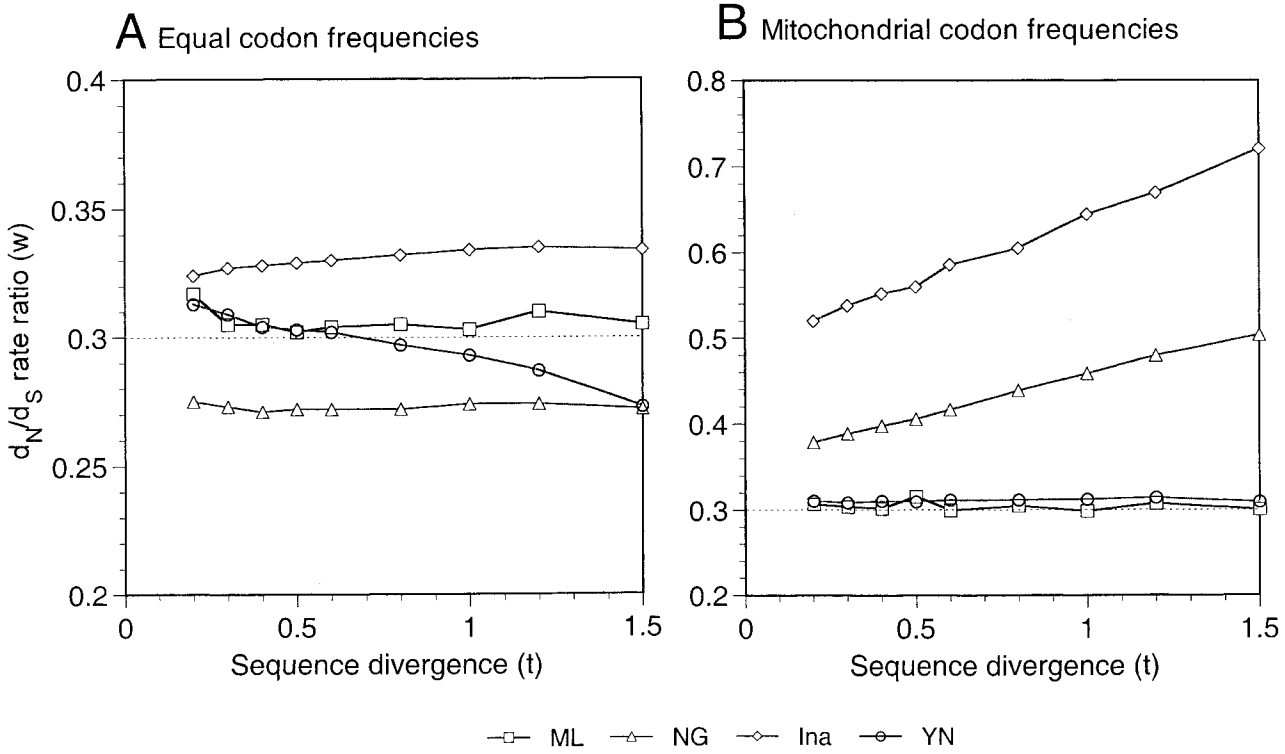

FIG. 4.—The averages of estimates of ω as a function of sequence divergence level *t*. The sequence length is $L_c = 500$ codons. *A*, Equal codon frequencies are assumed, with $\kappa = 2$ and $\omega = 0.3$. *B*, Codon frequencies from primate mitochondrial genes are used, with $\kappa = 20$ and $\omega = 0.3$. Each average was obtained by simulating 2,000 replicates for NG and YN (the method of this paper), 500 for the method of Ina (1995), and 100 for ML.

accounting for the transition/transversion rate bias (Fequal, κ estimated) increased the proportion of synonymous sites from 25% to 33%, and increased the d_N/d_S ratio from 0.093 to 0.130. Accounting further for the codon usage bias (F60, κ estimated) decreased the proportion of synonymous sites to 28%, with an estimate of $\omega = 0.045$. The pattern is the same as that in the consistency analysis and the computer simulation discussed above. The results demonstrate that the estimation method is important for estimating d_N and d_S (and their ratio ω) from real data and that methods accounting for both the transition bias and base/codon frequency bias should be used. This conclusion is consistent with Ina (1995), who found that none of the approximate methods he examined performed well when base/codon frequencies were extreme.

Furthermore, we note that estimates from the NG method are similar to those of ML assuming no transition bias ($\kappa = 1$ fixed) and no base frequency bias, and estimates obtained from Li (1993) are similar to ML accounting for the transition bias (κ estimated) and no base frequency bias. Ina's (1995) program did not run for this data set. Nevertheless, we expect Ina's method to be close to that of Li (1993) or ML accounting for the transition bias and no base frequency bias. Some minor differences in implementation between ML and the corresponding approximate methods were discussed by Yang and Nielsen (1998). Those results suggest that

Table 5
Proportions of Synonymous Sites (*S*%) and d_S and d_N Rates Between the Human and Orangutan Mitochondrial Genes

Method	*S*%	d_N	d_S	$\omega = d_N/d_S$	ℓ
NG	25.7%	0.060	0.668	0.090	
Li (1993)	NA	0.066	0.529	0.125	
YN ($\hat{\kappa} = 10.6$) (this paper)	24.5%	0.058	1.010	0.057	
ML (Fequal, $\kappa = 1$ fixed)	25.2%	0.057	0.614	0.093	−18744.7
ML (Fequal, $\hat{\kappa} = 6.1$)	32.8%	0.065	0.497	0.130	−18344.6
ML (F1 × 4, $\hat{\kappa} = 8.0$)	33.1%	0.064	0.784	0.082	−17647.5
ML (F3 × 4, $\hat{\kappa} = 12.6$)	23.2%	0.060	1.457	0.041	−16844.2
ML (F60, $\hat{\kappa} = 14.0$)	27.7%	0.062	1.388	0.045	−16447.0

NOTE.—Fequal: equal codon frequencies (=1/60) are assumed; F1 × 4: four nucleotide frequencies are used to calculate codon frequencies (three free parameters); F3 × 4: nucleotide frequencies at three codon positions are used to calculate codon frequencies (nine free parameters); F60: all codon frequencies are used as free parameters (59 free parameters). ℓ is the log likelihood value.

ad hoc treatments involved in the approximate methods may not have introduced too much bias and that failure to account for the transition bias and base frequency bias appears to be more important. However, Muse (1996) points out that at high sequence divergence levels, ad hoc treatments (such as those used in multiple-hit correction) in approximate methods may become a more serious problem (see also fig. 4).

Discussion

The approximate method of this paper accounts for two major features of DNA sequence evolution: transition bias and base/codon frequency bias. The consistency analysis of infinite data and the computer simulation of finite data suggest that the new method has smaller biases than either NG or Ina's (1995) method for almost all parameter combinations examined and produces estimates similar to those obtained with ML. The method may thus be useful for large-scale screening, when ML may be too time-consuming. Our analyses also suggest that NG and Ina's method may involve large biases when there exist transition/transversion rate bias and base/codon frequency bias, and that it is important to account for those features of DNA sequence evolution.

The ML method for pairwise comparison is less biased and has a lower MSE than the approximate methods for almost all parameter combinations. Only for short sequences and high ω ratios does it involve a positive bias and perform more poorly than some of the approximate methods. We suggest that, in general, the ML method, which accounts for both the transition bias and the codon usage bias, should be the preferred method for estimating d_S and d_N between two sequences. Only in the case of very short sequences may it be advantageous to use simpler models. In the course of this study, we realized that correcting for biases involved in the NG method is extremely complicated, despite the fact that the method is well known for its simplicity. In contrast, ML is conceptually much simpler, mainly because the probability theory employed by the method takes care of the difficult tasks of weighting evolutionary pathways and correcting for multiple hits, with no need for ad hoc approximations. Specifically, the Chapman-Kolmogorov theorem (e.g., Grimmett and Stirzaker 1992, p. 239) states that $p_{ij}(t) = \Sigma_k\, p_{ik}(s)p_{kj}(t - s)$ for any $0 \le s \le t$; that is, the probability that codon i changes to codon j over time t is a sum over all possible codons (k) at any intermediate time point s. This obvious result ensures that the likelihood calculation (eqs. 4–6) accounts for all possible pathways of changes between the two codons, weighting them appropriately according to their relative probabilities of occurrence.

The major advantage of ML appears to lie in its flexibility in simultaneous comparison of multiple sequences, taking into account their phylogenetic relationship. Hypotheses concerning variable d_N/d_S ratios among lineages (Yang 1998; Yang and Nielsen 1998) or among sites (Nielsen and Yang 1998) can be tested using the likelihood ratio test. The ML model can easily be extended to include important features of DNA sequence evolution such as the dependence of nonsynonymous rates on the chemical properties of the amino acids (Yang, Nielsen, and Hasegawa 1998).

Program Availability and Performance

A C program implementing the approximate method of this paper will be included in the PAML package, available at http://abacus.gene.ucl.ac.uk/software/paml.html. On a fast Pentium II, each pairwise comparison takes about 10–20 s by ML and a few seconds by the method of this paper. If pathways are weighted equally in counting differences in the new method, iteration will not be needed, and the method will be about as fast as other approximate methods such as NG, which seem to finish instantaneously.

Acknowledgments

We thank Hinrich Schulenburg, the two referees Keith Crandall and Spencer Muse, and Associate Editor Caro-Beth Stewart for many constructive comments. We thank X. Xia for the analysis using the method of Li (1993). This study is supported by a BBSRC grant to Z.Y. and NSF grant DEB 9815367.

LITERATURE CITED

AKASHI, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at ''silent'' sites in Drosophila DNA. Genetics **139**:1067–1076.

COMERON, J. M. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41**:1152–1159.

CRANDALL, K. A., and D. M. HILLIS. 1997. Rhodopsin evolution in the dark. Nature **387**:667–668.

EYRE-WALKER, A., and P. D. KEIGHTLEY. 1999. High genomic deleterious mutation rates in hominoids. Nature **397**:344–347.

GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford, England.

GOJOBORI, T. 1983. Codon substitution in evolution and the ''saturation'' of synonymous changes. Genetics **105**:1011–1027.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

GRIMMETT, G. R., and D. R. STIRZAKER. 1992. Probability and random processes. 2nd edition. Clarendon Press, Oxford, England.

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

INA, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40**:190–226.

———. 1996. Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. J. Genet. **75**:91–115.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–123 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

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

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

LEE, Y. H., T. OTA, and V. D. VACQUIER. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12**:231–238.

LEWONTIN, R. 1989. Inferring the number of evolutionary events from DNA coding sequence differences. Mol. Biol. Evol. **6**:15–32.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

———. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and non-synonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MESSIER, W., and C.-B. STEWART. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151–154.

MIYATA, T., and T. YASUNAGA. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. J. Mol. Evol. **16**: 23–36.

MORIYAMA, E. N., and J. R. POWELL. 1997. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. J. Mol. Evol. **45**:378–391.

MUSE, S. V. 1996. Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. **13**:105–114.

MUSE, S. V., and B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to chloroplast genome. Mol. Biol. Evol. **11**:715–724.

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

NIELSEN, R., and Z. YANG. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

OHTA, T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. **40**56–63.

OTA, T., and M. NEI. 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol. Biol. Evol. **11**:613–619.

PAMILO, P., and N. O. BIANCHI. 1993. Evolution of the *Zfx* and *Zfy* genes—rates and interdependence between the genes. Mol. Biol. Evol. **10**:271–281.

PEDERSEN, A.-M. K., C. WIUF, and F. B. CHRISTIANSEN. 1998. A codon-based model designed to describe lentiviral evolution. Mol. Biol. Evol. **15**:1069–1081.

PERLER, F., A. EFSTRATIADIS, P. LOMEDICA, W. GILBERT, R. KOLODNER, and J. DODGSON. 1980. The evolution of genes: the chicken preproinsulin gene. Cell **20**:555–566.

STUART, A., K. ORD, and S. ARNOLD. 1999. Kendall's advanced theory of statistics. Vol. 2a, 6th edition. Arnold, London.

TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **11**:261–277.

YANG, Z. 1994*a*. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

———. 1994*b*. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

———. 1999. Phylogenetic analysis by maximum likelihood (PAML). Version 2. University College London, England.

YANG, Z., and R. NIELSEN. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46**:409–418.

YANG, Z., R. NIELSEN, and M. HASEGAWA. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15**:1600–1611.

CARO-BETH STEWART, reviewing editor

Accepted September 6, 1999