Performance of Likelihood Ratio Tests of Evolutionary Hypotheses Under Inadequate Substitution Models

Jianzhi Zhang[1]

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

In recent years, likelihood ratio tests (LRTs) based on DNA and protein sequence data have been proposed for testing various evolutionary hypotheses. Because conducting an LRT requires an evolutionary model of nucleotide or amino acid substitution, which is almost always unknown, it becomes important to investigate the robustness of LRTs to violations of assumptions of these evolutionary models. Computer simulation was used to examine performance of LRTs of the molecular clock, transition/transversion bias, and among-site rate variation under different substitution models. The results showed that when correct models are used, LRTs perform quite well even when the DNA sequences are as short as 300 nt. However, LRTs were found to be biased under incorrect models. The extent of bias varies considerably, depending on the hypotheses tested, the substitution models assumed, and the lengths of the sequences used, among other things. A preliminary simulation study also suggests that LRTs based on parametric bootstrapping may be more sensitive to substitution models than are standard LRTs. When an assumed substitution model is grossly wrong and a more realistic model is available, LRTs can often reject the wrong model; thus, the performance of LRTs may be improved by using a more appropriate model. On the other hand, many factors of molecular evolution have not been considered in any substitution models so far built, and the possibility of an influence of this negligence on LRTs is often overlooked. The dependence of LRTs on substitution models calls for caution in interpreting test results and highlights the importance of clarifying the substitution patterns of genes and proteins and building more realistic models.

Introduction

Likelihood ratio tests (LRTs) of evolutionary hypotheses for DNA and protein sequences have been advocated by a number of authors in recent years (e.g., Felsenstein 1981; Muse and Weir 1992; Goldman 1993; Gaut and Weir 1994; Huelsenbeck and Rannala 1997; Huelsenbeck and Crandall 1997; Nielsen and Yang 1998). The suitability of LRTs of phylogenies, however, has been questioned because the statistical foundation of phylogeny estimation by likelihood has not been well established (Nei 1987, 1996; Yang, Goldman, and Friday 1995; Gaut and Lewis 1995; Yang 1996). Nevertheless, given the true tree, LRTs of hypotheses of nucleotide or amino acid substitution patterns (e.g., the molecular clock, equality of base frequencies, equality of transitional and transversional substitution rates, and uniformity of substitution rate among sites) are thought to be very useful. This is particularly so when the two hypotheses under comparison are nested, i.e., the null hypothesis H_0 is a subset or a special case of the alternative hypothesis H_1. In this case, the likelihood ratio (LR) $= 2(\ln L_1 - \ln L_0)$ is asymptotically χ^2 distributed, with the number of degrees of freedom (df) equal to $r = q - p$, where L_0 and L_1 are the maximum likelihoods of the null and alternative hypotheses, and p and q are the numbers of free parameters to be estimated under the two hypotheses, respectively.

Although LRTs are conceptually simple, mathematically sound, and operationally versatile (see Stuart and Ord [1991] for statistical properties of LRTs), they depend on the assumptions of models that describe the data. If the assumed models are incorrect, the useful distributional property of LR does not necessarily hold (e.g., Foutz and Srivastava 1977). This raises questions about the applicability of LRTs in the study of molecular evolution, because the actual pattern of nucleotide or amino acid substitution of a gene or protein is usually unknown. Furthermore, empirical studies have shown that LRTs of evolutionary hypotheses may give different results depending on the assumed models. For example, Huelsenbeck and Rannala (1997) showed that the hypothesis of a molecular clock was rejected for the mitochondrial cytochrome *c* oxidase subunit I (COI) genes of 13 gopher species when the F84 model (Felsenstein 1984; see fig. 1) of nucleotide substitution was assumed, but was not rejected when a discrete gamma (dG) distribution of among-site rate variation was considered (F84+dG model). It was also found in an analysis of the bacteriophage T7 sequences that the superiority of the general reversible (REV) model (fig. 1) to the HKY model (Hasegawa, Kishino, and Yano 1985; fig. 1) of nucleotide substitution in fitting the data depends on whether the among-site rate heterogeneity is considered (Cunningham, Zhu, and Hillis 1998). The causes of these observations have not been well studied, and confusion has been generated, such as the belief that the choice of the best-fit model relies on the parameter addition sequence in a series of LRTs (Cunningham, Zhu, and Hillis 1998). In general, how well LRTs perform under inadequate substitution models is unknown. In this paper, I first illustrate the influence of substitution models on LRTs by using a numerical example and then

[1] Present address: Laboratory of Host Defenses, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland.

Key words: likelihood ratio test, substitution models, molecular clock, transition/transversion bias, rate variation among sites, molecular evolution.

Address for correspondence and reprints: Jianzhi Zhang, Laboratory of Host Defenses, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Building 10, Room 11N104, 9000 Rockville Pike, Bethesda, Maryland 20892. E- mail: jzhang@atlas.niaid.nih.gov.

Mol. Biol. Evol. 16(6):868–875. 1999

$$\begin{pmatrix} . & 1 & 1 & 1 \\ 1 & . & 1 & 1 \\ 1 & 1 & . & 1 \\ 1 & 1 & 1 & . \end{pmatrix} \quad \begin{pmatrix} . & \kappa & 1 & 1 \\ \kappa & . & 1 & 1 \\ 1 & 1 & . & \kappa \\ 1 & 1 & \kappa & . \end{pmatrix} \quad \begin{pmatrix} . & \pi_C & \pi_A & \pi_G \\ \pi_T & . & \pi_A & \pi_G \\ \pi_T & \pi_C & . & \pi_G \\ \pi_T & \pi_C & \pi_A & . \end{pmatrix}$$

JC K80 F81

$$\begin{pmatrix} . & (1+\kappa/\pi_Y)\pi_C & \pi_A & \pi_G \\ (1+\kappa/\pi_Y)\pi_T & . & \pi_A & \pi_G \\ \pi_T & \pi_C & . & (1+\kappa/\pi_R)\pi_G \\ \pi_T & \pi_C & (1+\kappa/\pi_R)\pi_A & . \end{pmatrix}$$

F84

$$\begin{pmatrix} . & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & . & \pi_A & \pi_G \\ \pi_T & \pi_C & . & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & . \end{pmatrix} \quad \begin{pmatrix} . & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & . & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & . & \pi_G \\ c\pi_T & e\pi_C & \pi_A & . \end{pmatrix}$$

HKY REV

FIG. 1.—Rate matrices of various models of nucleotide substitution. π_T, π_C, π_A, and π_G are the equilibrium frequencies for bases T, C, A, and G, respectively, and $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. κ is the transition/transversion rate ratio, and *a*, *b*, *c*, *d*, and *e* are five substitution parameters of the REV model. The entry in the *i*th row and the *j*th column of a matrix is the relative rate of substitution from *i* to *j*, where *i* and *j* = 1, 2, 3, and 4, in place of T, C, A, and G, respectively. The references for these models are as follows: JC, Jukes and Cantor (1969); K80, Kimura (1980); F81, Felsenstein (1981); F84, Felsenstein (1984); HKY, Hasegawa, Kishino, and Yano (1985); REV, Yang (1997).

show by computer simulation that LRTs may become liberal or conservative depending on the conditions studied. In this study, I am primarily interested in the LRTs of nested hypotheses in which the null distribution of LR can be approximated by a χ^2 distribution with a known value of df. For those LRTs in which the null distribution of LR is unknown, parametric bootstrapping (simulation) is often used to generate the distribution (e.g., Goldman 1993). I briefly discuss the influence of wrong models on this type of LRT as well.

Dependence of LRTs on Substitution Models: A Numerical Example

Due to the complexity of parameter estimation under the maximum-likelihood criterion, it is not trivial to demonstrate analytically the influence of substitution models on LRTs. Therefore, I will use a very simple hypothetical example to illustrate the sensitivity of LRTs to substitution models. Suppose that paralogous proteins A and B of rodents are both 300 amino acids long. There are 15 amino acid differences between the mouse A and rat A sequences, and 5 differences between the mouse B and rat B sequences (see fig. 2). Using this information, we want to test the null hypothesis that the amino acid substitution rates for proteins A and B are equal (i.e., a molecular clock holds for the paralogous proteins). Let us first assume that all amino acid residues of a protein are variable and have the same substitution rate (model I) and that substitutions follow the Poisson process. Let $t/2$ equal the time since the separation of the mouse and rat, and let r_A and r_B be the (average) substitution rates per amino acid site per year for proteins A and B, respectively. We then have $p_A = 1 - e^{-r_A t}$, where p_A is the probability that a (potentially variable) site has different amino acids between the mouse A and the rat A sequences. Similarly, we have $p_B = 1 - e^{-r_B t}$. The likelihoods of H_0 and H_1 are as follows.

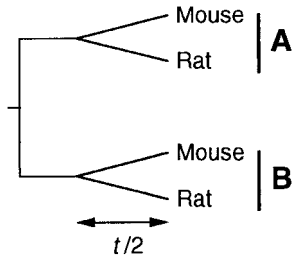

FIG. 2.—The evolutionary tree of paralogous proteins A and B of rodents used in the hypothetical example. There are 15 amino acid differences in protein A between the mouse and the rat, and there are 5 differences in protein B between these two species.

$$L_0 = \binom{300}{15} p_A^{15}(1 - p_A)^{285} \binom{300}{5} p_B^5(1 - p_B)^{295},$$

$$\text{with the restriction of } r_A = r_B, \qquad (1)$$

and

$$L_1 = \binom{300}{15} p_A^{15}(1 - p_A)^{285} \binom{300}{5} p_B^5(1 - p_B)^{295},$$

$$\text{without the restriction of } r_A = r_B. \qquad (2)$$

The likelihood estimates of the parameters are presented in table 1, and LR = 5.4. Since the null hypothesis is a special case of the alternative hypothesis, the χ^2 test is used to examine the statistical significance of the likelihood ratio LR, and H_0 is rejected at the 5% significance level (df = 1, $\chi^2_{0.05} = 3.84$).

In the above test, we assumed that there was no variation of substitution rate among different sites of a protein, which is unlikely to hold in reality (e.g, Fitch and Margoliash 1967; Uzzell and Corbin 1971; Zhang and Gu 1998). Now, suppose that in protein B only 10 sites are variable and that the rates are equal for these 10 sites, but in protein A all sites have the same rate (model II). Under this model, we may test the null hypothesis that the average substitution rates of the two proteins are equal. We now have

$$L_0 = \binom{300}{15} p_A^{15}(1 - p_A)^{285} \binom{10}{5} p_B^5(1 - p_B)^5 \binom{290}{0} 0^0 1^{290},$$

$$\text{with the restriction of } r_A = r_B, \qquad (3)$$

and

Table 1
Likelihood Ratio Tests of Equal Substitution Rates of Proteins A and B Under Three Different Models

	Model I		Model II		Model III	
	H_0	H_1	H_0	H_1	H_0	H_1
$r_A t$	0.034	0.051	0.040	0.051	0.039	0.069
$r_B t$	0.034	0.017	0.040	0.023	0.039	0.017
$2 \ln L$	−13.37	−7.97	−9.89	−7.31	−16.13	−7.33
LR = $2(\ln L_1 - \ln L_0)$	5.40*		2.58		8.80**	

* 5% significant.
** 0.5% significant.

$$L_1 = \binom{300}{15} p_A^{15}(1-p_A)^{285}\binom{10}{5} p_B^5(1-p_B)^5\binom{290}{0} 0^0 1^{290},$$
without the restriction of $r_A = r_B$, (4)

where $p_A = 1 - e^{-r_A t}$, and $p_B = 1 - e^{-30 r_B t}$. Because only 10 sites are variable in protein B, the rate at a variable site is 30 times higher than the average rate (r_B) per site for the entire protein. In this case, LR = 2.58, and the null hypothesis cannot be rejected at the 5% significance level (table 1).

Now, let us assume that in protein A only 30 sites are variable and that the rates are equal for these 30 sites, but in protein B all the sites have the same rate (model III). Under this model, we may also test the null hypothesis that the average substitution rates of the two proteins are equal. We then have

$$L_0 = \binom{30}{15} p_A^{15}(1-p_A)^{15}\binom{270}{0} 0^0 1^{270}\binom{300}{5} p_B^5(1-p_B)^{295},$$
with the restriction of $r_A = r_B$, (5)

and

$$L_1 = \binom{30}{15} p_A^{15}(1-p_A)^{15}\binom{270}{0} 0^0 1^{270}\binom{300}{5} p_B^5(1-p_B)^{295},$$
without the restriction of $r_A = r_B$, (6)

where $p_A = 1 - e^{-10 r_A t}$ and $p_B = 1 - e^{-r_B t}$. Note that since only 30 sites are variable in protein A, the rate at a variable site is 10 times higher than the average rate (r_A) per site for the whole protein. Now, LR = 8.80, and the null hypothesis is rejected at the 0.5% level ($\chi^2_{0.005} = 7.88$; table 1).

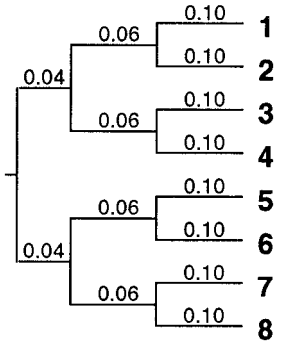

FIG. 3.—The model tree used in the computer simulation. The branch lengths (expected numbers of nucleotide substitutions per site) are given on the branches.

It is clear from this example that likelihood functions are model dependent and that likelihood estimates vary considerably depending on the model assumed. For example, the estimated average substitution rate of protein A under H_1 is about 40% higher in model III than in models I and II (table 1). It is also evident that the likelihood ratio does not remain the same under different models. Apparently, the effects of a change in the model on L_0 and L_1 are not canceled. For instance, compared with model I, use of model II improved the likelihoods for both H_0 and H_1, but the increase (3.48) of $2 \ln L_0$ is much greater than the increase (0.66) of $2 \ln L_1$. In contrast, use of model III improved the likelihood of H_1 but decreased the likelihood of H_0, such that H_1 becomes very favorable under model III. Although three different models have been considered in this example, it is likely that none of them are close to the true substitution pattern of the proteins. In practice, if the sequences of proteins A and B are obtained from many species, it may be possible to build more realistic substitution models, and the question of the equality of the substitution rates of proteins A and B may be answered with a higher certainty.

Performance of LRTs Under Inadequate Models: Computer Simulations

The extent of the influence of a wrong substitution model on LRTs of some generally interesting evolutionary hypotheses is unknown. In the following, I examine this problem by computer simulation. In all simulations, the true tree topology is used for an LRT, since the influence of a wrong tree topology on LRTs is beyond the scope of this study.

Test of the Molecular Clock

Because the LRT of a molecular clock gave different results under different substitution models in real data analysis (Huelsenbeck and Rannala 1997), I first examined the performance of the molecular-clock test. The tree of figure 3 was used as the model tree for all simulations in the paper, and a molecular clock was assumed in the simulation. The null hypothesis (H_0) is that sequence evolution follows a molecular clock, and the alternative hypothesis (H_1) is that sequence evolution is not constrained by a clock. Since H_0 is a special case of H_1, the two hypotheses are nested. LR is asymptotically χ^2 distributed with df = $n - 2 = 8 - 2 = 6$, where n is the number of sequences in the tree (Felsenstein 1981).

Table 2
The Rate Matrix of the REV Substitution Model Used in the Simulation

	T	C	A	G
T		0.70800	0.21214	0.06367
C	0.84040		0.33966	0.07807
A	0.27051	0.36489		0.31931
G	0.11670	0.12054	0.45895	
Frequency	0.31520	0.26560	0.24720	0.17200

NOTE.—An entry in the rate matrix (except in the last line) gives the rate of a particular substitution from a base at the left to a base along the top. The last line of the table gives the equilibrium base frequencies. This matrix, known as the Q matrix in Yang (1997), has been normalized so that the average substitution rate is 1. The parameters of the matrix are estimated from vertebrate mitochondrial genes (see text). The average transition/transversion ratio is 1.53.

The nucleotide substitution model used in the simulation was a general reversible (REV) model (see fig. 1) with a gamma distribution describing the rate heterogeneity among sites. The parameters of the rate matrix of the REV model (see table 2) were those estimated from the vertebrate mitochondrial genes, so that our simulation somewhat resembled the evolution of mitochondrial genes. In the estimation of the substitution parameters, the likelihood method with the REV+dG model was applied to the mitochondrial protein-coding genes of 11 vertebrate species. These species were the fin whale (*Balaenoptera physalus*; GenBank accession number X61145), the blue whale (*Balaenoptera musculus*; X72704), the cow (*Bos taurus*; V00654), the rat (*Rattus norvegicus*; X14848), the mouse (*Mus musculus*; V00711), the opossum (*Didelphis virginiana*; Z29573), the chicken (*Gallus gallus*; X52392), the African clawed frog (*Xenopus laevis*; X02890), the carp (*Cyrinus carpio*; X61010), the loach (*Crossostoma lacustre*; M91245), and the rainbow trout (*Oncorhynchus mykiss*; L29771). The sequence alignment was obtained from Russo, Takezaki, and Nei (1996), and Yang's (1997) PAML package was used for the parameter estimation. Only first and second codon positions were used in the estimation, because the substitution pattern for third codon positions is very different from that for first and second positions. Only 12 of the 13 mitochondrial protein-coding genes were used. We did not use the NADH-ubiquinone oxidoreductase subunit 6 (NADH6) gene, because it is encoded on the light chain, and its substitution pattern is quite different from those of the other 12 genes, which are encoded on the heavy chain of the DNA double helices. The shape parameter under the dG model (with eight rate categories) for this data set was estimated to be $\alpha = 0.37$. However, for simplicity, we used the continuous gamma (G) distribution with $\alpha = 0.5$ in our simulation.

The computer simulation of sequence evolution generated sequence data according to the model tree (with the clock) and the substitution model described above (see Zhang and Nei [1997] for a detailed description of the simulation). The simulated sequences were either 300 or 1,500 nt in length. We then computed L_0 (with the constraint of a clock) and L_1 (without the clock constraint) by using the known tree topology (fig. 3) and a given substitution model. The nucleotide substitution models used in the LRTs are listed in figure 1. All of these models have been commonly used in molecular evolutionary analyses. The program BASEML of the PAML package (Yang 1997) was used to compute the maximum-likelihood values in all of the simulations presented in this paper. The simulation was repeated 2,000 times for each model examined. The χ^2 test (with df = 6) was used in each replication to test if the null hypothesis of the molecular clock was rejected. The proportions of the 2,000 replications in which the null hypothesis was rejected at the 5% and 1% significance levels are denoted by $P_{5\%}$ and $P_{1\%}$. If the LRT is unbiased, $P_{5\%}$ and $P_{1\%}$ are expected to be equal to 0.05 and 0.01, respectively. However, the estimated $P_{5\%}$ and $P_{1\%}$ values may deviate from their corresponding expectations because of the limited sample size (2,000). Two-tail binomial tests were therefore performed to examine the statistical significance of these deviations. For example, if the hypothesis of the molecular clock is rejected at the 5% level in 64 of 2,000 replications (i.e., $P_{5\%} = 64/2{,}000 = 0.032$), then the probability of an event being as biased or more biased than observed is $2\sum_{m=0}^{64}\binom{2{,}000}{m}0.05^m 0.95^{2{,}000-m} = 0.00011$, where 0.05 is the expected value of $P_{5\%}$ when the LRT is unbiased. This result indicates that $P_{5\%}$ is significantly different from the expected value of 0.05, and therefore the LRT is biased. The power of the binomial test increases with the number of simulation replications.

The simulation results show that while some incorrect assumptions about the substitution parameters may not seriously affect the LRT of the molecular clock, some are detrimental to the test (table 3). The true substitution model under which the sequences were gener-

Table 3
Performance of the Likelihood Ratio Test of the Molecular Clock Under Different Substitution Models

	$100 \times P_{5\%}$		$100 \times P_{1\%}$	
MODELS USED	300 bp	1,500 bp	300 bp	1,500 bp
JC	3.20 (0.39)**	3.45 (0.41)**	0.50 (0.16)*	0.75 (0.19)
REV	3.35 (0.40)**	3.50 (0.41)**	0.45 (0.15)*	0.75 (0.19)
JC+dG	5.55 (0.51)	4.95 (0.49)	1.20 (0.24)	1.05 (0.23)
REV+dG	5.45 (0.51)	5.10 (0.49)	1.25 (0.25)	1.05 (0.23)

NOTE.—The standard errors of $P_{5\%}$ and $P_{1\%}$ are given in parentheses. Deviation of $P_{5\%}$ (or $P_{1\%}$) from 5% (or 1%) is examined by the binomial test. Significance level: * 0.05; ** 0.005.

Table 4
Performance of the Likelihood Ratio Test of Transition Bias Under Different Substitution Models

	$100 \times P_{5\%}$		$100 \times P_{1\%}$	
MODELS USED	300 bp	1,500 bp	300 bp	1,500 bp
JC vs. K80	5.05 (0.49)	9.50 (0.66)**	1.10 (0.23)	2.60 (0.36)**
F81 vs. HKY	5.00 (0.49)	4.55 (0.47)	0.80 (0.20)	0.90 (0.21)
JC+dG vs. K80+dG	7.40 (0.59)**	11.40 (0.71)**	1.35 (0.26)	3.50 (0.41)**
F81+dG vs. HKY+dG	5.90 (0.53)	5.50 (0.51)	1.00 (0.22)	1.30 (0.25)

NOTE.—Significance level: * 0.05; ** 0.005.

ated was REV+G. When the JC+dG or REV+dG model is used in the LRT of the molecular clock, the test is almost unbiased. For example, in the case of the JC+dG model with 300 nt, the molecular clock was rejected at the 5% level in 5.55% of the 2,000 replications, and the difference between the observed value of $P_{5\%} = 5.55\%$ and the expected value of 5% is not significant. But when among-site rate heterogeneity is ignored (e.g., the JC and REV models), the test becomes too conservative. For instance, in the case of the REV model with 300 nt, $P_{5\%} = 3.35\%$, which is significantly smaller than the expected value of 5%. This suggests that an incorrect hypothesis of the molecular clock may not be rejected as often as expected. This result seems to be understandable, because when among-site rate variation is ignored, the amount of evolution (branch lengths) is generally underestimated, particularly for long branches. Thus, rate variation among branches is diminished, and the molecular clock becomes difficult to reject. The simulation results suggest that consideration of rate variation among sites can increase the power of the molecular clock test and that the simple JC+dG model seems to be sufficient in the clock test to account for very complex substitution patterns. Furthermore, the test performs slightly better for longer sequences. The generality of the above conclusions, however, needs to be proved, because the present simulation is based on only one tree with a given set of branch lengths.

Test of Transition Bias

The rate of transitional nucleotide substitution (A↔G and T↔C) is often higher than that of transversional substitution (A↔T, A↔C, G↔T, and G↔C) in DNA sequence evolution. This phenomenon, known as the transition/transversion bias or transition bias, occurs because (1) transitional mutations occur more frequently than transversional mutations, and (2) transversions are more likely to change the encoded amino acids than transitions, such that the probability of fixation of a transversional mutation is lower than that for a transitional mutation under purifying selection (Li 1997). The transition bias of DNA evolution can be examined by an LRT. In this case, we have H_0: $\kappa = 1$, where κ is the ratio of the rate of transitional substitution to that of transversional substitution. To examine the influence of wrong models on the LRT of transition bias, I used the F81+G model of nucleotide substitution to generate sequence data which have no transition bias (see fig. 1), but I used other models in the LRT. The base frequencies used in the simulation were those observed from the first and second codon positions of the mitochondrial genes (table 2). The gamma parameter used was again 0.5.

The simulation results show that failure to take into account unequal base frequencies of the DNA sequences leads to rejection of the null hypothesis of no transition bias much more often than expected (see the results for the JC and JC+dG models), and the liberality of the test is enhanced when longer sequences are considered (table 4). For instance, in the case of the JC model versus the K80 model with 1,500 nt, the null hypothesis is rejected two to three times more often than expected. Consideration of the rate variation among sites appears to increase the power of the test, as previous studies suggest (Wakeley 1994; Yang, Goldman, and Friday 1995; Yang 1997), although the effect is not dramatic.

In order to determine the seriousness of the effects of ignoring base composition on the estimation of κ, we examined the distribution of the 2,000 κ values estimated under the K80 model (see fig. 1) with sequences of 1,500 nt in length. The mean κ was 1.048, with a standard error of the mean equal to 0.0015, indicating that κ was only slightly overestimated. This estimation bias can be explained intuitively as follows. Let π_T, π_C, π_A, and π_G be the frequencies of bases T, C, A, and G in a DNA sequence. Under the HKY model (see fig. 1), the number of transitional substitutions per site per unit time is

$$S = 2\kappa(\pi_T\pi_C + \pi_A\pi_G)u, \tag{7}$$

and the number of transversions is

$$\begin{aligned} V &= 2(\pi_T\pi_A + \pi_T\pi_G + \pi_C\pi_A + \pi_C\pi_G)u \\ &= 2(\pi_T + \pi_C)(\pi_A + \pi_G)u, \end{aligned} \tag{8}$$

where u is the substitution rate. Therefore,

$$\kappa = \frac{2(\pi_T + \pi_C)(\pi_A + \pi_G)S}{2(\pi_T\pi_C + \pi_A\pi_G)V} = \frac{2YS}{V} \tag{9}$$

where

$$Y = \frac{(\pi_T + \pi_C)(\pi_A + \pi_G)}{2(\pi_T\pi_C + \pi_A\pi_G)}. \tag{10}$$

To estimate κ, we need to know S, V, and Y. When a short period of time is concerned, S and V of equation (9) can be replaced by the observed numbers of transitions and transversions, respectively. $Y = 1$ when there is no base composition bias (i.e., $\pi_T = \pi_C = \pi_A = \pi_G = 0.25$). However, if the true Y is not equal to 1 but is

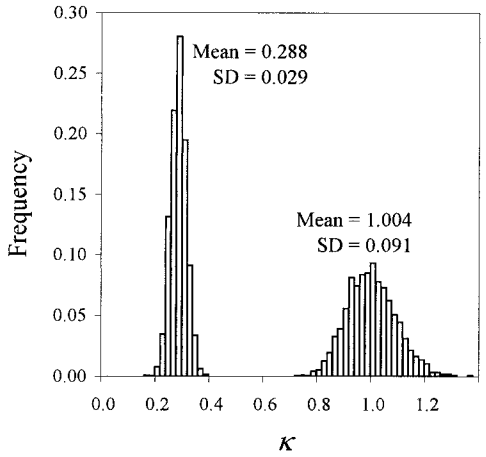

FIG. 4.—Distributions of the estimate κ (transition/transversion rate ratio) under an inadequate substitution model and under a correct substitution model. The substitution model in the simulation was F81, with base frequencies being $\pi_T = 0.50$, $\pi_C = 0.05$, $\pi_A = 0.05$, and $\pi_G = 0.40$. The distribution on the left is for estimates under an inadequate model (K80), whereas the distribution on the right is for estimates under a correct model (HKY).

assumed so in the likelihood estimation, the estimate of κ becomes biased. In our simulation, the *Y* value was actually 0.96, so when *Y* was erroneously assumed to be 1, as in the K80 model, κ was slightly overestimated.

From the above formulation, it can be seen that κ may be underestimated under the K80 model when the actual *Y* value is greater than 1. For example, when $\pi_T = 0.50$, $\pi_C = 0.05$, $\pi_A = 0.05$, and $\pi_G = 0.40$, $Y = 5.5$. We performed a simulation with 1,500 nt and the F81 model using these base frequencies. When the K80 model was used in the likelihood estimation, the mean value of the estimated κ was about 0.29 (fig. 4), which is substantially smaller than the true value of 1. Furthermore, the LRT of transition bias became extremely liberal. In 2,000 replications, all 2,000 LR values were greater than 170 and the mean was about 268, such that the null hypothesis of no transition bias was strongly rejected in each of the 2,000 replications when it was actually correct! When the correct model (HKY) was used, the mean value of the estimated κ approached 1 (fig. 4) and the test was unbiased.

Test of Among-Site Rate Variation

Substitution rate variation among sites is a general phenomenon for functional genes (e.g., Zhang and Gu 1998). The extent of rate variation was reported to be underestimated when complex substitution patterns were neglected (Yang, Goldman, and Friday 1995). I simulated sequence evolution by using the REV model described in table 2 and used a continuous gamma distribution of rate variation with the parameter $\alpha = 2.0$, which represents weak among-site rate variation. I then performed the LRT of H_0: $\alpha = 2.0$. The reason that I did not use $\alpha = \infty$ in the simulation is that the program I used (BASEML) searches for the maximum likelihood only for $\alpha < 40$ when the gamma distribution option is chosen so that the maximum likelihood of H_1 may not be found when the true α is infinity.

Table 5
Performance on the Likelihood Ratio Test of Rate Variation Among Sites Under Different Substitution Models

Models Used	$100 \times P_{5\%}$		$100 \times P_{1\%}$	
	300 bp	1,500 bp	300 bp	1,500 bp
JC	6.10 (0.54)*	8.25 (0.62)**	0.75 (0.19)	1.85 (0.30)**
REV	5.25 (0.50)	6.65 (0.56)**	1.10 (0.23)	1.55 (0.28)*

NOTE.—Significance level: * 0.05; ** 0.005.

The simulations show that when the simple JC model is used, the LRT becomes liberal, and this trend is clearer when the sequences are longer (table 5). Interestingly, even when the REV model was used, the test was still slightly liberal. This is probably due to the fact that the G distribution was used in generating the data, but the dG distribution with eight rate categories was used in the LRT. (Use of the dG model in the LRT substantially reduced the computational time and made the computer simulation feasible.) The medians, means, and standard deviations (square root of variance) of the estimates of the gamma shape parameter α under different models are given in table 6. When the JC+dG model was used, α was slightly overestimated. However, when the REV+dG model was used, α became underestimated. Again, use of the discrete gamma model in the estimation may have affected the results. It is interesting to observe that the estimates of α are generally closer to the true value when a simple model (JC+dG) is used than when a more realistic model (REV+dG) is used. The variance and coefficient of variation are, however, larger under the simple model.

LRTs with Parametric Bootstrapping

When the χ^2 approximation cannot be used in an LRT because of an unknown number of degrees of freedom or a very small sample size (Muse and Weir 1992; Goldman 1993), parametric bootstrapping (Monte Carlo simulation) is often used to generate the distribution of LR under the null hypothesis. The test is then performed by comparing the observed LR value from the real data with the null distribution of LR. In contrast to the usual nonparametric bootstrapping, the original data are discarded in the parametric bootstrapping; only the parameters estimated from the original data under the null hypothesis and an assumed substitution model are used (see Efron and Tibshirani 1993). In this case, if the assumed substitution model is inadequate, the estimates of

Table 6
Estimates of Gamma Shape Parameters Under Different Models

	JC+dG		REV+dG	
	300 bp	1,500 bp	300 bp	1,500 bp
Median	2.17	2.18	1.45	1.44
Mean	2.55	2.22	1.50	1.45
Standard deviation	1.72	0.37	0.37	0.16

NOTE.—The true substitution model was REV+G with the shape parameter $\alpha = 2$.

the substitution parameters will be biased. Thus, all sequence data simulated by using biased estimates and inadequate models will be insufficient to resemble the original data, such that the generated null distribution of LR will be distorted. This will probably seriously affect the result of the LRT.

To compare the performances of standard LRTs and those with parametric bootstrapping, a hypothesis that can be tested by both methods should be examined. Due to the large amount of computational time required by the LRT with parametric bootstrapping, I conducted simulations under one condition as a cursory examination of properties of this type of LRT. Sequence evolution was simulated according to the F81+G model with the base frequencies given in table 2 and $\alpha = 0.5$. The sequence length used was 300 nt. The null hypothesis of no transition bias was then tested under the assumption that the base frequencies were all equal. As presented in table 4, the standard LRT (with χ^2 test) is slightly liberal, with $P_{5\%} = 5.05\%$ and $P_{1\%} = 1.10\%$, neither being significantly different from the expected values of 5% and 1%, respectively. In contrast, the LRT with parametric bootstrapping is more liberal, with $P_{5\%} = 6.1\%$ and $P_{1\%} = 2.1\%$, both significantly different from the corresponding expectations. In this study, I conducted 2,000 simulation replications, and for each replication, 100 bootstrap pseudoreplications were used. Essentially the same result was obtained in another simulation of 1,000 replications, each with 300 bootstrap pseudoreplications. In short, the present simulation suggests that LRTs with parametric bootstrapping may be more sensitive to violations of assumptions of substitution models than are standard LRTs. This may largely limit the utility of this method in real data analysis. Nevertheless, it has to be pointed out that the above conclusion needs confirmation from more studies, because only limited conditions are considered in the present simulation.

Discussion

In both the numerical example and computer simulations, LRTs of various evolutionary hypotheses were found to be affected by the substitution models used, and the influence of a wrong model seems to be more serious in LRTs with parametric bootstrapping. Although only one tree topology with one set of tree branches was used in the simulation, it is already clear that use of inadequate substitution models may render the test too liberal or too conservative, depending on the hypotheses tested, the substitution models assumed, and the lengths of the sequences, among other things.

In the simulation, the true substitution pattern of the gene was known, and incorrect models were intentionally applied to examine their influence on LRTs. Since we now know that LRTs may be sensitive to violations of certain assumptions of evolutionary models, it is important to ask whether we can detect these violations if they happen and whether we can find more appropriate models. For instance, we have shown that among-site rate heterogeneity is an important factor in testing the molecular clock. We therefore examined whether the wrong assumption of no among-site rate heterogeneity can be rejected for the simulated data used in the LRT of the molecular clock. Note that the continuous gamma distribution with $\alpha = 0.5$ was used in generating the data. We found that the hypothesis of rate constancy among sites can be rejected (at the 1% significance level) by the LRT in all 2,000 simulation replications. This indicates that with this relatively high degree of rate variation among sites, the LRT can detect the wrong assumption of the model. Similarly, we found that the hypothesis of equal base frequencies can be rejected (at the 1% significance level) in about 96% of the 2,000 simulation replications when the mitochondrial base frequencies (table 2) are used as equilibrium frequencies in simulation. These results suggest that when a model is grossly wrong and there exists a more realistic model, LRTs can often reject the wrong model and choose the better one.

Nevertheless, the true substitution pattern of a gene is expected to be much more complex than the models commonly used in data analysis. While it does not seem to be necessary or possible to model every detail of the evolution of a gene, some aspects of gene evolution that are often neglected are almost sure to be important in certain LRTs. For instance, even the most complex substitution model that I used (REV+G) has an unrealistic assumption that all sites have an identical substitution matrix and that this matrix remains the same during all the evolutionary time concerned. This erroneous assumption will certainly affect the test of neutrality and will probably affect other LRTs as well. It should also be noted that the LRTs that we examined in the paper are relatively simple in the sense that many complex aspects of molecular evolution such as the structure of the codon table are not considered. Other LRTs, such as the tests of positive selection (Nielsen and Yang 1998) and host-parasite cospeciation (Huelsenbeck, Rannala, and Yang 1997), have been developed recently. These tests may be sensitive to factors that are not considered in the tests examined in the paper. For example, variations in rate of synonymous nucleotide substitution among different regions of a gene and among different species is not unusual (e.g., Alvarez-Valin, Jabbari, and Bernardi 1998; Deming et al. 1998). While this factor does not seem to be a significant problem for testing the equality of base frequencies, it will certainly affect the LRT of positive selection. It is possible that the overall performance of LRTs in everyday practice is worse than what has been shown in the simulation because of the complexity of the substitution process in reality. Thus, caution should be taken in interpreting results from LRTs.

I suggest that the performance of the LRT under inadequate substitution models be examined by simulation whenever a new application of the LRT is proposed. Particularly, information regarding the most important factors that might affect the test should be provided to users. If such information is lacking, users should perform the test under several different models and should not trust a test result with a marginal sig-

nificance level (e.g., 1%–5%) or a result that varies considerably with different models. Most importantly, more research on clarifying the nucleotide and amino acid substitution patterns of genes and proteins is needed, and more realistic models of molecular evolution should be built.

It should be noted that the performance of LRTs under inadequate substitution models is relatively easy to examine, because likelihood functions are explicitly dependent on substitution models. Although generally unknown, it is likely that statistical tests used in molecular evolutionary studies that are not based on likelihood (e.g., Rzhetsky and Nei 1995) are also subject to the same problems. It might be interesting to compare the power and robustness of LRTs and non-LRTs for a given hypothesis, such as the stationarity of base frequencies of homologous sequences.

Acknowledgments

I am grateful to Xun Gu, John Huelsenbeck, and Masatoshi Nei for helpful discussions and to Andy Clark, James Lyons-Weiler, Masatoshi Nei, Alex Rooney, Tom Whittam, and an anonymous referee for their comments on early versions of the manuscript. This work was supported by NIH and NSF research grants to Masatoshi Nei.

LITERATURE CITED

ALVAREZ-VALIN, F., K. JABBARI, and G. BERNARDI. 1998. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J. Mol. Evol. **46**:37–44.

CUNNINGHAM, C. W., H. ZHU, and D. M. HILLIS. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution **52**:978–987.

DEMMING, M. S., K. D. DYER, A. T. BANKIER, M. B. PIPER, P. H. DEAR, and H. F. ROSENBERG. 1998. Ribonuclease k6: chromosomal mapping and divergent rates of evolution within the RNase A gene superfamily. Genome Res. **8**:599–607.

EFRON, B., and R. J. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman and Hall, New York.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–372.

———. 1984. PHYLIP: phylogeny inference package. Verson 2.6. University of Washington, Seattle.

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

FOUTZ, R. V., and R. C. SRIVASTAVA. 1977. The performance of the likelihood ratio test when the model is incorrect. Ann. Stat. **5**:1183–1194.

GAUT, B. S., and P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. **12**:152–162.

GAUT, B. S., and B. S. WEIR. 1994. Detecting substitution-rate heterogeneity among regions of a nucleotide sequence. Mol. Biol. Evol. **11**:620–629.

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

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.

HUELSENBECK, J. P., and K. A. CRANDALL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. **28**:437–466.

HUELSENBECK, J. P., and B. RANNALA. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–232.

HUELSENBECK, J. P., B. RANNALA, and Z. YANG. 1997. Statistical tests of host-parasite cospeciation. Evolution **51**:410–419.

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.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

MUSE, S. V., and B. S. WEIR. 1992. Testing for equality of evolutionary rates. Genetics **132**:269–276.

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

———. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet. **30**:371–403.

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

RUSSO, C. A. M., N. TAKEZAKI, and M. NEI. 1996. Efficiencies of different genes and different tree-making methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. **13**:525–536.

RZHETSKY, A., and M. NEI. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. **12**:131–151.

STUART, A., and J. K. ORD. 1991. Kendall's advanced theory of statistics. Vol. 2. Oxford University Press, New York.

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

WAKELEY, J. 1994. Substitution rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. **11**:436–442.

YANG, Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. **42**:294–307.

———. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

YANG, Z., N. GOLDMAN, and A. FRIDAY. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44**:384–399.

ZHANG, J., and X. GU. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. Genetics **149**:1615–1625.

ZHANG, J., and M. NEI. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. **44**(Suppl. 1):S139–S146.

MARCY K. UYENOYAMA, reviewing editor

Accepted March 11, 1999