Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution

Maria Anisimova, Joseph P. Bielawski, and Ziheng Yang

Department of Biology, Galton Laboratory, University College London, London, England

The selective pressure at the protein level is usually measured by the nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$), with $\omega < 1$, $\omega = 1$, and $\omega > 1$ indicating purifying (or negative) selection, neutral evolution, and diversifying (or positive) selection, respectively. The ω ratio is commonly calculated as an average over sites. As every functional protein has some amino acid sites under selective constraints, averaging rates across sites leads to low power to detect positive selection. Recently developed models of codon substitution allow the ω ratio to vary among sites and appear to be powerful in detecting positive selection in empirical data analysis. In this study, we used computer simulation to investigate the accuracy and power of the likelihood ratio test (LRT) in detecting positive selection at amino acid sites. The test compares two nested models: one that allows for sites under positive selection (with $\omega > 1$), and another that does not, with the χ^2 distribution used for significance testing. We found that use of the χ^2 distribution makes the test conservative, especially when the data contain very short and highly similar sequences. Nevertheless, the LRT is powerful. Although the power can be low with only 5 or 6 sequences in the data, it was nearly 100% in data sets of 17 sequences. Sequence length, sequence divergence, and the strength of positive selection also were found to affect the power of the LRT. The exact distribution assumed for the ω ratio over sites was found not to affect the effectiveness of the LRT.

Introduction

Detecting positive Darwinian selection is a critical aspect of understanding the mechanisms of molecular evolution. Existing tests proposed in population genetics (see Wayne and Simonsen [1998] for a review) are powerful enough to reject the strictly neutral model. However, such tests are often not sufficient to distinguish different forms of natural selection or to detect adaptive molecular evolution (Yang and Bielawski 2000). A powerful method for detecting positive selection is through comparison of synonymous and nonsynonymous substitution rates. Selective pressure at the protein level is measured by $\omega = d_N/d_S$, where d_N and d_S are nonsynonymous and synonymous substitution rates, respectively. If amino acid changes are advantageous, they will be fixed at a higher rate than synonymous changes, with $d_N > d_S$. Thus, a significantly higher nonsynonymous substitution rate ($\omega > 1$) is evidence of adaptive molecular evolution. If amino acid changes are deleterious, purifying selection will reduce their fixation rate, such that $d_N < d_S$ and $\omega < 1$. Neutral mutations result in $\omega = 1$, as selection on the protein has no effect on fitness.

Until recently, cases of positive selection have been difficult to demonstrate. A large-scale database search performed by Endo, Ikeo, and Gojobori (1996) identified only 17 out of 3,595 genes that might have undergone adaptive evolution. Endo, Ikeo, and Gojobori (1996) considered a gene to be under positive selection if the average d_N was greater than d_S in more than half of the pairwise sequence comparisons. This approach computes the ω ratio as an average over both amino acid sites and time; although popular, it has little power. For example, Crandall et al. (1999) found that the approach of pairwise comparison failed to detect positive selection in the protease gene of HIV-1 despite clear evidence of parallel evolution. Crandall et al. (1999) suggested that the ω ratio averaged over sites was a poor indicator of positive selection. Indeed, the assumption that all sites in a sequence are under equal selective pressure is unrealistic. Typically, adaptive evolution occurs at only a few sites, as most amino acids in a protein are under structural and functional constraints with d_N, and hence ω, close to 0 (e.g., Li 1997). Thus, calculating ω as an average over all amino acid sites substantially reduces the power to detect positive selection.

Codon-based models recently developed by Nielsen and Yang (1998) and Yang et al. (2000) account for variation of the ω ratio among sites. They are implemented in the maximum-likelihood (ML) framework and can be used (1) to test for the presence of codon sites affected by positive selection and (2) to identify such sites when they exist. The idea is to allow the ω ratio to take values from a number of discrete site classes or from a continuous distribution. The application of such models has led to detection of positive selection in many genes for which it has not previously been suggested. For example, using the ML model of Nielsen and Yang (1998), Zanotto et al. (1999) detected positive selection in the *nef* gene of HIV-1, whereas in earlier studies of the same gene, the average ω ratio over sites provided no evidence for adaptive evolution (Plikat, Nieselt-Struwe, and Meyerhans 1997; da Silva and Huges 1998). Yang et al. (2000) detected diversifying positive selection in six out of ten genes from nuclear, mitochondrial, and viral genomes, while the ω ratio averaged over sites was less than one in all of those genes. Similarly, in an analysis of the fertility gene *DAZ*, Agulnik et al. (1998) found similar average synonymous and nonsynonymous rates and similar rates at the three codon positions and thus concluded that the *DAZ* gene family was not under any selective constraint. However,

Key words: positive selection, nonsynonymous/synonymous rate ratio, likelihood ratio test (LRT), molecular adaptation, type I error, type II error.

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

Mol. Biol. Evol. 18(8):1585–1592. 2001

using models of variable ω ratios, Bielawski and Yang (2001) found that most amino acids in the *DAZ* gene were under strong functional constraints, while a few sites were under diversifying selection.

While the new models have been successfully applied to real data, the accuracy and power of the likelihood ratio test (LRT) have not been examined. Here, we use computer simulation to investigate the accuracy and power of the LRT in detecting positive selection. In cases considered here, the LRT statistic does not follow the χ^2 distribution due to the so-called boundary problem. This problem arises because the null hypothesis is equivalent to the alternative hypothesis with some parameters fixed at the boundary of the parameter space. The sample size (i.e., the sequence length) also affects the distribution of the LRT statistic; the χ^2 approximation is asymptotic and reliable for large samples only (e.g., Silvey 1970, pp. 112−114). We attempted to characterize the minimum sample size required for the χ^2 approximation to be acceptable. Furthermore, we examined how the power of the LRT depends on the sequence divergence, the sequence length, the number of taxa, and the strength of positive selection. Finally, we tested the sensitivity of the LRT to misspecification of the ω distribution among sites.

Theory and Methods

Codon Substitution Models for Detecting Positive Selection at Sites

The Markov model of codon substitution proposed by Goldman and Yang (1994; see also Muse and Gaut 1994) was modified recently to account for heterogeneous ω ratios among sites (Nielsen and Yang 1998; Yang et al. 2000). Here, we present an overview of these models. Let h denote a site in the sequence and N denote the number of codons in the sequence ($h = 1, 2, \ldots, N$). The relative instantaneous substitution rate from codon i to codon j ($i \neq j$) at site h is given by

$$q_{ij}^{(h)} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three nucleotide positions} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by one synonymous transversion} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by one synonymous transition} \\ \omega^{(h)}\pi_j, & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transversion} \\ \omega^{(h)}\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transition,} \end{cases} \quad (1)$$

where π_j is the equilibrium frequency of codon j, κ is the transition/transversion rate ratio, and $\omega^{(h)}$ is the d_N/d_S ratio at site h. The transition probability matrix over time t is given by $P(t) = e^{Qt}$, where $Q = \{q_{ij}^{(h)}\}$ (e.g., Lio and Goldman 1998).

Following the recommendations of Yang et al. (2000), we consider the following models of ω ratio distribution among sites: M0 (one-ratio), M3 (discrete), M7 (beta), and M8 (beta&ω) (see table 1). M0 (one-ratio) assumes one ω ratio for all sites, so $\omega^{(h)} = \omega$ for any h. Model M3 (discrete) classifies sites in the sequence into K discrete classes, with both the ω ratios $\omega_0, \omega_1, \ldots, \omega_{K-1}$ and the proportions $p_0, p_1, \ldots, p_{K-1}$ estimated from the data. Three classes ($K = 3$) were used in this paper. Under model M7 (beta), the ω ratio varies according to the beta distribution $B(p, q)$ with parameters p and q. The beta distribution is bounded within the interval (0, 1) and thus does not allow for positively selected sites. Model M8 (beta&ω) adds a discrete ω class to the beta model to account for sites under positive selection with $\omega > 1$. A proportion p_0 of sites have ω drawn at random from the beta distribution $B(p, q)$, while the rest (with proportion $p_1 = 1 - p_0$) have the same ratio ω. M0 (one-ratio) and M3 (discrete) are nested models and can be compared using an LRT. Similarly, models M7 (beta) and M8 (beta&ω) are nested and can be compared using an LRT.

Accuracy of the LRT

The type I error occurs if the null hypothesis H_0 is rejected when it is true. A test is accurate if the type I error rate is not greater than the chosen significance level α. If H_0 holds, the LRT statistic $2\Delta\ell$ (twice the log likelihood difference) can be approximated by the χ^2 distribution with the degree of freedom ν equal to the difference in the number of free parameters in the two nested models (e.g., Stuart, Ord, and Arnold 1999, p. 241). This, however, is only true for large samples and under certain regularity conditions. For example, if the null model H_0 is equivalent to an alternative model H_1 with some parameters fixed at the boundary of the parameter space, the regularity conditions are not satisfied and the χ^2 approximation is not expected to apply. Such is the case with the LRTs considered here. For example, M0 (one-ratio) is a special case of M3 (discrete) by constraining two of the five free parameters in M3 (p_0 and p_1) to 0. This breaches the regularity conditions, as $p_0 = 0$ and $p_1 = 0$ lie on the boundary of the parameter space. Moreover, parameters ω_0 and ω_1 become undefined when $p_0 = p_1 = 0$. Comparison between M7 and M8 poses a similar problem. The transformation from M8 to M7 forces the parameter ω to become inestimable by fixing p_1 at 0, which is on the boundary of the parameter space. Therefore, in neither of our cases is the LTR statistic expected to follow the χ^2 distribution.

We assessed the accuracy of the test by simulating replicate data sets under the null hypothesis and analyzing them using both the null and the alternative hypotheses. The distribution of the test statistic $2\Delta\ell$ among replicates was then compared with the χ^2_ν distribution, with $\nu = 4$ for the M0-M3 comparison and $\nu = 2$ for the M7-M8 comparison (table 1). The settings of the simulation experiments are summarized in table 2. Trees used to simulate the data are shown in figure 1. We do not assume the molecular clock (rate constancy over time), and all trees are unrooted. While the d_N/d_S rate ratio ω is the same among branches, the total rate, measured by the expected number of nucleotide substitutions per codon, varies among branches. We used codon frequencies empirically estimated from 17 vertebrate β-globin genes and

Table 1
Models of Variable ω Ratios Among Sites Used to Investigate the Accuracy and Power of the Likelihood Ratio Test

Model	Description	Free Parameters	No. of Free Parameters
M0 (one-ratio).....	One ω ratio for all sites	ω	1
M3 (discrete).....	$K = 3$ site classes	ω_0, ω_1, ω_2, p_0, p_1	5
M7 (beta).....	$\omega \sim B(p, q)$	p, q	2
M8 (beta&ω).....	Proportion p_0 of sites $\sim B(p, q)$, p_1 of sites from discrete ω class	p, q, p_0, ω	4

from 23 HIV-1 *pol* genes (see table 2). The vertebrate β-globin gene is biased against adenine at third codon positions, whereas the HIV-1 *pol* gene is G-C rich at third positions. Simulation parameters were taken to represent the range of estimates from real data (Yang et al. 2000). We simulated sequences of $N = 100$ and 500 codons using trees of $T = 5$, 6, or 17 taxa. Sequence divergence was measured by the tree length S, the expected number of nucleotide substitutions per codon along the tree, and three values (''low,'' ''medium,'' and ''high'') were used for each tree (table 2).

Power of the LRT

The type II error of a test occurs if the test fails to reject H_0 when it is false. The power of a test is defined as 1 − type II error rate and is equal to the probability of rejecting H_0 given that H_0 is wrong and that the alternative hypothesis H_1 is correct. To examine the power of the LRT, we simulated replicate data sets under H_1 and analyzed them using both H_0 and H_1 to see whether H_0 was rejected by the LRT. We considered two measures. First, we counted the replicates for which positive selection was indicated by the parameter estimates in the alternative model, and we denote the proportion of such replicates by P_+. Formally, $P_+ = \Pr(\text{there exists an } \hat{\omega} > 1 \,|\, H_1 \text{ is true})$, where $\hat{\omega}$ is the ML estimate of any of the parameters ω_i ($i = 0, 1, 2$) under M3 (discrete) or of the single ω parameter in model M8 (beta&ω) (see table 1). The second measure is more stringent and requires that positive selection is indicated by the parameter estimates in the alternative model and that the LRT is significant. We denote the proportion of such replicates by P_{+S} and refer to it as the power of the LRT. As P_{+S} depends on the significance level α, we also use the notation $P_{+S,\alpha}$. In other words, $P_{+S,\alpha} = \Pr(\text{there exists } \hat{\omega} > 1 \text{ and } 2\Delta\ell > \chi^2_{\nu,\alpha} \,|\, H_1 \text{ is true})$. Note that $P_+ \geq P_{+S}$.

We also investigated the sensitivity of LRTs to misspecification of the distribution of the ω ratio among sites. We simulated data sets under M3 (discrete) and analyzed them using M7 (beta) and M8 (beta&ω). Similarly, we simulated data sets under M8 (beta&ω) and analyzed them using M0 (one-ratio) and M3 (discrete). Parameter settings used are listed in table 3. As before, we used a number of parameter combinations to represent a variety of real data situations.

All sequence data sets were generated using the evolver program. Log likelihood values were calculated with the codeml program. Both programs are from the PAML package (Yang 2000).

Results

Accuracy

Results obtained from simulations examining the accuracy of the LRTs are presented in table 2. In experiments A–C, data were simulated under M0 (one-ratio) and analyzed using M0 (one-ratio) and M3 (dis-

Table 2
Type I Error Rate: Numbers of Cases out of 100 for Which the Null Hypothesis Was Rejected at the α = 1% (5%) Significance Levels

Experiment	Simulation	Analysis	Simulation Parameters				Type I Error at α = 1% (5%)	
			T	κ	ω	S	$N = 100$	$N = 500$
A.....	M0	M0 & M3	6	2	0.40	0.11	0 (0)	0 (0)
						1.1	0 (0)	0 (0)
						11	0 (0)	0 (0)
B.....	M0	M0 & M3	17	2	0.40	2.11	0 (0)	0 (1)
						8.44	0 (0)	0 (1)
						16.88	0 (1)	0 (0)
C.....	M0	M0 & M3	5	5	0.25	0.91	0 (0)	0 (0)
						9.1	0 (0)	0 (1)
						18.2	0 (1)	2 (3)
D.....	M7	M7 & M8	6	2	$p = 0.41$	0.11	NA	0 (0)
					$q = 1.10$	1.1	NA	1 (5)
						11	NA	1 (4)

Note.—Codon frequencies from the vertebrate β-globin gene were used in all experiments except experiment C, in which those from the HIV-1 *pol* gene were used. In experiment D, simulation was not conducted for $N = 100$.

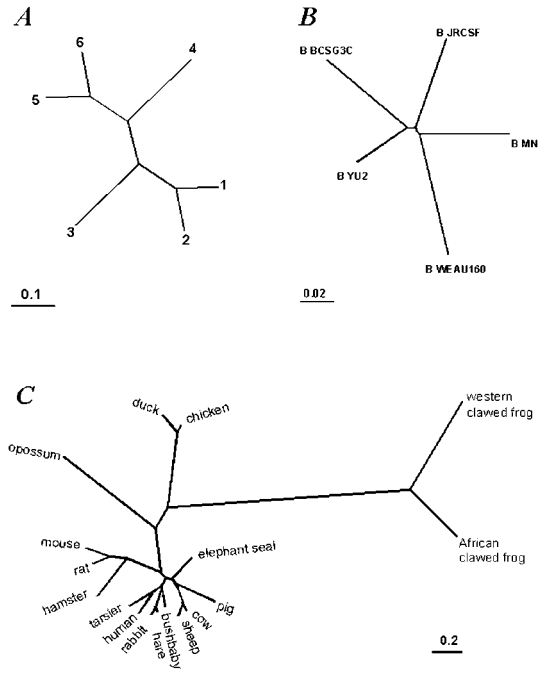

FIG. 1.—Tree topologies used in the simulations. *A*, Artificial six-taxon tree. *B*, Five-taxon subtree from a tree constructed for 23 HIV-1 *pol* gene sequences (Yang et al. 2000). *C*, A β-globin tree for 17 vertebrate species from Yang et al. (2000).

crete), with χ^2_4 used to test significance. If χ^2_4 were the correct null distribution, H_0 would be rejected (type I error) in 5% of the replicates at the $\alpha = 0.05$ significance level and in 1% of the replicates at $\alpha = 0.01$. However, the regularity conditions for the χ^2 approximation are not satisfied. Results of table 2 (experiments A–C) suggest that the null hypothesis was rejected less often than allowed by the significance level. In most cases, the estimated type I error rate was 0 for $\alpha = 0.05$ (table 2). Even at $\alpha = 0.1$, the estimated probability of rejecting the null hypothesis never exceeded 6% and was often much lower than the expected 10% (results not shown). Thus, use of χ^2_4 to compare M0 and M3 makes the LRT conservative.

The shapes of the $2\Delta\ell$ distribution were similar for all parameter combinations in experiments A–C, in which the LRT compared M0 (one-rate) against M3 (discrete). One example is shown in figure 2*A* for the combination $N = 500$ and $S = 1.1$ in experiment A. The simulated distribution has a skewed L-shape, while χ^2_4 has a peak in the middle with a long tail to the right. The two distributions are very different. At very low sequence divergence ($S = 0.11$ in experiment A), there was a substantially higher peak near $2\Delta\ell = 0$, such that M0 was rejected even less often and the LRT was even more conservative. Short sequences had an effect similar to that of low divergence, and the LRT was more

Table 3
Power of the Likelihood Ratio Test (LRT): Numbers of Replicates out of 100 in Which Positive Selection Was Indicated by Parameter Estimates (P_+) or Detected by the LRT at the 1% ($P_{+s,0.01}$) and 5% ($P_{+s,0.05}$, in parentheses) Significance Levels

Experiment	Simulation	Analysis	Simulation Parameters				P_+		$P_{+s,0.01(0.05)}$	
			T	κ	ω Distribution	S	N = 100	N = 500	N = 100	N = 500
1.....	M3	M0 & M3	6	2	$\omega_0 = 0.018$, $p_0 = 0.386$;	0.11	33	48	0 (1)	4 (8)
					$\omega_1 = 0.304$, $p_1 = 0.535$;	1.1	73	95	54 (66)	95 (95)
					$\omega_2 = 1.691$, $p_2 = 0.079$	11	52	80	52 (52)	80 (80)
						55	28	11	28 (28)	11 (11)
2.....	M3	M0 & M3	17	2	Same as in experiment 1	0.38	61	80	10 (17)	66 (72)
						2.11	93	100	91 (92)	100 (100)
						8.44	99	100	99 (99)	100 (100)
						16.88	99	99	99 (99)	99 (99)
						105.5	31	58	31 (31)	58 (58)
3.....	M3	M0 & M3	6	2	Same as in experiment 1,	0.11	39	82	3 (4)	43 (59)
					except **$\omega_2 = 4.739$**	1.1	85	100	81 (83)	100 (100)
						11	48	43	48 (48)	43 (43)
4.....	M3	M0 & M3	5	5	$\omega_0 = 0.049$, $p_0 = 0.838$;	0.91	72	97	28 (43)	93 (95)
					$\omega_1 = 0.849$, $p_1 = 0.142$;	9.1	78	97	78 (78)	97 (97)
					$\omega_2 = 4.739$, $p_2 = 0.020$	18.2	70	83	70 (70)	83 (83)
5.....	M8	M7 & M8	6	2	$p = 0.572$, $q = 2.172$;	0.11	60	58	0 (0)	0 (2)
					$p_0 = 0.943$;	1.1	78	96	10 (21)	65 (77)
					$\omega = 2.081$, $p_1 = 0.057$	11	55	60	2 (10)	28 (36)
6.....	M3	M7 & M8	6	2	Same as in experiment 1	0.11	60	56	0 (0)	1 (3)
						1.1	79	94	4 (13)	48 (69)
						11	61	41	0 (1)	1 (4)
7.....	M8	M0 & M3	6	2	Same as in experiment 5	0.11	37	53	1 (2)	5 (11)
						1.1	78	96	69 (72)	96 (96)
						11	51	35	51 (51)	35 (35)

NOTE.—Codon frequencies from the vertebrate β-globin gene were used in all experiments except experiment 4, where those from the HIV-1 *pol* gene were used. Simulation parameters representing positive selection are indicated in bold.

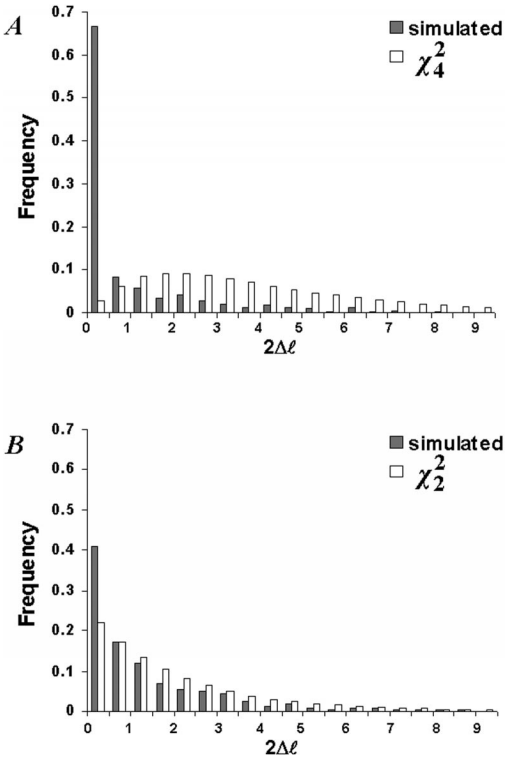

FIG. 2.—Comparison of the χ^2 distribution with the distribution of the likelihood ratio test (LRT) statistic $2\Delta\ell$ in 500 simulated replicates. *A*, The LRT compares M0 (one-ratio) and M3 (discrete) for $N = 500$ and $S = 1.1$ (table 2, experiment A). *B*, The LRT compares M7 (beta) and M8 (beta&ω) for $N = 500$ and $S = 1.1$ (table 2, experiment D).

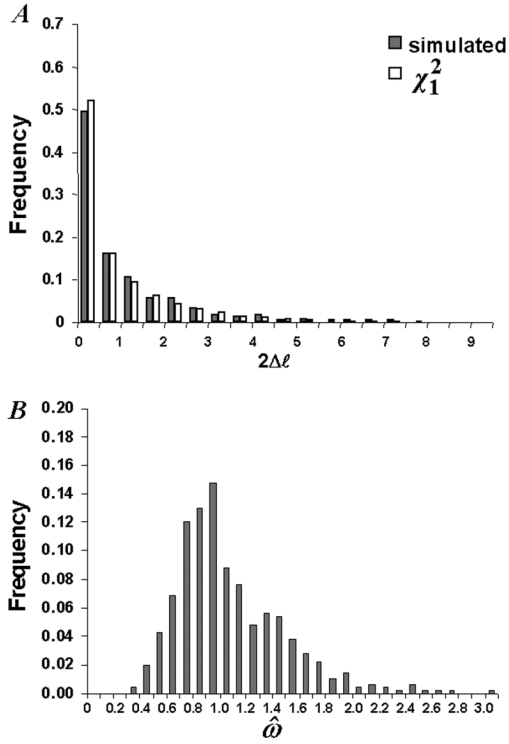

FIG. 3.—Accuracy of the asymptotic theory for the likelihood ratio test of H_0: $\omega = 1$ against H_1: $\omega \neq 1$. One ω ratio (model M0) is assumed for all sites in both H_0 and H_1. Five hundred data sets were simulated using parameters taken from experiment A of table 2, except that $\omega = 1$. The six-taxon tree of figure 1*A* and the codon frequencies from the vertebrate β-globin gene were used. The tree length is $S = 1.1$ substitutions per codon along the tree. The sequence length is $N = 50$ codons. *A*, Comparison of χ^2_1 with the simulated distribution of $2\Delta\ell$. The two distributions are not significantly different from each other. *B*, The distribution of maximum-likelihood estimates of ω under H_1.

conservative in data sets of 100 codons than in data sets of 500 codons (results not shown). The number of taxa did not appear to affect the shape of the distribution.

We simulated data sets under M7 (beta) in order to check whether the χ^2_2 approximation was reliable for comparing M7 (beta) and M8 (beta&ω). ML estimation under M7 and M8 is time-consuming; hence, only three parameter combinations were used (experiment D in table 2). For $S = 0.11$, M7 was never rejected at $\alpha = 0.05$, whereas for $S = 1.1$ and $S = 11$, M7 was rejected approximately as often as expected from the significance level α when $\alpha = 0.01$, 0.05, and 0.1. Figure 2*B* compares the distribution of the $2\Delta\ell$ statistic with χ^2_2 for the combination $N = 500$ and $S = 1.1$. The match is not good, and the simulated distribution is left-skewed. Therefore, use of the χ^2_2 makes the LRT conservative. Furthermore, the LRT was even more conservative for data sets of highly similar sequences ($S = 0.11$), as in the comparison of M0 (one-ratio) and M3 (discrete).

The reliability of the χ^2 approximation could have been affected by both the boundary problem and a small sample size. To distinguish between these two factors, we conducted a simple experiment free from the boundary problem. One ω ratio was assumed for all sites (M0), and the hypothesis H_0: $\omega = 1$ was tested against the alternative H_1: $\omega \neq 1$. The LRT statistic $2\Delta\ell$ was compared with χ^2_1. The tree in figure 1*A* was used, and the parameters (with the exception of ω) were the same as in experiment A (table 2). The distribution of $2\Delta\ell$ fitted the expected χ^2_1 distribution for all values of *S* and *N*. Figure 3*A* shows one case where the tree length $S = 1.1$ and the sequence length was only $N = 50$ codons. It is remarkable that the χ^2 distribution appears reliable for such short sequences. An equally good fit was observed for $N = 100$. Data sets of 50 codons with $S = 0.11$ were not analyzed, as such data carry little information and cause convergence problems. These results are compatible with those of Zhang (1999), who found in nucleotide-based simulations that the χ^2 approximation is reliable in fairly small data sets. Besides the χ^2 approximation to the LRT statistic, asymptotic theory also predicts that ML estimates of parameters are normally distributed (e.g., Stuart, Ord, and Arnold 1999, pp. 57–59). For $N = 50$, the distribution of $\hat{\omega}$ was left-skewed (fig. 3*B*), and in 47% of the replicates, $\hat{\omega}$ was greater than 1. The mean of the distribution was 1.09, indicating that the ML estimate involves a positive bias in small samples (Yang and Nielsen 2000). This pattern was found to be typical for small samples. With an increase of *N*, the distribution looked much more concentrated and symmetrical. Compared with the χ^2 approximation to the LRT statistic, the normal approximation to ML parameter estimates appeared to require larger samples to be reliable.

To examine the performance of the LRT on a neutral gene, we also applied the LRT comparing M0 and M3 to data sets simulated under M0 (one-rate) with $\omega = 1$. The parameter settings were the same as in figure 3 except that the sequence length was $N = 500$. In 72% of the replicates, estimates of at least one of the ω ratios under M3 were greater than 1, indicating positive selection. However, in most of them, the LRT was insignificant, and the type I error rate was only 0.004 at $\alpha = 0.05$. Thus, the LRT was reliable.

Power Analysis

Results obtained from simulations examining the power of the LRT are summarized in table 3. In experiment 1, we simulated data under M3 (discrete) using a six-taxon tree and analyzed them using M0 (one-ratio) and M3 (discrete). Both P_+ (probability of parameter estimates indicating positive selection) and P_{+s} (power of the LRT) were consistently higher when $N = 500$ than when $N = 100$. This effect of the sequence length was expected. The level of sequence divergence had a significant effect on the power of the test. At low sequence divergence ($S = 0.11$), ML parameter estimates under M3 suggested positive selection (P_+) in 33 data sets for $N = 100$ and in 48 data sets for $N = 500$ (table 3). However, in only a few of these cases was the evidence statistically significant (P_{+s}). For example, at $\alpha = 0.05$, the LRT was significant in only one case for $N = 100$ and in only eight cases for $N = 500$ (table 3). Note that $S = 0.11$ means that the sequences are highly similar, with 2.4% of total divergence along the tree at a nonsynonymous site ($d_N = 0.024$) and 7.8% of divergence at a synonymous site ($d_S = 0.078$). The transformation from tree length S to d_N and d_S can be made using the relationships $S = 3d_S p_S + 3d_N(1 - p_S)$, and $d_N/d_S = \bar{\omega}$ (e.g., Yang and Nielsen 2000). Here, the average ω ratio $\bar{\omega} = 0.018 \times 0.386 + 0.304 \times 0.535 + 1.691 \times 0.079 = 0.303$ (see table 3), and the proportion of synonymous sites is $p_S = 23.84\%$ for the vertebrate β-globin gene (Yang et al. 2000). Increasing sequence divergence to the intermediate level ($S = 1.1$) yielded a substantial increase in both P_+ and P_{+s}. For example, with $N = 500$, parameter estimates in $P_+ = 95\%$ of replicates suggested positive selection, and in all of them, the LRT was significant at the 1% level ($P_{+s,0.05} = P_{+s,0.01} = 95\%$) (table 3). The power decreased when S was increased to 11 (e.g., for $N = 500$, $P_+ = 80\%$ and $P_{+s,0.05} = 80\%$). At $S = 55$ nucleotide substitutions per codon, both P_+ and P_{+s} decreased dramatically (e.g., for $N = 500$, $P_+ = 11\%$ and $P_{+s,0.05} = 11\%$). Note that $S = 55$ represents unrealistically high sequence divergence, with $d_N = 11.8$ substitutions per nonsynonymous site and $d_S = 39.1$ substitutions per synonymous site along the tree. In summary, the power increased with increasing S, peaked at a medium level of S, and fell when sequences became highly divergent.

In experiment 2, we examined the effect of increasing the number of taxa to 17 (table 3). Here, P_{+s} was very high for most values of S and N. For example, even for the short sequences ($N = 100$) of rather low divergence ($S = 2.11$), ML estimates suggested positive selection in 93 data sets, with most of these cases being statistically significant ($P_{+s,0.01} = 91\%$). The LRT reached full power ($P_{+s} = 100\%$) for long sequences and realistic S in the range 2.11–8.44. As in experiment 1, the power increased with the initial increase of S, peaked at a medium level of S, and thereafter decreased with a further increase of S. For example, increasing S to an unrealistically high value ($S = 105.5$) for the short sequences ($N = 100$) resulted in $P_+ = 31\%$ and $P_{+s,0.05} = 31\%$.

Experiment 3 examined the influence of the strength of positive selection; ω_2 was increased from 1.69 in experiment 1 to 4.74 (table 3). As expected, there was a rise in the power of the LRT as compared with experiment 1. For every combination of S and N, the power in experiment 3 was higher than the corresponding result in experiment 1. Once again the power was low for either very similar or highly divergent sequences and was highest at intermediate levels of sequence divergence (around $S = 1.1$). As before, increasing sequence length from 100 to 500 yielded an increase in the power.

Experiment 4 examined the power of the LRT using the tree topology, simulation parameters, and codon frequencies derived from the HIV-1 *pol* gene (Yang et al. 2000) (table 3). As before, the power of the LRT was higher for longer sequences. Moreover, the power increased with the increase of S, peaked, and then decreased with a further increase in S. However, the level of sequence divergence at which the power began to fall differed from previous experiments. To enable a qualitative comparison, we used the average number of nucleotide changes per codon per branch as a relative measure of sequence divergence. This is $S/(2T - 3)$, where $2T - 3$ is the number of branches of an unrooted tree of T taxa. Unlike experiment 1, in which the highest power was observed at the medium level of sequence divergence ($S = 1.1$ and $T = 6$, or $S/(2T - 3) = 0.12$), here the highest power was obtained for relatively divergent data sets ($S = 9.1$ and $T = 5$, or $S/(2T - 3) = 1.3$). Hence, the optimal sequence divergence depends on the properties of the data and appears to be within the medium-to-high range.

In experiment 5, we simulated data under M8 (beta&ω) and analyzed them with M7 (beta) and M8 (beta&ω) (table 3). Although $\hat{\omega}$ derived from M8 often suggested positive selection, the power of the LRT was substantially lower than in experiment 1. For example, when $N = 500$ and $S = 1.1$, the power was $P_{+s,0.05} = 95\%$ in experiment 1 but only 77% in experiment 5. This difference is due to the fact that M0 is less realistic than M7 and easier to reject (see below).

In experiment 6, we examined whether the LRT was sensitive to the true distribution of ω by simulating data under M3 (discrete) and analyzing them with M7 (beta) and M8 (beta&ω) (table 3). The results were compared with those of experiment 1, where the data were analyzed with M0 and M3. The null model M0 was rejected much more frequently than the null model M7. For example, for the combination $N = 500$ and $S = 1.1$, the power was $P_{+s,0.01} = 100\%$ in experiment 1 and 48% in exper-

iment 6. Comparison between M7 and M8 is clearly a more stringent test of positive selection than comparison between M0 and M3. In contrast to P_{+s}, P_+ was often higher in experiment 6 than in experiment 1 except for the combination $N = 500$ and $S = 11$ (table 3). In sum, parameter estimates under M8 tend to suggest positive selection more often than M3, but the LRT based on M8 is significant less often than the LRT based on M3.

In experiment 7, we simulated data under M8 (beta&ω) and analyzed them using M0 (one-ratio) and M3 (discrete) (table 3). The results were compared with those of experiment 5, in which the data were analyzed using models M7 (beta) and M8 (beta&ω). We observed the same pattern as in the comparison between experiments 1 and 6. First, the null model M0 (one-ratio) was rejected more frequently than the null model M7 such that the power P_{+s} was always higher for the LRT comparing M0 and M3 than for the LRT comparing M7 and M8. For example, for $N = 500$ and $S = 1.1$, the power was $P_{+s,0.01} = 100\%$ in experiment 7 and 65% in experiment 5 (table 3). Second, the proportion of replicates in which positive selection was indicated by parameter estimates (P_+) was generally higher under M8 than under M3 (table 3).

Discussion

Accuracy of the χ^2 Approximation

If the type I error rate of a test is greater than α, the test is liberal and unreliable. If the type I error rate is less than α, the test is conservative and might lack power. It would be best to use the correct distribution of the LRT statistic $2\Delta\ell$ under the null hypothesis, or its close approximation, as then the type I error rate would match the significance level α. However, finding such a distribution for the two LRTs considered in this paper is problematic, mainly because of the boundary problem.

A number of special cases of LRTs under nonstandard conditions are discussed in Self and Liang (1987), which remains the latest reference on this issue. If only one parameter is on the boundary of the parameter space, the LRT statistic is approximately distributed as a mixture $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ if no other parameter is tested (case 5 of Self and Liang 1987). Here, χ_0^2 is the distribution that takes the value 0 with probability 1. An example is the comparison of the one-rate and gamma-rates models of among-sites rate variation. In this case, the null model (one-rate) is equivalent to fixing the shape parameter α of the gamma distribution at infinity (Yang 1996). Recent simulations (Goldman and Whelan 2000; Ota et al. 2000) showed that the LRT statistic fits the above mixture distribution very well even when the sample size is not very large. However, increasing the number of boundary parameters complicates the case and, in some situations, might cause the LRT statistic not to be expressible as a mixture of χ^2 distributions (e.g., case 8 of Self and Liang 1987). Moreover, the existence of a consistent ML estimator is one of the main assumptions for the LRT statistic to asymptotically converge to the χ^2 or its mixture distributions (Self and Liang 1987). In the LRTs considered in this paper, some parameters are not estimable, so none of the known distributions or their mixtures are expected to apply.

Consequently, we used χ_4^2 to compare M0 (one-ratio) and M3 (discrete), and we used χ_2^2 to compare M7 (beta) and M8 (beta&ω), as suggested by Yang et al. (2000). This approach makes the LRT conservative and leads to loss of power. This might be particularly important for data sets of highly similar sequences, as failure to detect positive selection might be due to the lack of power of the LRT. Note that when we examined the accuracy of the LRT (table 2), we considered the statistic $2\Delta\ell$ only, but when we examined the power of the test P_{+s} (table 3), we further required that parameter estimates in the alternative model (M3 or M8) suggested positive selection. Thus, the LRTs used in detecting positive selection as examined in table 3 are even more conservative than the results of table 2 suggest.

Besides the boundary problem, the χ^2 approximation can also be affected by insufficient sample sizes. However, our simulation with no boundary problem, as well as previous studies (e.g., Whelan and Goldman 1999; Zhang 1999), suggests that even with relatively short sequences (e.g., with 50 codons), the distribution of $2\Delta\ell$ fits the χ^2 quite well. Hence, analysis of short sequences appears feasible, although it might be difficult to get significant results. We should note that when the χ^2 approximation is unreliable, Monte Carlo simulation can be used to obtain the correct null distribution (Goldman 1993).

Power of LRT

Our simulations show several patterns of the power function, all of which are intuitively justified. Longer sequences exhibit an increased probability of detecting adaptive evolution, while for short sequences the power can be almost 0%. Very similar sequences carry little information, causing low power of the LRT. The power increases with sequence divergence until it reaches its maximal value, after which further increases of sequence divergence lead to reduced power. With multiple substitutions at the same site, the most recent changes might overwrite previous substitutions, causing loss of information. Thus, very divergent sequences do not contain much information.

The most efficient way of obtaining high power appears to be to use many sequences. Adding more sequences causes a spectacular rise in power, even when the sequence divergence is low. Increasing the strength of positive selection also leads to improved power. Increasing the proportion of positively selected sites should have a similar effect, although no simulations were performed to examine it.

Differences Between the Two LRTs

We obtained significant results much more often with the LRT that compares M0 (one-ratio) and M3 (discrete) than with the LRT that compares M7 (beta) and M8 (beta&ω). We note that M7 is a very flexible null model and accounts for both neutral and deleterious mutations with $0 < \omega < 1$. As a result, the M7-M8 com-

parison is a very stringent test of positive selection. The M0-M3 comparison, however, is more a test of variable selective pressure among sites (indicated by the ω ratio) than a test of positive selection. Since the selective pressure seems to be variable among sites in every functional protein, M0 is a very unrealistic model. For example, in all 10 data sets analyzed by Yang et al. (2000), M0 was easily rejected when compared with M3, although in four of them positive selection was not detected. Thus, if by chance parameter estimates under M3 indicate positive selection, we might falsely claim positive selection using the LRT comparing M0 and M3. We performed one such simulation experiment where the assumption of M0 was violated. We simulated 500 replicate data sets, each with $N = 500$ codons, using parameter settings of experiment A in table 2 except that we used the neutral model (M1) for the ω distribution. M1 (neutral) assumes two site classes with the ω ratios $\omega_0 = 0$ and $\omega_1 = 1$. We set the proportions for the two site classes at $p_0 = 0.5$ and $p_1 = 0.5$. The simulated data were then analyzed using M0 and M3. In 75% of replicates, at least one of the three ω parameters in M3 was estimated to be greater than 1, and the LRT was also significant, leading to false detection of positive selection. The LRT comparing M7 and M8 applied to the same data sets were found to be robust to violation of assumptions and falsely detected positive selection in only 5% of the replicates at $\alpha = 0.05$. Furthermore, if the data were analyzed using M1 (neutral) and M3 (discrete), the false-positive rate was 0.02 at $\alpha = 0.05$. Following Yang et al. (2000), we thus recommend that multiple models and tests be used in real data analysis and that caution be exercised when only the M0-M3 comparison suggests positive selection.

Acknowledgments

We thank Willie Swanson and two anonymous referees for constructive comments. This study was supported by a Biotechnology and Biological Sciences Research Council grant to Z.Y. M.A. was supported by a Medical Research Council studentship.

LITERATURE CITED

AGULNIK, A. I., A. ZHARKIKH, H. BOETTGER-TONG, T. BOURGERON, K. MCELREAVEY, and C. E. BISHOP. 1998. Evolution of the *DAZ* gene family suggests that Y-linked *DAZ* plays little, or a limited, role in spermatogenesis but underlines a recent African origin for human populations. Hum. Mol. Genet. **7**:1371–1377.

BIELAWSKI, J. P., and Z. YANG. 2001. Positive and negative selection in the *DAZ* gene family. Mol. Biol. Evol. **18**:523–529.

CRANDALL, K. A., C. R. KELSEY, H. IMAMICHI, H. C. LANE, and N. P. SALZMAN. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. **16**: 372–382.

DA SILVA, J., and A. L. HUGES. 1998. Conservation of cytotoxic T lymphocyte (CTL) epitopes as a host strategy to constrain parasite adaptation: evidence from the *nef* gene of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **15**:1259–1268.

ENDO, T., K. IKEO, and T. GOJOBORI. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

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

GOLDMAN, N., and S. WHELAN. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. **17**:975–978.

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

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

LIO, P., and N. GOLDMAN. 1998. Models of molecular evolution and phylogeny. Genome Res. **8**:1233–1244.

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

NEILSEN, 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.

OTA, R., P. WADDELL, M. HASEGAWA, H. SHIMODAIRA, and H. KISHINO. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol. Biol. Evol. **17**:798–803.

PLIKAT, U., K. NIESELT-STRUWE, and A. MEYERHANS. 1997. Genetic drift can determine short-term human immunodeficiency virus type 1 *nef* quasispecies evolution in vivo. J. Virol. **71**:4233–4240.

SELF, S., and K.-Y. LIANG. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. **82**:605–610.

SILVEY, S. 1970. Statistical inference. Penguin Books, Middlesex, England.

STUART, A., J. K. ORD, and S. ARNOLD. 1999. Kendall's advanced theory of statistics. Vol. 2A. Oxford University Press, New York.

WAYNE, M., and K. SIMONSEN. 1998. Statistical tests of neutrality in the age of weak selection. TREE **13**:236–240.

WHELAN, S., and N. GOLDMAN. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol. Biol. Evol. **16**:1292–1299.

YANG, Z. 1996. Maximum likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. **42**:587–596.

———. 2000. Phylogenetic analysis by maximum likelihood (PAML). Version 3.0. University College London, London, England.

YANG, Z., and J. BIELAWSKI. 2000. Statistical methods for detecting molecular adaptation. TREE **15**:496–503.

YANG, Z., and R. NEILSEN. 2000. Estimating synonymous and nonsynonymous rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32–43.

YANG, Z., R. NEILSEN, N. GOLDMAN, and A.-M. K. PEDERSEN. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

ZANOTTO, P. M., E. G. KALLAS, R. F. SOUZA, and E. C. HOLMES. 1999. Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics **153**:1077–1089.

ZHANG, J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol. Biol. Evol. **16**:868–875.

FUMIO TAJIMA, reviewing editor

Accepted April 30, 2001