J Mol Evol (2004) 60:426–433
DOI: 10.1007/s00239-004-0137-1

JOURNAL OF MOLECULAR EVOLUTION

Accuracy and Power of the Likelihood Ratio Test for Comparing Evolutionary Rates Among Genes

Jan Erik Aagaard, Patrick Phillips

Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, OR 97403, USA

Received: 29 April 2004 / Accepted: 11 October 2004 [*Reviewing Editor:* Dr. Rosmus Nielsen]

Abstract. Sequences for multiple protein-coding genes are now commonly available from several, often closely related species. These data sets offer intriguing opportunities to test hypotheses regarding whether different types of genes evolve under different selective pressures. Although maximum likelihood (ML) models of codon substitution that are suitable for such analyses have been developed, little is known about the statistical properties of these tests. We use a previously developed fixed-sites model and computer simulations to examine the accuracy and power of the likelihood ratio test (LRT) in comparing the nonsynonymous-to-synonymous substitution rate ratio (ω = dN/dS) between two genes. Our results show that the LRT applied to fixed-sites models may be inaccurate in some cases when setting significance thresholds using a χ^2 approximation. Instead, we use a parametric bootstrap to describe the distribution of the LRT statistic for fixed-sites models and examine the power of the test as a function of sampling variables and properties of the genes under study. We find that the power of the test is high ($>80\%$) even when sampling few taxa (e.g., six species) if sequences are sufficiently diverged and the test is largely unaffected by the tree topology used to simulate data. Our simulations show fixed-sites models are suitable for comparing substitution parameters among genes evolving under even strong evolutionary constraint ($\omega \approx 0.05$), although relative rate differences of 25% or less may be difficult to detect.

Key words: Likelihood ratio test — Nonsynonymous/synonymous rate ratio

Introduction

A major theme in the study of molecular evolution has been the comparison of evolutionary rates among genes. This includes comparisons among broad classes of loci, for example, nuclear and plastid genes, as well as comparisons among families of molecules such as the globin genes (Li 1997). More recently, the completion of several large-scale sequencing projects and the relative ease of cloning and sequencing from multiple populations or species has allowed for specific hypotheses regarding the selective pressures under which different genes evolve to be tested. These comparisons have implications for understanding basic biological processes such as identifying the control points of signal transduction pathways (e.g., Riley et al. 2003) and the degree of evolutionary constraint among components of biosynthetic pathways (e.g., Lu et al. 2003).

A number of methods of analyzing sequence data from coding regions have been developed (reviewed in Yang and Bielawski 2000; Yang 2002). Prominent among these have been maximum likelihood (ML) methods employing the codon substitution models developed by Goldman and Yang (1994; see also Muse and Gaut 1994). These codon models are intuitively appealing in their use of the nonsynonymous-to-synonymous substitution rate ratio (ω = dN/dS)

Correspondence to: Jan Erik Aagaard, Department of Genome Sciences, University of Washington, Box 357730, Seattle, WA 98195-7730, USA; *email:* jaagaard@gs.washington.edu

to define the type and strength of selection. ML methods are extremely flexible, allowing for multiple models of sequence evolution to be fit to the data while incorporating a variety of substitution parameters which may vary across a phylogeny (Yang 1998) or along a sequence (Nielsen and Yang 1998; Yang et al. 2000). Codon models that allow for heterogeneous selection pressures among sites have received particular interest because of their power to detect selection (Yang and Nielsen 2002) and their potential utility in predicting which sites are under selection (Anisimova et al. 2002). Yang and Swanson (2002) introduced a subset of heterogeneous-sites models called fixed-sites models wherein a sequence is partitioned *a priori* based on previous knowledge of functional domains. Significantly, it was noted that fixed-sites models were also readily applicable to the analysis of multiple genes from the same species so that selection pressures influencing the evolution of different genes can be compared (Yang and Swanson 2002).

To our knowledge, the fixed-sites models of Yang and Swanson (2002) have yet to be applied to the comparison of ω's among genes. Although power analyses using other ML models have been conducted (e.g., Anisimova et al. 2001), the accuracy and power of the likelihood ratio test (LRT) in comparing fixed-sites models have not been examined. In this study we use simulated data sets to describe the distribution of the LRT statistic and the accuracy of the LRT based on a χ^2 approximation for fixed-sites models. In addition, we conduct analyses of the power of the LRT using fixed-sites models to detect different ω values as a function of common sampling variables (sequence divergence, number of taxa sampled, sequence length, tree topology) and properties of the genes under study (ω values). Given that sequencing efforts are finite in most studies, efficient allocation of sequencing resources among taxa (that are more or less closely related) is an important consideration in experimental design. Our results are intended to aid in both the selection and analysis of existing data sets and the design of further sequencing experiments in order to compare selection pressures among genes using fixed-sites models.

Theory and Methods

Fixed-Sites Models

Yang and Swanson (2002) present a few simple modifications to the codon substitution model of Goldman and Yang (1994) in order to implement their fixed-sites models. Fixed-sites models allow an *a priori* partition of nucleotide sequence, which may correspond to discrete domains within a gene or multiple concatenated genes. The simplest fixed-sites models assume all site partitions (genes in our case) have identical substitution parameters including the same absolute rates (branch lengths), the same transition/transversion rate ratio (κ), the same nonsynonymous/synonymous rate ratio (ω), and the same parameters for codon frequencies (π_s). Successively more complex models allow individual parameters to vary among partitions. For example, in Yang and Swanson's (2002) Model C, only rate ratios (κ and ω) are assumed to be the same among site partitions; in Model E all parameters including κ and ω are assumed to be different among partitions (unfortunately, κ and ω cannot be decoupled in current implementations of PAML; Z. Yang, personal communication). Twice the difference in log-likelihood values for these two models constitutes a test of the hypothesis that κ and ω are equal among site partitions, where a χ^2 distribution with two degrees of freedom (the difference in the number of parameters between models with two genes or partitions; Yang and Swanson 2002) is typically used to set significance levels in hypothesis testing.

Distribution of the LRT Statistic

We generated replicate simulated data under a null hypothesis (H_0, genes evolving at equal rates) and examined the distribution of the LRT statistic relative to χ^2 in order to assess the frequency of type I errors (incorrectly rejecting H_0). We refer to these as accuracy experiments. From a table of codon usage for the *Drosophila melanogaster* ADH gene (accession no. M17827) and a simple star phylogeny with equal branch lengths, we generated between 250 and 2000 replicate data sets (depending on the variable being examined) for each of two genes using the evolver program in the PAML computer package (version 3.13; Yang 1997). Although simulation parameter values differed among accuracy experiments (see below), they were identical between the two genes in all cases. Four simulation parameters were independently varied: (1) sequence divergence, measured as branch lengths of the star phylogeny; (2) number of species (sequences for each of two genes) in the phylogeny; (3) sequence length (number of codons); and (4) the degree of selective constraint under which sequences evolve, as measured by the nonsynonymous-to-synonymous rate ratio (ω_1 and ω_2 for genes 1 and 2, respectively; ω held constant among codons and branches of the phytogeny for a gene).

(1) Five levels of sequence divergence were simulated (the expected number of nucleotide substitutions per codon along each branch of the star phylogeny = 0.05, 0.10, 0.20, 0.50, and 0.80), holding other simulation parameters constant (12 species; 300 codons; $\omega_1 = \omega_2 = 0.25$). (2) Three levels of taxon sampling were simulated (6,12, or 18 species; branch lengths = 0.05 to 0.80; 300 codons; $\omega_1 = \omega_2 = 0.25$). (3) Three levels of sequence length were simulated (150, 300, or 600 codons; branch lengths = 0.05 to 0.80; 12 species; $\omega_1 = \omega_2 = 0.25$). (4) Three levels of selective constraint were simulated ($\omega_1 = \omega_2 = 0.05$, $\omega_1 = \omega_2 = 0.25$, or $\omega_1 = \omega_2 = 0.50$; branch lengths = 0.05 to 0.80; 12 species; 300 codons). The transition/transversion rate ratio ($\kappa = 2$) was kept constant in all simulations. Sequences for the two genes were concatenated and analyzed with the codeml program in the PAML package (Yang 1997) using options Mgene = 2 and Mgene = 4 (Models C and E in Yang and Swanson [2002], respectively) and the same star phylogeny as used to simulate data. Equilibrium codon frequencies for both models were calculated from the average nucleotide frequencies at the three codon positions (CodonFreq = 2) in most cases, though a subset of simulated data (see below) was also analyzed by estimating codon frequencies as free parameters (CodonFreq = 3). The LRT statistic was calculated as twice the difference in log-likelihood values for the two models, and the cumulative distribution of test statistics for each set of simulated data was plotted against a χ^2 distribution with two degrees of freedom (the difference in the number of parameters between Model C and Model E).

In addition to the four simulation parameters described above, we also tested the effect of tree topology on the distribution of the LRT. Three simplified topologies were used to model the effect of unequal branch lengths and shared phylogenetic history: a simple star phylogeny with equal branch lengths as in (1) to (4) above, a star phylogeny with all branches of unequal length, and a maximally symmetrical branched phylogeny (all branch lengths equal). Total tree lengths (the expected number of substitutions per codon along the tree) of 0.40, 0.80, 0.16, 4.0, and 6.4 were used holding other parameters constant (8 species; 300 codons; $\omega_1 = \omega_2 = 0.25$). Concatenated sequences were again analyzed with the codeml program in the PAML package (Yang 1997) using options Mgene = 2 and Mgene = 4 as above (Models C and E in Yang and Swanson [2002], respectively). However, for the maximally symmetrical branched phylogeny, we employed the correct phylogeny (the same tree as used to simulate data) in the codeml analysis as well as two incorrect phylogenies by swapping two taxa either between one terminal node of the tree or over the basal node of the tree.

Power Analysis

In order to assess the power of the LRT, we generated simulated data under an alternative hypothesis (H_A, genes evolving at different rates) and examined the frequency of type II errors (incorrectly failing to reject H_0). We refer to these as power experiments. Power experiments were identical to the accuracy experiments described above for all simulation parameters except that ω for one of the two genes was increased by 50% relative to the other in all simulations (e.g., $\omega_1 = 0.25$ vs. $\omega_2 = 0.375$). Because this rate difference adds an additional simulation parameter, we also examined its affect by simulating three levels of rate difference [$\Delta\omega = (\omega_2 - \omega_1)/\omega_1$; 10, 25, and 50%] while holding other parameters constant (branch lengths = 0.05 to 0.80; 12 species; 300 codons; $\omega_1 = 0.25$). Simulated data in power experiments were also analyzed with the codeml program in the PAML package (Yang 1997) as described above. Power of the LRT under different simulation parameters was calculated using the empirical distribution of LRT statistics from simulated data sets generated under H_0 to specify a significance threshold of $\alpha = 5\%$ (a parametric bootstrap; Goldman 1993).

Results and Discussion

Accuracy Experiments

In our accuracy experiments we examine the fit between a commonly applied probability distribution (χ^2) and the empirical distribution of LRT statistics for fixed-sites models as a function of a variety of simulation parameters. The distribution of LRT statistics from accuracy experiments generally follows a χ^2 distribution with two degrees of freedom. However, as simulation parameter values increase (number of species, sequence divergence, number of codons, ω), the distributions typically shift farther to the right of χ^2. For example, as the degree of sequence divergence (the expected number of nucleotide substitutions per codon along each branch of a star phylogeny) increases from 0.05 to 0.80, the empirical distribution of LRT statistics appears to deviate more strongly from χ^2 (Fig. 1A). Rerunning analyses using different starting values for ω and κ results in identitical likelihood scores, suggesting that our results represent the true maximum likelihoods and are not due to computational problems. The tree topology used to simulate data also does not appear to contribute towards this trend, as the magnitude of the shift is similar for data based on star phylogenies with uneven branch lengths or maximally symmetrical phylogenies (data not shown). Increasing the number of taxa sampled from 6 to 18 results in a similar shift in the distribution of the test statistic relative to χ^2 (Fig. 1B). In both cases H_0 is rejected more frequently than the specified α when using the χ^2 approximation as determined by a two-tailed binomial test (Table 1) (Zhang 1999). Similar trends were observed for our accuracy experiments examining the effect of increasing sequence length and increasing ω (data not shown).

Based on statistical theory (Stuart et al. 1999), the distribution of LRT statistics from a comparison of nested ML models should be asymptotically χ^2 distributed. Previous studies of the χ^2 approximation for nucleotide (Whelan and Goldman 1999) and codon models (Anisimova et al. 2001) generally agree with this theory. Whelan and Goldman (1999) found the transition/transversion rate ratio (κ) behaved as expected for ML estimators regardless of the additional substitution parameters in the model, suggesting κ is unlikely to contribute to the significant departure from χ^2 we observed for codon models. Similarly, because ω is also estimated by ML and is not constrained at the boundary of the parameter space in either of the models we used, this parameter is an unlikely source of the bias (Anisimova et al. 2001).

A more likely cause of the significant increase in type 1 errors we found has to do with methods used to calculate equilibrium codon frequencies. In most cases, we used models that estimate equilibrium codon frequencies indirectly from ML estimates of nucleotide frequencies at the three codon positions (CodonFreq = 2 in PAML; Yang 1997). Because data were simulated from empirical codon frequencies for ADH, this could introduce a systematic bias that becomes more extreme as the sample size (sequence divergence, sequence length, etc.) increases. Accordingly, we reanalyzed a subset of the simulated data using models that estimate equilibrium codon frequencies directly from observed frequencies in simulated data sets (CodonFreq = 3). These results show a very similar pattern to that in Figs. 1A and B, where the bias appears to increase with increasing sequence length or species number, resulting in a significant increase in type 1 errors (Table 1). Whelan and Goldman (1999) also found that nucleotide substitution models that use non-ML estimates of nucleotide frequencies resulted in biased test statistics. This suggests that the pattern we observed could be a common feature when parameters such as codon and nucleotide frequencies are not calculated directly using ML estimators.

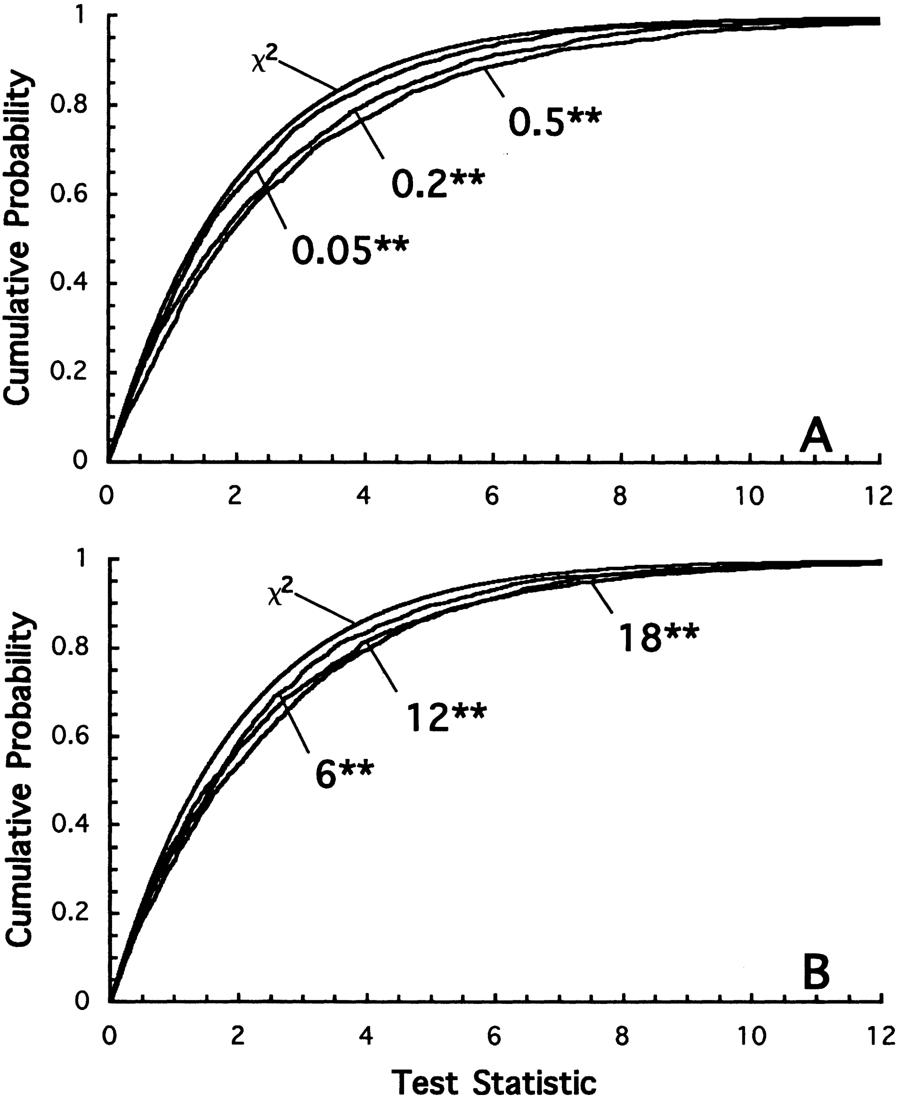

Fig. 1. Distribution of the cumulative probabilities of test statistics for χ^2 (2 df) and the LRT from accuracy experiments. **A** Two thousand data sets each of sequences for two genes using a star phylogeny with branch lengths of 0.05, 0.20, or 0.80 were generated under H_0, holding other simulation parameters constant (12 species; 300 codons; $\omega_1 = \omega_2 = 0.25$). **B** Two thousand data sets each of sequences for two genes from 6, 12, or 18 species were generated under H_0, holding other simulation parameters constant (branch lengths = 0.20; 300 codons; $\omega_1 = \omega_2 = 0.25$). The LRT statistics were calculated as twice the difference in log likelihoods between codeml Models C and Model E (Yang and Swanson 2002; CodonFreq = 2). Distributions reflect the probability of observing a test statistic as large or smaller by chance alone. Accuracy experiments in which the proportion of simulated data significantly exceed the specified $\alpha = 0.05$ (**0.005) were determined by a two-tailed binomial test.

To test the effects of approximations of codon frequencies on the distribution of LRT statistics in codeml models (CodonFreq = 2 or 3) relative to χ^2, we generated a replicate set of accuracy experiments as in Fig. 1. However, simulation parameters differed from those used previously by substituting equal codon frequencies in the PAML simulation program evolver (Yang 1997) for the codon frequencies from ADH used previously. Simulated data were again analyzed under models C and E of Yang and Swanson (2002), but both models assumed that all codon frequencies were equal (CodonFreq = 0) rather than calculating codon frequencies using approximations. Regardless of the simulation parameter values (branch length or species number), the type 1 error rate did not significantly exceed the specified α, in sharp contrast with our earlier simulations (Table 1). This is strong evidence that the approximation of codon frequencies alone is responsible for the bias in the LRT statistic.

Because using the χ^2 approximation when setting significance thresholds may lead to an increase in the type I error rate in some cases, we suggest the following caveat regarding LRTs comparing nested versions of Goldman and Yang's (1994) codon models as currently implemented in PAML (Yang 1997): when LRT statistics are marginally significant based on χ^2 approximation (e.g., 0.01 to 0.05 for $\alpha = 0.05$), a parametric bootstrap is necessary to firmly establish significance thresholds. In addition, because our simulations show the bias in the χ^2 approximation is consistent (underestimates the type 1 error rate), use of the parametric bootstrap is unlikely to increase the frequency of significant test results. When using nested ML models such as the fixed-sites models of Yang and Swanson (2002), applying the parametric distribution is straightforward. Substitution parameters are estimated from the original data under the null hypothesis (in our case, equal ω and κ among genes) and are then used to generate simulated data under the same null hypothesis (Goldman 1993). Simulated data are then analyzed under both null and alternative models (in

Table 1. Type 1 error rate of the LRT using the α^2 approximation

	$\alpha = 0.05$			$\alpha = 0.01$		
	CodonFreq = 0	CodonFreq = 2	CodonFreq = 3	CodonFreq = 0	CodonFreq = 2	CodonFreq = 3
Sequence divergence						
0.05	0.059^{ns}	0.067^{**}	0.072^{**}	0.015^{ns}	0.014^{ns}	0.016^{*}
0.20	0.049^{ns}	0.087^{**}	0.086^{**}	0.010^{ns}	0.021^{**}	0.023^{**}
0.80	0.046^{ns}	0.112^{**}	0.99^{**}	0.010^{ns}	0.034^{**}	0.32^{**}
Species number						
6	0.056^{ns}	0.068^{**}	0.082^{**}	0.010^{ns}	0.020^{**}	0.020^{**}
12	0.048^{ns}	0.087^{**}	0.074^{**}	0.014^{ns}	0.020^{**}	0.023^{**}
18	0.038^{ns}	0.089^{**}	0.099^{**}	0.008^{ns}	0.027^{**}	0.027^{**}

Note. Shown is the proportion of data sets simulated under H_0 for which H_0 is rejected using the χ^2 approximation (2 df). Likelihoods were estimated using codeml models which assume equal codon frequencies (CodonFreq = 0) for data simulated under models of equal codon frequencies or calculate codon frequencies using one of two different approximations (CodonFreq = 2 or CodonFreq = 3) for data simulated under empirical codon frequencies (see text). Models calculating codon frequencies using approximations significantly exceed the specified α in most cases as determined by a two-tailed binomial test. Significant level: *0.05; **0.005; ^{ns}not significant.

our case, different ω and κ among genes), and the distribution of LRT statistics from simulated data (twice the difference in log likelihoods between the models) used as the distribution of the test statistic under the null hypothesis. If the LRT statistic from the original data exceeds the largest 5% of test statistics from simulated data (α, an arbitrary threshold), the null hypothesis is rejected under the parametric bootstrap criterion. We apply this parametric bootstrap approach below to describe the power of the LRT statistic for fixed-sites models.

Power Experiments

In our power experiments, we explored the effect of sampling variables (sequence divergence, species number, sequence length, tree topology) and properties of the genes under study (degree of selective constraint, rate difference) on the power of the LRT applied to fixed-sites models. The range of simulation parameters was constrained based on previous power analyses of other codon substitution models (Anisimova et al. 2001) and expectations for the distribution of ω among broad surveys of genes (e.g., Barrier et al. 2003) in order to describe the test's performance for realistic sampling strategies. Accordingly, our results are not intended as a general description of the performance of the test and fixed-sites models but, rather, as a guide in selecting taxa and genes for which meaningful comparisons about substitution rates can be made.

Taxon Sampling and Sequence Length

We explored the effect of taxon sampling on power by simulating data for three levels of sampling (6, 12, and 18 species) across a range of sequence divergence. The degree of selective constraint and rate difference in simulations for the two genes were held constant ($\omega_1 = 0.25$, $\omega_2 = 0.375$; $\Delta\omega = 50\%$), as was the length of sequences (300 codons). Our simulations show that power increases as branch lengths of the star phylogeny increase for the range of simulation parameters we explored (Fig. 2A), although the rate of approach is markedly slower for the six-species simulations. Power when sampling 12 taxa with branch lengths of 0.2 is nearly the same as when sampling 6 taxa with branch lengths of 0.8 (83 and 87%, respectively). Accordingly, by focusing sampling at an appropriate level of divergence for the genes under study (see below), sequencing effort may be dramatically reduced (by 50% or more) without loss of power for hypothesis testing. However, as sequence divergence increases beyond the range in our simulations, power is expected to decrease due to multiple substitutions (Anisimova et al. 2001; see below).

One caveat of this asymptotic increase in power is the limitations of the test when sequence divergence is low. In our simulations using branch lengths of 0.05 and a star phylogeny, power was at or below 50% regardless of the number of taxa sampled. Transforming the scale for sequence divergence from branch length to the proportion of silent sites with substitutions (dS) suggests that species that have substitutions at roughly 4% or fewer silent sites provide little information for testing rate variation among genes using fixed-sites models (assuming $\omega = 0.25$ with 26% of positions silent for ADH; see Yang and Nielsen, [2000] for a transformation between branch length and dS).

Sequence length affects the power of the LRT in a similar fashion as the number of species sampled (Fig. 2B). In our simulations, we independently varied the length of sequences fourfold (150, 300, or 600 codons), holding other simulation parameters constant (12 species; $\omega_1 = 0.25$ and $\omega_2 = 0.375$) across a range of sequence divergence. Power again increases

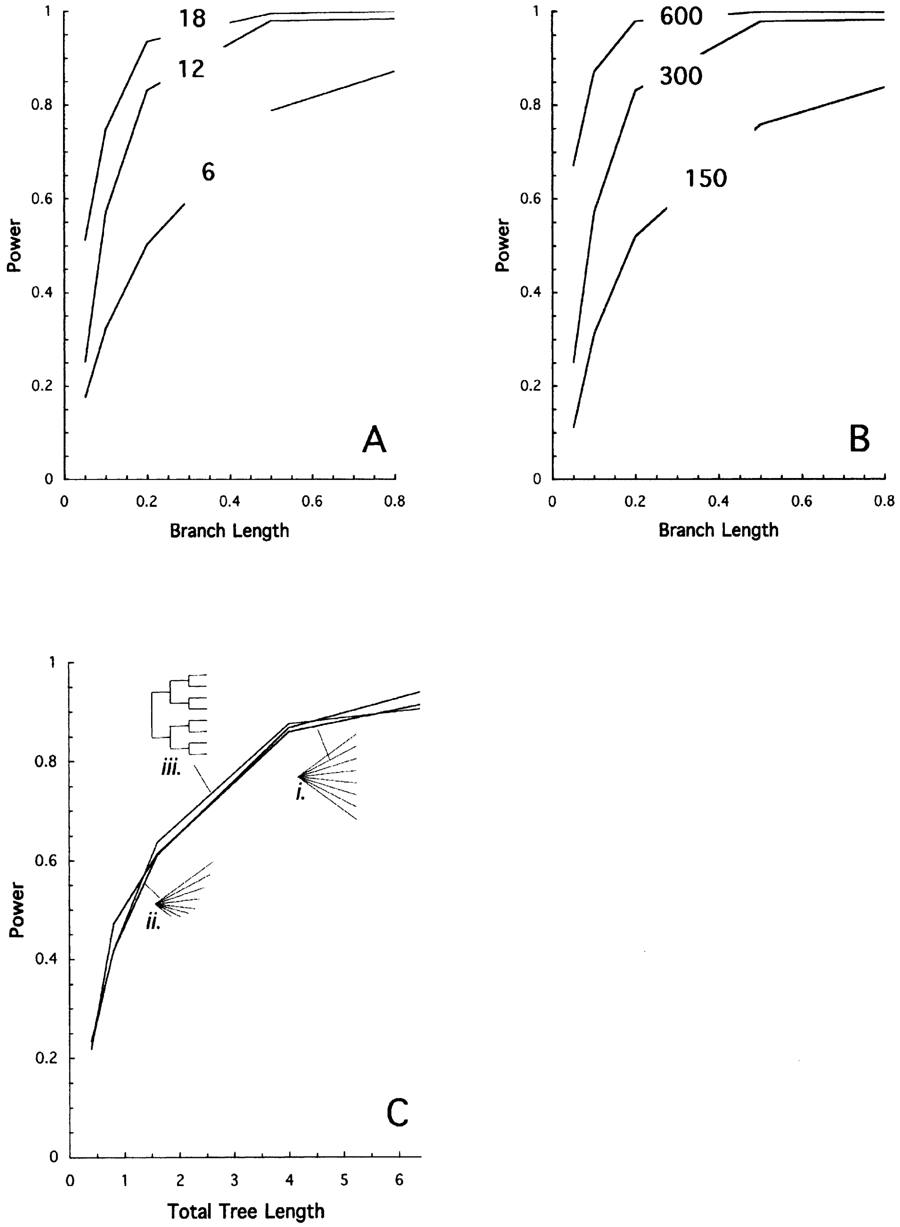

Fig. 2. Power of the LRT to detect differences in ω between two genes as a function of (**A**) the number of taxa sampled, (**B**) the length of the sequences being compared, and (**C**) tree topology. **A** Two hundred fifty data sets each of sequences for two genes evolving at a 50% rate difference from 6, 12, or 18 species were generated under H_A using a star phylogeny with equal branch lengths of 0.05 to 0.80, holding other simulation parameters constant (300 codons; $\omega_1 = 0.25$ and $\omega_2 = 0.375$). **B** Two hundred fifty data sets each of sequences for two genes 150, 300, or 600 codons in length were generated holding other variables constant as in A. **C** Two hundred fifty data sets each of sequences for two genes generated under H_A evolving along (*i*) a star phylogeny with equal branch lengths, (*ii*) a star phylogeny with unequal branch lengths, or (*iii*) a maximally symmetrical phylogeny with equal branch lengths (8 taxa; 300 codons; $\omega_1 = 0.25$ and $\omega_2 = 0.375$). In all cases power was calculated from the parametric bootstrap ($\alpha = 0.05$).

more slowly for shorter sequences as a function of sequence divergence such that 300 codon sequences have nearly identical power as 150 codon sequences when increasing branch lengths from 0.2 to 0.8 (83 and 84%, respectively). Doubling sequence length has a nearly identical effect on power as doubling the number of taxa (sequences) sampled. This is an expected result as we held ω constant among sites and branches for a gene when simulating data. However, for empirical data ω is likely to vary both among sites within a gene and among branches of the phylogeny. Yang and Nielsen (2002) suggest that for random-sites models, variation in ω within a gene has a stronger effect on power of the LRT than variation among lineages, presumably because the variance in ω is greater among sites. If true, sampling more sites (longer sequences) would also have a larger effect on power of the LRT for fixed-sites models than sampling more taxa, a phenomenon which is not reflected in our results. We suggest modeling variation in ω both among sites and among lineages will be important in future studies of fixed-sites models.

In most of our power experiments we used a star phylogeny with equal branch lengths. Because relationships among species are typically more complex, we also examined the effect of tree topology on power of the LRT for fixed-sites models. In order to test the effects of inequality of branch lengths and hierarchical relationships separately, we simulated data using the simplified star phylogeny with equal branch lengths as above, a star phylogeny with all branches of unequal length, and a maximally symmetrical (nested) phylogeny with equal branch lengths, holding other simulation parameters constant (8 species;

$\omega_1 = 0.25$ and $\omega_2 = 0.375$; 300 codons; total tree length = 0.4 to 6.4). Power appears to be nearly identical for all three topologies we used to generate simulated data (Fig. 2C) and exhibits the characteristic rise as a function of sequence divergence described previously. This suggests that the topological relationships reflected in species phylogenies should not be the central determining factor when selecting taxa for comparing evolutionary rates of genes. Rather, taxa should be selected primarily based on branch lengths of phylogenies.

Surprisingly, the LRT applied to fixed-sites models appears quite robust to violations in the phylogenetic relationships assumed among taxa. We also analyzed our simulated data generated under the maximally symmetrical tree using incorrect phylogenies which swapped two taxa across (*i*) terminal or (*ii*) basal nodes (data not shown). When simulated data were analyzed using fixed-sites models under (*i*), we found no apparent change in the power of the LRT to detect a significant difference between genes. When analyzed under (*ii*), power decreased by only about 7%. Though we did not examine how incorrect phylogenies influence the parameter estimates themselves, clearly the models and LRT are robust for purposes of comparing relative rate ratios among genes, at least for the range of simulation parameters we examined (8 species; $\omega_1 = 0.25$ and $\omega_2 = 0.375$; 300 codons; total tree length = 0.4 to 6.4).

Degree of Selective Constraint and Rate Difference

Most protein-coding genes evolve under strong purifying selection. This constraint is reflected in a typically low level of nonsynonymous substitution ($\omega << 1$) for the vast majority of such loci (e.g., Endo et al. 1996). In our simulations, we explored how the degree of selective constraint acting on loci affects the power of the LRT for fixed-sites models by varying ω's 10-fold ($\omega_1 = 0.05$ and $\omega_2 = 0.075$, $\omega_1 = 0.25$ and $\omega_2 = 0.375$, or $\omega_1 = 0.50$ and $\omega_2 = 0.75$; $\Delta\omega = 50\%$ in all cases) across a range of sequence divergence, while holding other simulation parameters constant (12 species; 300 codons). The range of ω's we used was intended to bracket plausible values estimated for functionally diverse categories of loci (Barrier et al. 2003). Power of the test applied to genes evolving under the strongest selective constraint ($\omega_1 = 0.05$ and $\omega_2 = 0.075$) is low at low sequence divergence but approaches that for genes evolving under weaker constraint ($\omega_1 = 0.25$ and $\omega_2 = 0.375$, or $\omega_1 = 0.50$ and $\omega_2 = 0.75$; Fig. 3A) as sequence divergence increases (84% for branch lengths of 0.8). This suggests fixed-sites models are applicable for testing whether even strongly constrained genes evolve at different rates, assuming that taxa are sufficiently diverged. Not surprisingly, the greatest limitation of fixed-sites models appears to be their limited power to detect small relative differences in ω among genes ($\Delta\omega$). In order to explore the effect of ω on power of the LRT for fixed-sites models, we varied the rate difference [$\Delta\omega = (\omega_2-\omega_1)/\omega_1$] for the two genes by 10, 25, and 50%, holding other simulation parameters constant (12 species; 300 codons; $\omega_1 = 0.25$) across a range of sequence divergence. Regardless of branch lengths, power of the LRT was low (never exceeding 60%; Fig. 3B) for $\Delta\omega = 10$ and 25%. This is in marked contrast to the case of $\Delta\omega = 50\%$ (held constant in Figs. 2A–C and Fig 3A), where power rises rapidly as a function of sequence divergence. In short, differences in ω of 25% or less are unlikely to be detectable using fixed-sites models unless very divergent sequences are compared (where multiple substitutions may cause problems; Anisimova et al. 2001) or large numbers of taxa are sampled.

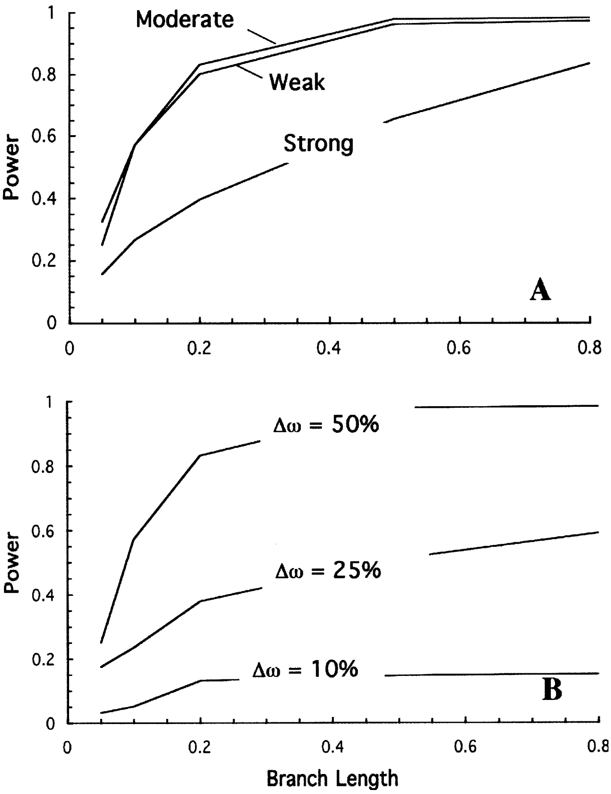

Fig. 3. Power of the LRT to detect differences in ω between two genes as a function of (**A**) the degree of evolutionary constraint acting on genes and (**B**) the rate difference between two genes. **A** Five hundred data sets each of sequences for two genes evolving at a 50% rate difference under strong constraint ($\omega_1 = 0.05$, $\omega_2 = 0.075$), moderate constraint ($\omega_1 = 0.25$, $\omega_2 = 0.375$), or weak constraint ($\omega_1 = 0.50$, $\omega_2 = 0.75$) were generated holding other variables constant (star phylogeny with equal branch lengths; 12 species; 300 codons). **B** Two hundred fifty data sets each of sequences for two genes evolving with rate differences of 10% ($\omega_1 = 0.25$, $\omega_2 = 0.275$), 25% ($\omega_1 = 0.25$, $\omega_2 = 0.313$), or 50% ($\omega_1 = 0.25$, $\omega_2 = 0.375$) were generated holding other variables constant as in A. In all cases power was calculated from the parametric bootstrap ($\alpha = 0.05$).

Conclusions

Codon substitution models such as the fixed-sites models of Yang and Swanson (2002) are likely to see broad application in future studies comparing evolutionary rates among genes because they provide estimates of substitution parameters that have clear evolutionary interpretation (e.g., ω's) and directly incorporate a statistical approach for testing the significance of parameters (LRTs). Our simulations suggest that a number of issues should be considered prior to applying these models. First, our accuracy experiments demonstrate that the distribution of LRT statistics for fixed-sites models deviate from the expected χ^2 distribution for a range of simulation parameters, resulting in a significant (though slight) increase in type 1 errors. This finding suggests that a parametric bootstrapping procedure (Monte Carlo simulations) may be necessary for setting significance thresholds in some cases. Based on our simulations, we suggest first using a χ^2 approximation to estimate the significance of the observed test statistics, followed by parametric bootstrapping for marginally significant results. Second, our power experiments show that fixed-sites models have limited power to detect differences in ω of 25% or less. This limitation does not mean that fixed-sites models cannot be used to compare ω's between slowly evolving genes, however, as power is high when sequences have sufficiently diverged. Third, taxa should be selected for comparisons based primarily on the extent of sequence divergence among species, as topological relationships reflected in species phylogenies do not appear to strongly affect the test. Finally, sampling effort (the number of taxa from which genes are sequenced) may be greatly reduced by identifying taxa at an optimal level of divergence for the genes under study. In general, our simulations show that comparing sequences from at least 12 taxa having substitutions at 20% or more of codons gives a high power ($>80\%$) using the LRT with fixed-sites models.

Acknowledgments. We thank Z. Yang, W. Swanson, and J. Felsenstein for helpful discussions. This work was supported by an NSF IGERT predoctoral fellowship and NSF dissertation improvement grant (DEB-0105176) to J.E.A. and by a grant from the NIH (GM54185) to P.C.P.

References

Anisimova M, Bielawski JP, Yang Z (2001) The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. Mol Biol Evol 18:1585–1592

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958

Barrier M, Bustamante CD, Yu J, Purugganan MD (2003) Selection on rapidly evolving proteins in the *Arabidopsis* genome. Genetics 163:723–733

Endo T, Ikeo K, Gojobori T (1996) Large-scale selection 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, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Huelsenbeck JP, Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol Syst 28:437–466

Huelsenbeck JP, Rannala B (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276:227–231

Li W (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Lu Y, Rausher MD (2003) Evolutionary rate variation in anthocyanin pathway genes. Mol Biol Evol 20:1844–1853

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

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

Riley R, Jin W, Gibson G (2003) Contrasting selection pressures on components of Ras-mediated signal transduction in *Drosophila.* Mol Ecol 12:1315–1323

Stuart A, Ord K, Arnold S (1999) Kendall's advanced theory of statistics, 6th ed, Vol 2a. Arnold, London

Whelan S, Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol Biol Evol 16:1292–1299

Yang Z (1997) PAML: a program for package for phylogenetic analysis by maximum likelihood. CABIOS 15:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z (2002) Inference of selection from multiple species alignments. Curr Opin Genet Dev 12:688–694

Yang Z, Bielaeski JP (2000) Statistical methods for detecting molecular adaptation. TREE 15:496–501

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917

Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19:49–57

Yang Z, Nielsen R, Goldman N, Petersen A (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–499

Zhang J (1999) Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol Biol Evol 16:868–875