Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative

Maria Anisimova, Olivier Gascuel

Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative

MARIA ANISIMOVA[1,2] AND OLIVIER GASCUEL[1]

[1]*Equipe Méthodes et Algorithmes pour la Bioinformatique LIRMM-CNRS, Université Montpellier II, Montpellier 34392, France;
E-mail: manisimova@hotmail.com (M.A.); gascuel@lirmm.fr (O.G.)*
[2]*Current Address: Biology Department, University College London, Darwin building, Gower Street, London WC1E 6BT, United Kingdom*

Abstract.—We revisit statistical tests for branches of evolutionary trees reconstructed upon molecular data. A new, fast, approximate likelihood-ratio test (aLRT) for branches is presented here as a competitive alternative to nonparametric bootstrap and Bayesian estimation of branch support. The aLRT is based on the idea of the conventional LRT, with the null hypothesis corresponding to the assumption that the inferred branch has length 0. We show that the LRT statistic is asymptotically distributed as a maximum of three random variables drawn from the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution. The new aLRT of interior branch uses this distribution for significance testing, but the test statistic is approximated in a slightly conservative but practical way as $2(\ell_1 - \ell_2)$, i.e., double the difference between the maximum log-likelihood values corresponding to the best tree and the second best topological arrangement around the branch of interest. Such a test is fast because the log-likelihood value ℓ_2 is computed by optimizing only over the branch of interest and the four adjacent branches, whereas other parameters are fixed at their optimal values corresponding to the best ML tree. The performance of the new test was studied on simulated 4-, 12-, and 100-taxon data sets with sequences of different lengths. The aLRT is shown to be accurate, powerful, and robust to certain violations of model assumptions. The aLRT is implemented within the algorithm used by the recent fast maximum likelihood tree estimation program PHYML (Guindon and Gascuel, 2003). [Accuracy; branch support; likelihood-ratio test; phylogeny reconstruction; power.]

The increased interest in tree reconstruction in recent years (e.g., "Tree of Life" project: http://tolweb.org/tree/phylogeny.html; phylogenetic database TreeBASE: http://www.treebase.org/treebase/index.html) prompts further methodological developments. One common task in phylogenetic inference is to test various phylogenetic relationships statistically and to measure the support in favor of one or another hypothesis for the given data and the chosen model. Here, we focus on this task of statistically evaluating branch support in phylogenies. Several tests and support measures of phylogenetic relationships were proposed and explored more than two decades ago, including the parametric and nonparametric bootstrap and jackknife, the Bremer support, the likelihood-ratio test, the interior-branch test, the conditional probability of reconstruction, the relative support, and spectral plots (for review see Swofford et al., 1996; Siddall, 2002; Felsenstein, 2004). Some branch support measures have since been shown to be inaccurate, or difficult to interpret, whereas others, such as nonparametric bootstrap supports (Felsenstein, 1985) and Bayesian posterior probabilities (Li, 1996; Rannala and Yang, 1996; Mau et al., 1997; Larget and Simon, 1999; Huelsenbeck et al., 2001), have become standard practice. However, even for these widely used methods, appropriate interpretation is controversial, and reliability has been examined in a lengthy and intensive debate (Hillis and Bull, 1993; Felsenstein and Kishino, 1993; Sanderson, 1995; Berry and Gascuel, 1996; Suzuki et al., 2002; Wilcox et al., 2002; Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003a; Erixon et al., 2003; Simmons et al., 2004; Taylor and Piel, 2004).

Although the theory of nonparametric bootstrap is well established (Efron and Tibshirani, 1993), the full implementation is too complicated to be applied in phylogenetics due to the discontinuous nature of the tree variable. Felsenstein's bootstrap (Felsenstein, 1985), by far the most commonly used implementation, is only a first-order approximation, which may be poor in some cases due to the curvature of the tree space (Efron et al., 1996). Depending on local configuration of the topological space around the inferred tree, it may be conservative or liberal, and correcting for this effect is hard to achieve (Efron et al., 1996). However, the nonparametric bootstrap (including Felsenstein's approximation) does not depend on a priori specified hypotheses about the underlying evolutionary processes and therefore generally does not suffer from false assumptions (but see Galtier, 2004). Felsenstein's bootstrap is often thought to be consistently conservative, as bootstrap proportions tend to underestimate the probability for the clades to be true (Zharkikh and Li, 1992; Hillis and Bull, 1993), but this finding and its interpretation was contested by Efron et al. (1996) and Durbin et al. (1998) (see also our results below). At least three different interpretations of bootstrap were proposed (summarized in Yang and Rannala, 2005), which illustrates that the debate is still open.

Statistical theory behind the Bayesian phylogenetic inference is well defined. The Bayesian branch support represents the probability that the clade in question is true conditional on the data, the model, and the parameter priors (e.g., Huelsenbeck et al., 2002; Huelsenbeck and Rannala, 2004). In practice, MCMC chains are used to approximate the Bayesian tree inference procedure, but it is not always easy to decide how long (and how many) MCMC chains should be run (e.g., Geyer, 1992; Cowles and Carlin, 1996). In phylogenetics, large-taxon data sets have much larger tree spaces, making it more difficult to achieve convergence. Moreover, as any parametric method, Bayesian inference is sensitive to model assumptions (Huelsenbeck and Rannala, 2004; Yang and Rannala, 2005). Finally, some studies showed that the

Bayesian analysis may produce high supports for nonexisting clades (Cummings et al., 2003; Erixon et al., 2003; Suzuki et al., 2002) or simultaneously support contradicting relationships (Buckley et al., 2002; Douady et al., 2003b). Lewis et al. (2005) suggest that such phenomena could often be due to the failure of current Bayesian methods to account for polytomies and offer a simple solution by introducing unresolved trees into the tree space.

Numerous studies explored the association between the two measures, and the possibility of using bootstrap supports and Bayesian probabilities as lower and upper bounds of node reliability, respectively (Suzuki et al., 2002; Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003a; Erixon et al., 2003; Taylor and Piel, 2004). All authors agreed that Bayesian probabilities were on average higher than nonparametric bootstrap support values, but the relationship between Bayesian and bootstrap supports was variable. Thus, direct comparison of the two measures is difficult, especially considering the sensitivity of Bayesian posterior probabilities to prior assumptions (Yang and Rannala, 2005).

In this study we do not pursue the debate on the advantages, limitations, and relationships of bootstrap re-sampling and Bayesian methods, but we explore an alternative. The new test for branches we propose is a modification of the standard likelihood-ratio test (LRT; e.g., Stuart et al., 1999), which compares an alternative hypothesis of a positive branch length ($t \geq 0$) to the nested null hypothesis with the branch of interest being constrained to a zero-length ($t = 0$). The standard LRT statistic is calculated as double the difference of the maximum log-likelihood values under the alternative and the null hypotheses, $2(\ell_1 - \ell_0)$ or $2\Delta\ell$. Under regularity conditions, the LRT statistic is asymptotically distributed as χ_1^2 with degrees-of-freedom (d.f.) equal to the difference in the number of free parameters allowed by the alternative and the null hypotheses (Chernoff, 1954; Stuart et al., 1999). The performance of the LRT for non-zero interior branch length was previously assessed by Gaut and Lewis (1995). The authors used χ_1^2 for significance testing and found the test to have high type I error rate. This fault was attributed to data limitations. Moreover, the authors suggested that the distribution of the LRT statistics under the null may vary from χ_1^2. Indeed, the null hypothesis has only one fewer parameter ($t = 0$), but it is fixed at the boundary, since branch lengths are non-negative, causing the asymptotic distribution to take shape of a 50:50 mixture of χ_0^2 and χ_1^2 (Chernoff, 1954; Self and Liang, 1987). This has been confirmed by simulations when a tree topology in the alternative hypothesis is a priori fixed, i.e., not inferred from the data at hand (Goldman and Whelan, 2000; Ota et al., 2000). However, this is not generally the case in phylogeny reconstruction as a single data set is usually used both to infer the tree and to test the branches of the inferred tree.

Here we describe the theoretical shape of the distribution of the standard LRT statistics when the alternative hypothesis allows different branching arrangements around the branch of interest. This distribution is used for significance testing in our new approximate LRT (aLRT), where the standard null hypothesis "the branch has 0-length" is *approximated* by the more general hypothesis "the branch is incorrect." More specifically, the aLRT compares the likelihoods of the best and the second best alternative arrangements around the branch of interest. We explain the rationale for such an approximation and show that our aLRT statistic has a null distribution close to the theoretical null distribution of the standard LRT statistic when multiple alternatives are accounted for.

Ideally a branch test should be fast for large trees, accurate, powerful, and robust to misspecifications of key assumptions. We therefore test partial optimization of likelihood scores, which greatly reduces the computational time while retaining good accuracy and power achieved with full optimization. Because aLRT is parametric, we evaluate the robustness of aLRT to model violations as these can have a negative effect on both the ML phylogenetic inference and parameter estimation (e.g., Sullivan and Swofford, 2001). Performance of the aLRT is studied on 4-, 12-, and 100-taxon simulated data sets. We compare the accuracy and power of the aLRT and branch tests based on ML bootstrap supports and Bayesian posterior probabilities. The benefits and drawbacks of these methods are discussed.

METHODS

Multiple Testing Correction

There exist only three alternative arrangements around a branch of interest (Fig. 1A–C) and, therefore, only three alternative topologies are allowed under the alternative hypothesis ($t > 0$). Let X_1, X_2, and X_3 be random variables representing the LRT statistics calculated for the three possible topological arrangements. In each such LRT, the topology is fixed and so X_1, X_2, and X_3 are asymptotically distributed as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = f(x)$, under the null hypothesis. Let $F(x)$ be the cumulative probability function corresponding to density $f(x)$. If, under the null hypothesis, the topology around the branch of interest is not fixed but inferred by ML using the (unique) data set at hand, then the "LRT" statistic of such test is distributed as the maximum of X_1, X_2, and X_3. Denote $f^*(x)$ and $F^*(x)$ as the corresponding density and cumulative functions, respectively. By definition,

$$F^*(x) = \Pr\{\text{AND}[\max(X_i)] \leq x\} = 1 - \Pr[\text{OR}(X_i > x)],$$

and so

$$F^*(x) \geq 1 - \sum_i \Pr(X_i > x) = 1 - 3[1 - F(x)]. \quad (1)$$

Assume now that X_i variables are independent, we have:
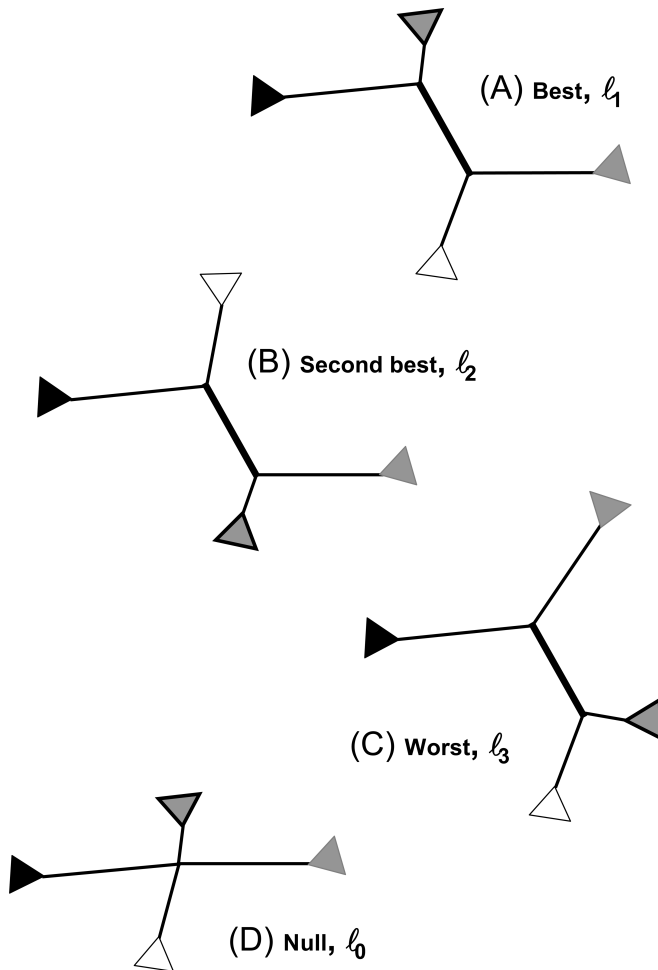
$$F^*(x) = F(x)^3. \quad (2)$$

FIGURE 1. Representation of alternative topological arrangements around the branch of interest for any tree with ≥ 4 taxa: (A) The best ML tree with the ML score ℓ_1; (B) the best ML tree rearranged around the branch of interest with the second best ML score ℓ_2; (C) the worst rearrangement of the best ML around the branch of interest with the ML score ℓ_3; (D) the tree representing the null hypothesis (branch of interest collapsed to zero length) with the ML score ℓ_0. The triangles of different shades represent different subtrees that remain unchanged in each rearrangement. The branch of interest is in bold, the four adjacent branches are in sharp.

Using Equation (1), to get a confidence level $1 - \alpha^*$ it is sufficient to chose a threshold x^* such that

$$F^*(x^*) = 1 - \alpha^* = 1 - 3[1 - F(x^*)],$$

which is equivalent to

$$F(x^*) = 1 - \alpha^*/3.$$

In other words, to test for the significance at level α^* of the maximum of three identically distributed variables with distribution f, it is sufficient to apply the standard test with $\alpha = \alpha^*/3$. This is known as the Bonferroni correction (Miller, 1981: pp. 67–70). It should be noted that this correction applies whether or not the variables are independent. However, consider Equation (2), which assumes

independence and set

$$F(x^*) = 1 - \alpha,$$

we get

$$F^*(x^*) = (1 - \alpha)^3 = 1 - 3\alpha + O(\alpha^2),$$

which means that Bonferroni correction closely fits the independence case. When the variables are positively correlated, Bonferroni correction tends to be conservative. But here correlations among X_i should be negative, as when one topology has high likelihood, the other two are usually poor and close to the null hypothesis. This means that our case is "between" independence and the extreme case represented by Equation (1), and that the Bonferroni correction should not be conservative in practice. To test for the significance of the inferred branch, we thus use the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution but apply Bonferroni correction. For example, to achieve 0.05 and 0.01 significance levels, we use $\alpha = 0.01666$ and 0.00333, which corresponds to statistic values 4.529 and 7.361, respectively.

However, Equation (1) and Bonferroni correction are only applicable for small α values (as usual in statistical testing) and should not be used to estimate branch support when the statistics value becomes low, as it may produce negative values. To estimate branch support we then use cubic approximation (2). For example, assuming that the statistic has value 1.0, $F(1.0) = 0.84$ and $F^*(1.0) = 0.84^3 = 0.59$, whereas using Equation (1) we get $1 - 3(1 - 0.84) = 0.52$. However, as explained above, both cubic and Bonferroni solutions become identical with higher value of the statistic.

Approximate LRT for Branches

The three possible topological arrangements (Fig. 1A–C) can be ordered according to their maximum log-likelihood values from the best tree to the worst: $\ell_1 \geq \ell_2 \geq \ell_3$. The standard LRT compares the ML value ℓ_1 of the best tree with the ML value ℓ_0 of a tree representing the null hypothesis (branch of interest is collapsed; Fig. 1D), relying on the calculation of the LRT statistic $2(\ell_1 - \ell_0)$. Consider instead using the statistic $2(\ell_1 - \ell_2)$, which compares the maximum likelihood value of the best tree with a maximum likelihood value of a less likely tree. In comparison with the null (tree D; Fig. 1), the hypothesis corresponding to tree B has an extra free parameter: the length of the branch under consideration. Thus, the ML value ℓ_2 can never be lower than ℓ_0, and the inequality $2(\ell_1 - \ell_0) \geq 2(\ell_1 - \ell_2)$ follows. This means that using the statistic $2(\ell_1 - \ell_2)$ instead of $2(\ell_1 - \ell_0)$ should result in a more conservative test, i.e., with lower type I error rate but possibly with lower power. Intuitively, the statistics $2(\ell_1 - \ell_0)$ and $2(\ell_1 - \ell_2)$ should be close, as the second best topology B usually does not provide a much better fit than does the star tree. Additionally, the statistic $2(\ell_1 - \ell_2)$ avoids conflicts reported for the standard LRT, which, in some rare cases, can be significant for all three or two possible topological arrangements around the branch of interest (Tateno et al., 1994).

Although the alternative hypothesis (and hence the ML value ℓ_1) remains the same for every internal branch, the null hypothesis changes with a change of the internal branch, and so the ML values ℓ_0 and ℓ_2 have to be calculated for each internal branch. Full optimization of those statistics is slow for large trees, so partial optimization should be advantageous. However, if approximation is not good enough, the LRT statistic is overestimated and the test becomes too liberal. We performed simulation experiments to determine the number of branches to be reestimated such that the approximation of ℓ_0 and ℓ_2 is reliable. We first explored the possibility of estimating ℓ_0 and ℓ_2 by recalculating the likelihood value with the branch of interest collapsed to zero length (in case of ℓ_0), or optimizing only the internal branch of interest (in case of ℓ_2). Next, we considered more accurate estimates of statistics by optimizing the internal branch of interest (but not in case of ℓ_0) and the four branches adjacent to it (branches shown in Fig. 1). In both experiments all model parameters were kept as estimated for the best tree.

Accuracy and Power of Tree Inference Based on a Branch Test

Under the null hypothesis ($t = 0$), no substitutions have occurred along the branch of interest. This does not fully correspond to what users envision when they perform branch tests. The basic question is whether the studied branch is correct or not. We therefore define measures of accuracy and power of *tree inference* based on a branch test, which are distinct from the accuracy and power of the LRT of non-zero branch length.

As usual, Accuracy = 1 – type I error rate, and power = 1 – type II error rate, but

Type I error rate = Pr(test is significant|the branch is

not correct), and

Type II error rate = Pr(test is *not* significant|

the branch is correct).

These definitions are similar to the standard measures of accuracy and power of statistical tests. Recall that for the standard LRT, the type I error rate = Pr(LRT is significant $|t = 0$), and the type II error rate = Pr(LRT is not significant $|t \geq 0$). Thus, if the null hypothesis "the branch length $t = 0$" can be approximated by the hypothesis "the branch is incorrect," then the accuracy and power of the LRT of non-zero branch length should be similar to the accuracy and power of tree inference. This means that for the hard (i.e., short) branches, which are the branches of interest when testing, we should have a strong correlation between tree inference and the standard test measures. Moreover, tree inference accuracy and power correspond to the practical measures that are biologically relevant, and they can be applied to estimate the accuracy and power of branch tests other than LRT. Type I and II error rates of tree inference were evaluated by simulation.

Computer Simulations

To explore properties of the *null distribution of the standard LRT* for internal non-zero branch length, we simulated 10,000 star trees with 4 taxa and branches drawn from the exponential distribution with expectation 0.25 substitutions per site. For each tree we simulated a data set under HKY+Γ with 4 rate categories and shape parameter 1.0, $\kappa = 4.5$, and unequal base frequencies: $f_A = 0.18$, $f_C = 0.24$, $f_G = 0.32$, $f_T = 0.26$. The sequence length was 1000 nucleotides (nt) in all simulations, unless otherwise stated. The simulated data were analyzed assuming first the star tree (the null) and then a fixed non-star tree (the alternative). We also analyzed data when the alternative non-star topology was unknown and therefore inferred from data, which more closely corresponds to actual phylogenetic studies. As we shall see, the Bonferroni correction is quite satisfactory in such (realistic) case. In all other simulations the alternative hypothesis assumed the topology to be unknown and the Bonferroni correction was used to account for multiple testing.

We explored the *effect of model misspecification* on the null distribution by analyzing these 4-taxon data sets with the correct model HKY+Γ as well as with simpler models JC, JC+Γ, and HKY. For each of 10,000 simulated star trees, we also generated data under more complex models and analyzed them using a simpler model HKY+Γ with all parameters being estimated from data. These more complex simulation models were (1) GTR+Γ with arbitrary parameters: nucleotide frequencies $f_A = 0.18$, $f_C = 0.24$, $f_G = 0.32$, $f_T = 0.26$, 4 rate categories of Γ shape parameter 1.0, and rates of nucleotide changes $r_{A\leftrightarrow C} = 3.0$, $r_{A\leftrightarrow G} = 10.5$, $r_{A\leftrightarrow T} = 1.3$, $r_{C\leftrightarrow G} = 1.4$, $r_{C\leftrightarrow T} = 15.0$, $r_{G\leftrightarrow T} = 1$; (2) GTR+$\Gamma$ with estimates from an HIV data set (Posada and Crandall, 2001): nucleotide frequencies $f_A = 0.40$, $f_C = 0.20$, $f_G = 0.22$, $f_T = 0.18$, four rate categories of Γ shape parameter 0.969, and rates of nucleotide changes $r_{A\leftrightarrow C} = 1.72$, $r_{A\leftrightarrow G} = 5.03$, $r_{A\leftrightarrow T} = 0.84$, $r_{C\leftrightarrow G} = 0.91$, $r_{C\leftrightarrow T} = 7.70$, $r_{G\leftrightarrow T} = 1$; (3) discrete codon model M3 with positive selection (Yang et al., 2000), with codon frequencies estimated from sperm lysine of 25 abalone species (as supplied in an example file in PAML package of Yang, 1997), transition/transversion ratio $\kappa = 4$, and three ω-classes with $\omega_0 = 0.1$, $\omega_1 = 0.8$, and $\omega_2 = 4.0$ in proportions $p_0 = 0.6$, $p_1 = 0.3$, and $p_2 = 0.1$, respectively.

To evaluate and compare the *type I error rate and the power of tree inference* based on the standard and approximate LRTs we simulated data under the alternative hypothesis (i.e., the true tree was non-star). With 4 taxa, we simulated 10,000 topologies with branches drawn from the exponential distribution with expectation of 0.15 changes per nucleotide. For each tree, data were simulated under HKY+Γ with parameters used to generate the null distribution of the standard LRT statistic for 4 taxa (see above). The results of 10,000 LRTs were reordered according to the tree length (S), the long-branch attraction (LBA), and the interior branch length (t) of the true tree, and the type I error rate and power were plotted

against these characteristics of the true tree. Both S and t were measured in expected nt changes per site along the tree or branch, respectively. Note that when t is close to 0, we expect results close to those obtained with star trees and standard statistical error measures. The LBA measure was calculated as the difference of the sum of the two longest branches (one on each side of the internal branch) and the sum of the two shortest branches (again, one on each side of the internal branch). For 4-taxon data, all branch lengths were optimized when computing the ML values ℓ_0 and ℓ_2.

Next, we evaluated *properties of the new aLRT* using 12- and 100-taxon data sets with 10,000 and 500 replicates, respectively. Data generation was analogous to Guindon and Gascuel (2003), to be referred to for more explanations and details. Non-star trees were simulated using the standard speciation process. Deviations from molecular clock were created by multiplying every branch length by $(1 + X)$, where X was exponentially distributed with expectation μ, which varied between the trees and was equal to $0.2/(0.001 + U)$ with uniform U drawn from [0, 1]. The greater the μ, the greater is the deviation from molecular clock. Finally, the tree length was rescaled to be uniformly distributed in the range $[S_{min}, S_{max}]$ by multiplying every branch length by $[S_{min} + V \times (S_{max} - S_{min})]/T$, where T was the tree length and V varied between the trees and was uniformly drawn from [0, 1]. The tree length range was [0.1, 2.5] for 12-taxon data and [0.5, 10] for 100-taxon data, so that both 12- and 100-taxon data had comparable maximum pairwise divergence. Phylogenies simulated this way reflect variability of evolutionary rates and differences in deviations from molecular clock, which is observed in real data sets. For each 12-taxon tree, sequences containing 1000, 500, 250, and 100 nt were simulated under the model HKY+Γ and the codon model M3 with positive selection as described above for 4-taxon simulations. For 100-taxon data, sequences with 1000 nt were simulated under HKY+Γ, as described earlier. To further check *robustness of our aLRT to over- and underparametrization*, we also evaluated 100-taxon data sets used in Desper and Gascuel (2004) and publicly available from http://www.lirmm.fr/mab/sommaire_english.php3 (together with details of simulation). These test data were simulated under (1) K2P+Γ model and (2) the covarion model and contained 500 replicates simulated under each model with sequences of 600 nt. All 12- and 100-taxon data sets were analyzed assuming HKY+Γ.

Nucleotide data described in this section were simulated using Seq-Gen (Rambaut and Grassly, 1997), but evolver from PAML package was used to generate data under codon model. Although all our simulations involve nucleotide sequences, the same approach can be applied to amino acid sequences.

Comparison with Branch Tests Based on ML Bootstrap Supports and Bayesian Posterior Probabilities

The first 1500 replicates of 12-taxon data simulated previously under HKY+Γ were used to compare the accuracy and power of tree inference based on the aLRT, ML bootstrap supports, and Bayesian posterior probabilities. All three methods used the correct analysis model. The aLRT branch supports were calculated using cubic approximation (2) and the mixture distribution $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, as explained earlier. PHYML (Guindon and Gascuel, 2003) was used to estimate ML bootstrap branch support; only 100 replicates were performed to minimize the computational costs during our simulation. Although a higher replicate number is desirable, only 100 replicates are often used in research papers for computing reasons (e.g., Wilcox et al., 2002; Rokas et al., 2003). MrBayes v3.1 (Huelsenbeck and Ronquist, 2001) was employed to estimate the Bayesian posterior probabilities of inferred internal branches. We used two independent runs of 3×10^4 generations (after a 5000-generation burnin), with 4 differently heated MCMC chains (as specified by default) and a sampling frequency of 10. Despite the short chains, we achieved good convergence as assessed by the average standard deviation of split frequencies between the two runs, which averaged 0.004. In addition, for the first 10 replicates we verified the convergence and the estimates of posterior probabilities using longer runs (10^6 generations with 4 heated chains and sampling every 100). In all 10 replicates the same trees were inferred, and the correlation between the posterior probabilities for inferred branches during short and long runs was as high as 0.995. Consequently, we assumed a similar general behavior in the full sample. The inferred internal branches were compared to the true tree and the corresponding Bayesian posterior probabilities were used to calculate the type I error and the power of tree inference.

RESULTS AND DISCUSSION

Null Distribution with Fixed and Inferred Topology

Using simulated 4-taxon star trees, we compared the null distributions of the standard LRT statistics when (1) the tree topology was fixed a priori and when (2) the tree topology was inferred from data. The distribution observed in the first case closely matched the expected $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution (result not shown). This confirms and generalizes the result of Ota et al. (2000), who simulated 4-taxon trees with fixed branch lengths (recall that in our simulation branch lengths were drawn from the exponential distribution with expectation 0.25). When the topological arrangement around the branch of interest was not fixed a priori but inferred from data, the null distribution clearly varied from $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, as expected (Fig. 2A). For confidence levels above 90%, the observed, the Bonferroni, and the cubic corrected mixture distributions were very similar, while the cubic correction gave better fit in the 80% to 90% range. For example, in our simulations the uncorrected type I error rate was 0.031 at the significance level $\alpha = 0.01$ and 0.143 at $\alpha = 0.05$, which confirms that correction for multiple testing is necessary. The Bonferroni and cubic corrections reduced the type I error rates to almost perfect 0.012 at $\alpha = 0.01$ and 0.052 at $\alpha = 0.05$. Thus, the Bonferroni correction seems to be well suited in the standard range $\alpha < 0.1$.
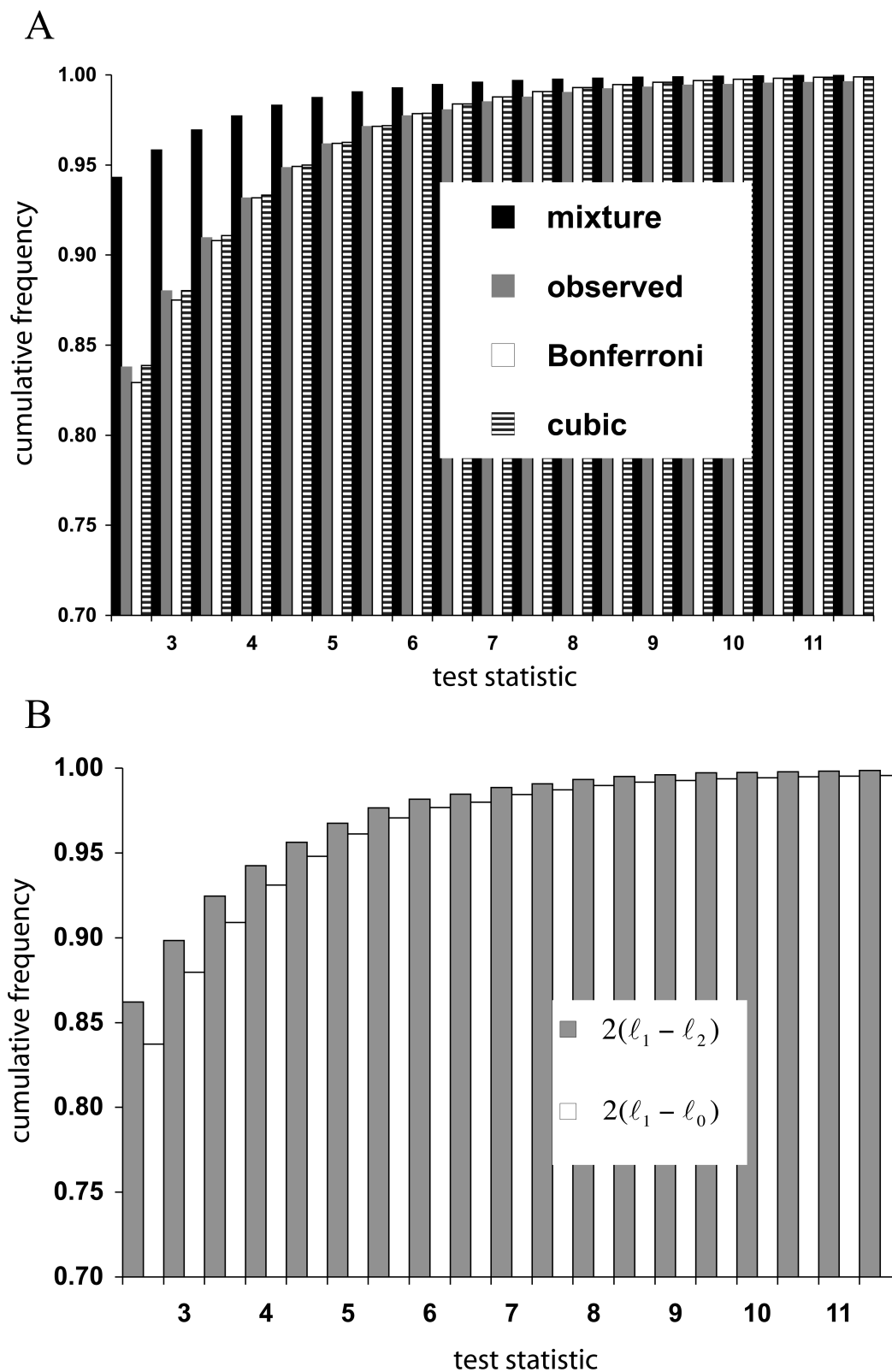
FIGURE 2.   Cumulative frequency graphs comparing (A) the uncorrected mixture distribution $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, the observed null distribution when tree is not fixed a priori, and corrected mixture distributions using Bonferroni and cubic corrections; (B) distributions of statistics $2(\ell_1 - \ell_2)$ and $2(\ell_1 - \ell_0)$, observed under the null hypothesis. Data were simulated and analyzed using HKY+Γ. The ML values ℓ_2 and ℓ_0 were fully optimized.

To check the robustness to model misspecifications, we repeated the analysis of the same 4-taxon data, simulated under HKY+Γ, assuming the oversimplified nucleotide models: JC, JC+Γ, and HKY. As the tree was assumed unknown under the alternative hypothesis, the Bonferroni correction was applied to estimate type I error rates. Even then, these were unacceptably high: 0.93 for model JC, 0.67 for HKY, and 0.33 for JC+Γ at $\alpha = 0.05$. The shape of the null distribution varied considerably (not shown). Such behavior was not unexpected since very important factors, transition/transversion bias, unequal nucleotide frequency bias, and unequal rates of evolution amongst sites were not accounted for. This shows the (expected) sensitivity of the LRT to model assumptions.

We further checked the extent to which the test was sensitive to model violations by simulating under a more complex model (GTR+Γ or codon model with positive selection) but analyzing with a simpler heterogeneous rates model (HKY+Γ). The resultant null distributions were relatively close in shape to the distribution obtained when the same (HKY+Γ) model was used for both data generation and analysis (curves not shown). Type I error rates were within acceptable limits: 0.012 at $\alpha = 0.01$ and 0.054 at $\alpha = 0.05$ for model GTR+Γ with arbitrarily chosen parameters; 0.01 at $\alpha = 0.01$ and 0.045 at $\alpha = 0.05$ for GTR+Γ with parameter estimates from HIV but slightly elevated 0.016 at $\alpha = 0.01$ and 0.082 at $\alpha = 0.05$ for codon data with positive selection. In these three cases the LRT performs well, despite significant model violation ($P < 0.001$ using the Goldman-Cox test; Goldman, 1993).

We can then draw conclusions about accuracy and robustness of the standard LRT with 4 taxa. When the analysis model describes data well, the type I error rate obtained using Bonferroni corrected mixture distribution is close to the significance level α, so that the standard LRT remains accurate. Moreover, our results suggest that minor (but detectable) deviations from model assumptions do not significantly affect its accuracy. However, when important factors (e.g., transition/transversion ratio, rate variation among sites) are not accounted for, the test can become very inaccurate.

Using $2(\ell_1 - \ell_2)$ Statistic

Next, we considered whether a more convenient statistic $2(\ell_1 - \ell_2)$ provides a suitable approximation of the standard LRT statistics. Figure 2B shows the observed cumulative null distributions (i.e., data simulated using star trees) of fully optimized statistics $2(\ell_1 - \ell_0)$ and $2(\ell_1 - \ell_2)$ for 4 taxa; both distributions are very close when the cumulative frequency is larger than 0.9 (i.e., in the region of practical interest). These results confirm that using $2(\ell_1 - \ell_2)$ makes the test slightly more conservative than using $2(\ell_1 - \ell_0)$: at the 0.05 significance level, the type I error rate was 0.044 when using $2(\ell_1 - \ell_2)$ and 0.052 when using $2(\ell_1 - \ell_0)$. Although the difference for 4 taxa is very small, we expect the more conservative nature of $2(\ell_1 - \ell_2)$ to be somewhat compensated in

larger trees by partial optimization of the ML values (see below).

In sum, the null hypothesis can be conveniently replaced by the second best-fitting alternative arrangement around the branch of interest. A branch test formulated using ℓ_2 uniquely chooses between the three competing arrangements around the branch of interest and cannot statistically support all three or two at the same time.

Accuracy and Power of Tree Inference of the aLRT

The aLRT should be able to offer an objective statistical test of an interior branch, as well as a way to calculate branch supports. We tested this for 4-, 12-, and 100-taxon trees, using our definitions of accuracy and power of tree inference.

First, we evaluated the standard and approximate LRTs using 4-taxon non-star trees. The standard power (the proportion of internal branches with significant LRT) was high: 0.833 at $\alpha = 0.01$ and 0.871 at $\alpha = 0.05$. The standard power of the aLRT was almost identical: 0.830 at $\alpha = 0.01$ and 0.867 at $\alpha = 0.05$. For the aLRT, the type I error rate of tree inference (the proportion of incorrectly inferred branches supported by significant aLRT) was close to the significance level: 0.007 at $\alpha = 0.01$ and 0.052 at $\alpha = 0.05$, with very similar results for the standard LRT. The power of tree inference (the proportion correctly inferred branches supported by significant aLRT) was as high as 0.866 at $\alpha = 0.01$ and 0.905 at $\alpha = 0.05$, with the power of the LRT being similarly high. Because the results for 4 taxa were very close with both the standard and approximate LRTs, all further results are presented only for the aLRT.

In our simulations, the type I error rate of tree inference did not seem to depend on tree characteristics such as tree length (S), long branch attraction (LBA) and length (t) of the true interior branch (Fig. 3). But the power of tree inference, intuitively, decreased as tree inference became harder. For example, at $\alpha = 0.05$ we observed an increase of power as S increased from 0.1 to 0.5, since more and more informative sites were available. For $S \approx 0.5$, the power reached an optimum of 0.92 (Fig. 3A); a further increase of S from 0.5 to 3.0 caused a decrease of power to 0.83 (Fig. 3A). The power decreased with an increase of LBA (Fig. 3B), because tree inference becomes more difficult for larger LBA (Felsenstein, 1978). An increase of t facilitated a rapid increase of power (Fig. 3C). However, for trees with a very short interior branch, the power was not absolutely lost (Fig. 3C). In sum, the aLRT based on the Bonferroni-corrected mixture distribution has an acceptable accuracy and a high power of tree inference for 4-taxa trees.

Further, we evaluated properties of the aLRT on larger data sets. First, for 12-taxon trees we compared the performance of the aLRT using partial and full optimization of ℓ_2. When ℓ_2 was optimized only over the branch of interest, the test had unsatisfactory accuracy (results not shown). However, when we optimized ℓ_2 over five branches, the branch of interest and the four adjacent branches, and analyzed the data with the generating
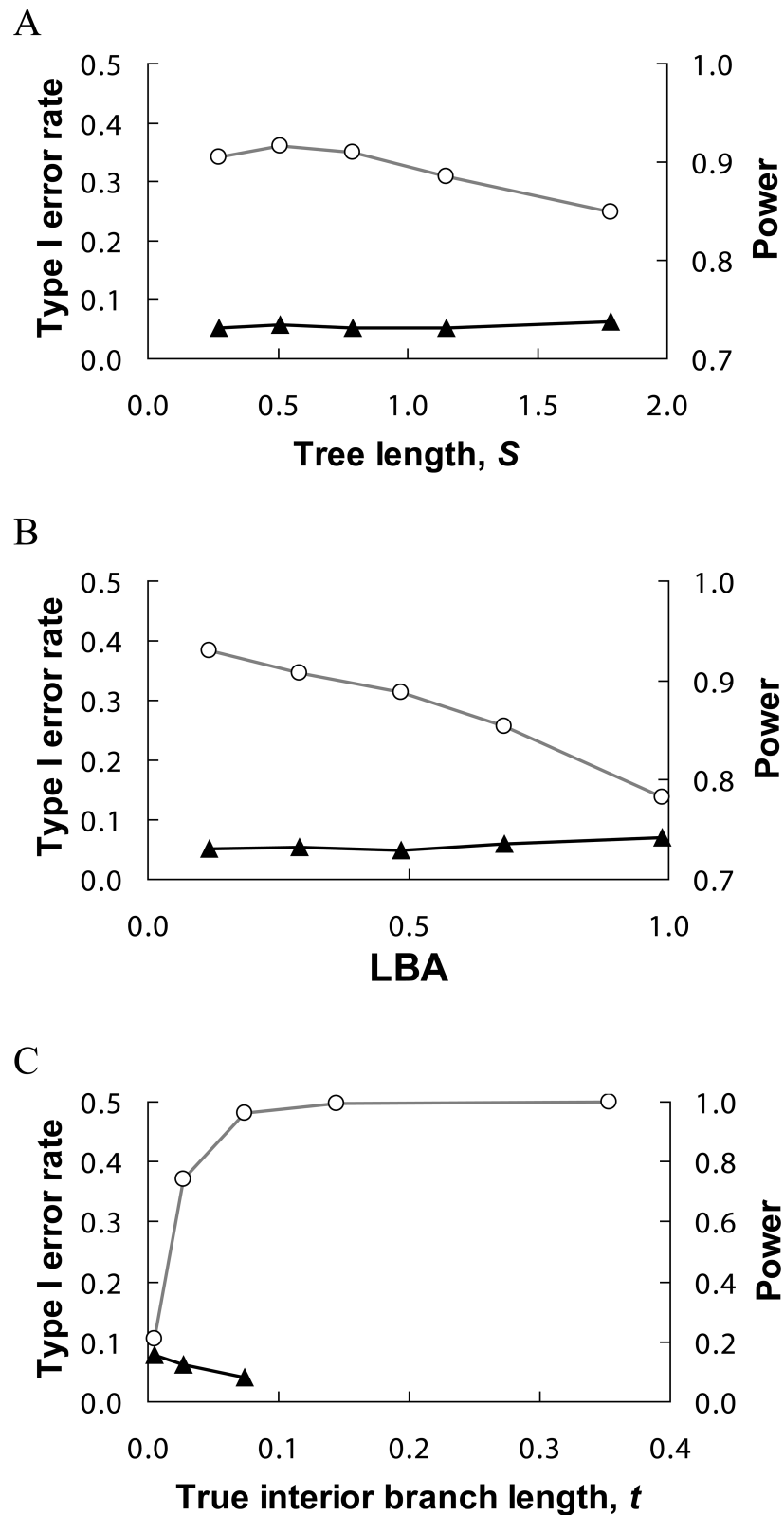
FIGURE 3. The power and the type I error rate of tree inference based on the aLRTs for data with 4 taxa ($\alpha = 0.05$), plotted versus: (A) tree length, S; (B) LBA measure of long branch attraction; (C) interior branch length, t. Lines connecting data points in shape of a hollow circle ($\circ$) illustrate power; lines connecting data points in shape of a black triangle ($\blacktriangle$) illustrate type I error rate. Note that two data points in graph C are missing for the type I error rate line as error could not be calculated due to 100% correct inference.

(HKY$+\Gamma$) model, the type I error rate of tree inference was 0.015 at $\alpha = 0.01$ and 0.048 at $\alpha = 0.05$, whereas the power of tree inference was 0.847 at $\alpha = 0.01$ and 0.890 at $\alpha = 0.05$. When all branches and parameters were reoptimized, the type I error rate was surprisingly similar: 0.015 at $\alpha = 0.01$ and 0.049 at $\alpha = 0.05$, and the power was slightly lower: 0.846 at $\alpha = 0.01$ and 0.859 at $\alpha = 0.05$. Moreover, for 100-taxon trees, the optimization of ℓ_2 over five branches resulted in an accurate test: the type I error rate of the aLRT was 0.052, very close to $\alpha = 0.05$. Given this result, it seems that optimization of more branches is unlikely to bring noteworthy benefits, but would make the calculation slower. Thus, all following results are presented for the aLRT with ℓ_2 optimized only over five branches.

The type I error rate and power of the aLRT ($\alpha = 0.05$) for 12- and 100-taxon data were plotted against maximum pairwise divergence, deviation from molecular clock, and sequence length (Fig. 4). We observed a mild increase of type I error rates for trees with high maximum pairwise divergence (Fig. 4A) or very short sequences (Fig. 4C). But the type I error rate was always close to the significance level. As with 4 taxa, the power reduced for more "difficult" trees or for data sets that were either too similar or too divergent. These patterns were mild but still discernable. For example, there was a tendency for the power to be lower for trees with very small or very large maximum pairwise divergence when data had either little information or essentially randomized sequences (Fig. 4A). Deviations from molecular clock were measured by the ratio of the length of the longest to the length of the shortest lineages, so that perfect molecular clock was indicated by ratio ≈ 1.0. We observed a decline of power with the increasing deviation from molecular clock (Fig. 4B). However, even in the worst cases, the power was as high as 0.86 at $\alpha = 0.05$ (Fig. 4A, B). Other properties relating to tree shape may also influence the power, as tree shape affects the complexity of tree reconstruction. For example, we noticed that the power was higher for more symmetrical trees than for unbalanced trees, but this did not seem to affect the type I error rate (results not shown). However, as expected the most influential parameter is sequence length: for 12 taxa, the power of tree inference was reduced from 0.89 at $\alpha = 0.05$ for 1000 nt to 0.60 at $\alpha = 0.05$ for 100 nt (Fig. 4C).

To test robustness of our aLRT to oversimplification of model assumptions, we considered 12-taxon data sets simulated under the codon model M3 with positive selection. This was the strongest model violation simulated with 4 taxa (see above). Data were analyzed with the incorrect HKY$+\Gamma$ model, which was rejected by the Goldman-Cox test ($P < 0.001$), indicating significant violation. Nevertheless, for the longest sequences (1000 nt), assuming a wrong model did not visibly affect the type I error rate, nor the power of tree inference. For example, at $\alpha = 0.05$ the type I error rate was 0.055 and the power was 0.88 (compared to 0.049 and 0.89, respectively, when the correct model was used in the analysis). For shortest sequences (100 nt), we observed a slightly elevated type I error rate, whereas power was still very similar:

at $\alpha = 0.05$ the type I error rate was 0.079 and the power was 0.56 (compared to 0.066 and 0.60, respectively, when the correct model was used in the analysis).

We also compared the accuracy and the power using 100-taxon data simulated under K2P$+\Gamma$ and the covarion model and analyzed them with HKY$+\Gamma$. Data simulated under the covarion model rejected the analysis model with the Goldman-Cox test ($P < 0.001$). The type I error rate was satisfactory: when the K2P$+\Gamma$ data were analyzed with an overparameterized model, at $\alpha = 0.05$ the type I error rate and power were 0.06 and 0.79, respectively. When the covarion data were analyzed with an incorrect model, at $\alpha = 0.05$ the type I error rate and power were 0.066 and 0.79, respectively.

From the above experiments we may conclude that our fast aLRT, based on $2(\ell_1 - \ell_2)$ statistic and partial optimization, (1) has accuracy and power similar to the standard LRT; (2) provides an accurate branch test even with certain (mild but discernible) model misspecifications. Although none of the models can incorporate full biological complexity, it is advisable to perform the aLRT using a model that reflects the most important trends present in data. For example, such model selection could be done using ModelTest (Posada and Crandall, 1998). Finally, the power of the test is generally high and mostly depends on sequence length, but also may be influenced by factors affecting the complexity of tree reconstruction, such as long branch attraction, elevated maximum pairwise divergence and departure from molecular clock.

Comparison with Branch Tests Based on ML Bootstrap Supports and Bayesian Posterior Probabilities

Regardless of differences in interpretation of ML bootstrap supports and Bayesian posterior probabilities, many researchers subconsciously use these values to make a rule-based decision (e.g., Leaché and Reeder, 2002; Rokas et al., 2003). In other words, the support values are typically compared to a certain threshold, and branches with supports higher than this threshold are considered to be reliable. As a consequence of such decision rule, it is natural to evaluate performance of the branch tests based on ML bootstrap supports or Bayesian posterior probabilities by estimating the type I error rate and the power of tree inference as defined in this paper.

We compared the type I error rates and power of tree inference of the aLRT and branch tests based on nonparametric ML bootstrap supports and Bayesian posterior probabilities using 1500 data sets with 12 taxa. The aLRT outperformed the ML bootstrap with respect to both accuracy and power (Fig. 5). For example, for the aLRT the type I error rate was 0.015 at $\alpha = 0.01$ and 0.049 at $\alpha = 0.05$ (the aLRT is close to exact), whereas the power was 0.848 at $\alpha = 0.01$ and 0.889 at $\alpha = 0.05$. For the ML bootstrap the type I error rate was 0.056 at $\alpha = 0.01$ and 0.086 at $\alpha = 0.05$. Moreover, the error rate remained as high as 0.052 at $\alpha = 0.001$, due to the significant proportion of incorrectly inferred branches with 100% support. This may be attributed to low number of replicates (100) used in our simulation. However, only 100 replicates are
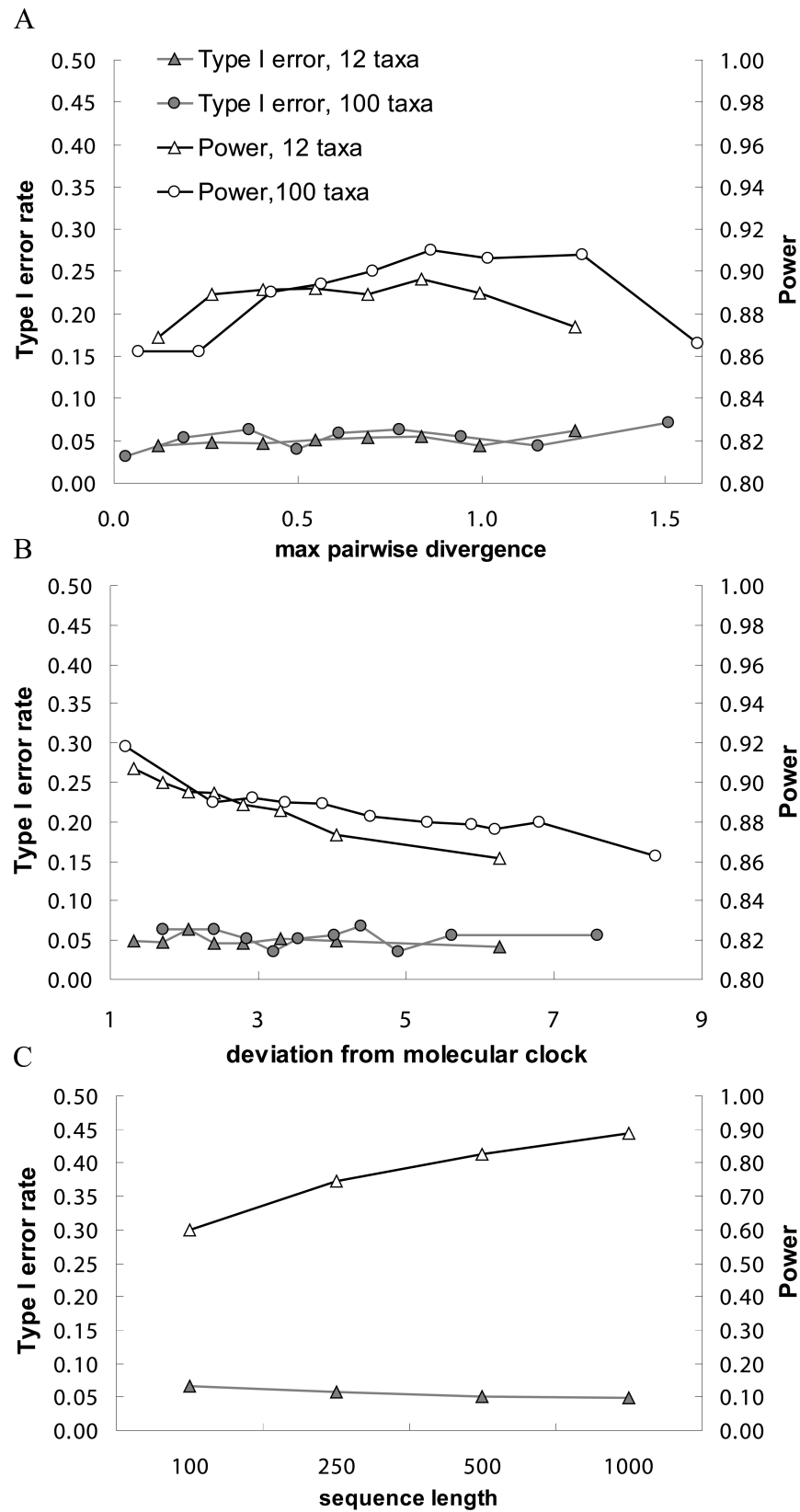
FIGURE 4. The power and the type I error rates of tree inference based on the aLRTs for data with 12 and 100 taxa ($\alpha = 0.05$), plotted against: (A) maximum pairwise divergence; (B) deviation from molecular clock; (C) sequence length.
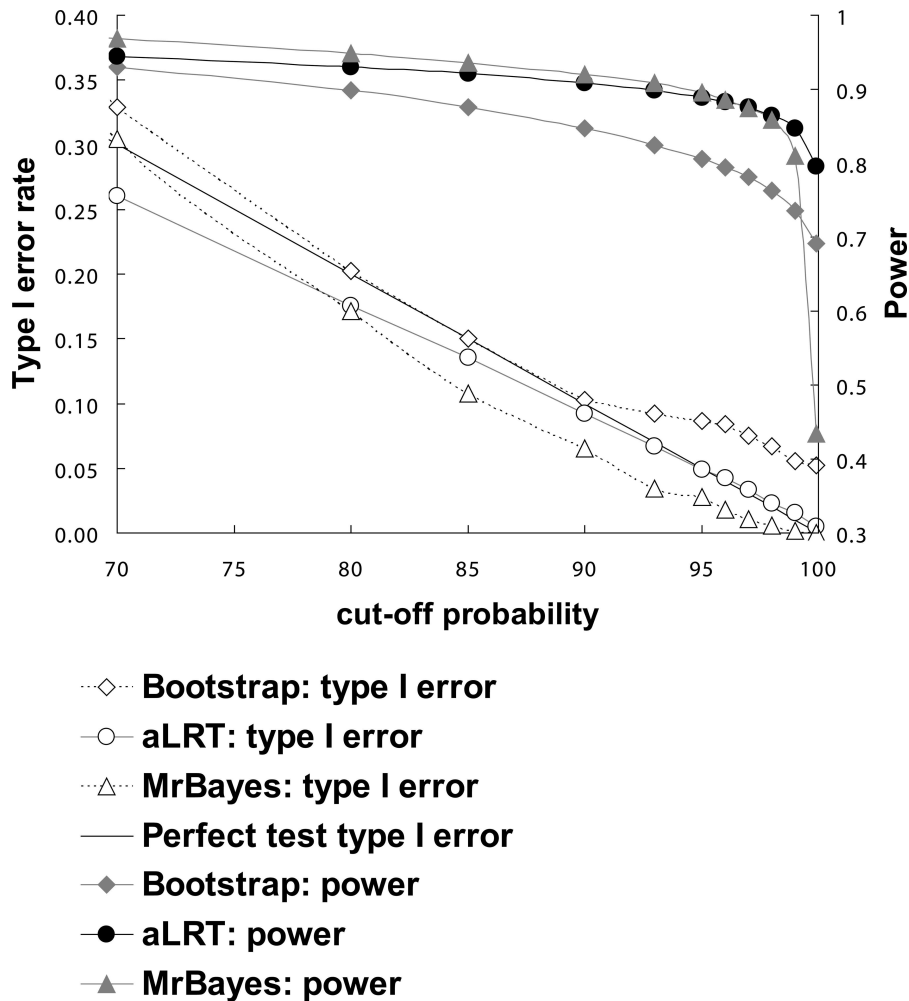
FIGURE 5.   Comparison of the power and the type I error rates of tree inference based on nonparametric ML bootstrap supports, Bayesian posterior probabilities, and the aLRT for 12 taxa, plotted against the cut-off probability $(1 - \alpha)$.

often used in publications as this still requires heavy computing time. Note that although the test was liberal for the ML bootstrap, its power was also lower than the corresponding power of the aLRT (Fig. 5). In sum, for $\alpha \leq 0.1$ the aLRT of an interior branch is almost exact (with the type I error rate $\approx \alpha$), but the ML bootstrap (as a branch test with a cut-off $\geq 1 - \alpha$) is liberal. This might seem contrary to a general belief that nonparametric bootstrap is conservative (although it is accepted that this might not be true in situations that cause inconsistency). However, we suggest that our conclusions are not at odds with previous reports, which focused on the meaning of the bootstrap probabilities rather than the performance of non-parametric bootstrap as a decision rule branch test. The seeming discrepancy is due to differences in measures used to assess the accuracy of bootstrap branch supports in this article and in previous studies. Previous studies (e.g., Hillis and Bull, 1993) plotted the probability of the branch to be true depending on its bootstrap proportion, usually showing that for bootstrap proportions $\geq 80\%$ or even 70% the probabil-

ity of a clade to be true was much higher. Whereas Hillis and Bull (1993) showed this for parsimony bootstrap, we observed a similar behavior for the ML bootstrap in our simulations (Fig. 6). But the resulting curve strongly depends on the simulation settings (e.g., tree shape, deviation from molecular clock), so it is much more preferable to estimate the type I error rate and power on the basis of conditional probabilities, as is done here and as is common in the statistics literature.

In contrast to bootstrap, using Bayesian probabilities made the test slightly conservative for significance levels $\alpha < 0.2$ (Fig. 5). For example, the type I error rate was 0.002 at $\alpha = 0.01$ and 0.027 at $\alpha = 0.05$. Although at $\alpha = 0.05$ the power was almost identical to that of the aLRT, at $\alpha = 0.01$ the power of the Bayesian test decreased to 0.81 (lower than for the aLRT); and for $\alpha = 0.001$, it fell to 0.435, significantly lower than for both, bootstrap and aLRT. In this simulation, branch tests based on Bayesian posteriors appear to be conservative, which may seem surprising. Yet it has been noted previously that, despite being higher than nonparametric
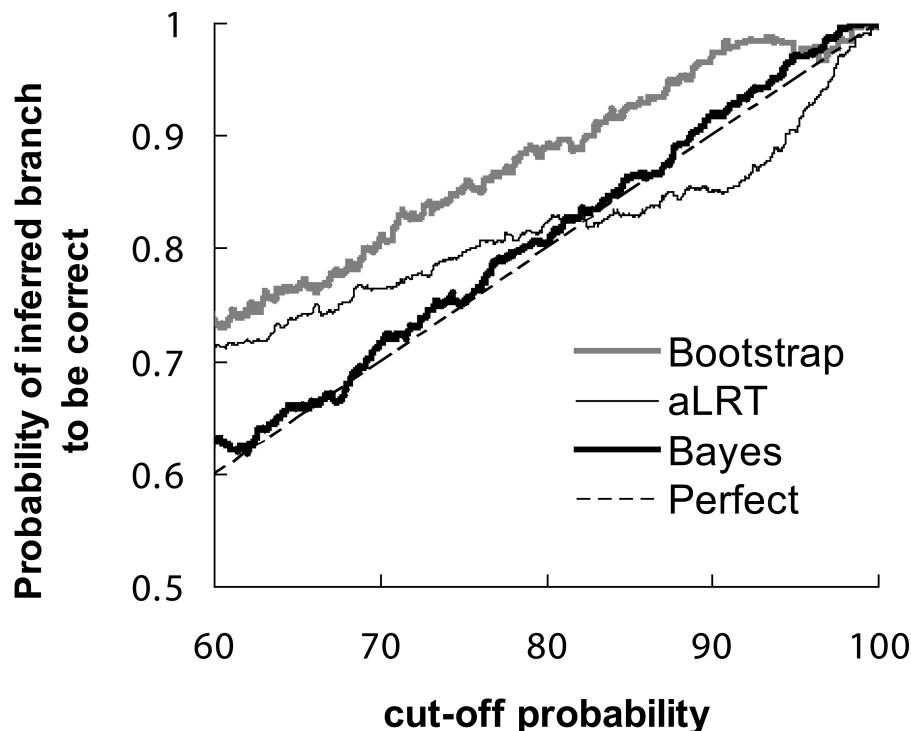
FIGURE 6.   Probability of the inferred branch to be true based on nonparametric ML bootstrap proportions, aLRT supports, and Bayesian posterior probabilities calculated using cubic approximation. Curves are obtained using a sliding window with length 500.

bootstrap values, the Bayesian posterior probabilities should be accurate estimates that a clade is correct if the assumptions of the method are satisfied (Wilcox et al., 2002; Alfaro et al., 2003). Just like in the case of bootstrap, previous discussions about the accuracy of the branch supports concentrated around their meaning. In our simulations, Bayesian posteriors for branches come very close to the actual probabilities (Fig. 6), as indicated by the Bayesian theory and was already shown by simulation (Huelsenbeck and Rannala, 2004). However, this excellent behavior could be affected by model violation (Waddell et al., 2002; Huelsenbeck and Rannala, 2004), including unrealistic branch priors (Yang and Rannala, 2005) and insufficiently long MCMC chains, especially for large taxon samples. Being model-based, the aLRT can also be affected by strong model violations (see above), but with mild violations, elevated type I error rates may be avoided (i.e., the violated model may nevertheless be sufficient). This is critical, because although only some model violations were explored in this study, a variety of unaccounted processes might occur in real data. Finally, it must be understood that aLRT supports and Bayesian posteriors are fundamentally different. There is no theory to suggest that Bayesian posteriors should provide a valid statistical test and therefore fit with our expectations in Figure 5. In turn, there is no theoretical foundation to suggest that aLRT supports should reflect probabilities of a clade to be true, and fit with the scheme in Figure 6. It is thus quite satisfactory to observe that within their own theoretical frameworks

aLRT and Bayesian posteriors are excellent. However, it is also important to keep in mind that both approaches are parametric, and that bootstrap, being nonparametric, could be more robust to serious model violations, which in some way could compensate the fact that it is so hard to give it any simple statistical interpretation.

The main advantage of the aLRT is that it is much faster than either the ML bootstrap or the Bayesian inference. Even though the implementation of the aLRT for this study was largely suboptimal, for 12 taxa, the aLRT was about 5 times faster than performing the ML bootstrap with 100 replicates, and about 10 times faster than the Bayesian method with 2 × 30,000 generations. However, many more bootstrap replicates and much longer MCMC runs are usually required, which dramatically increases the computational time rendering the aLRT especially useful for large-taxon data. It would be of interest to compare this aLRT to the recent fast approximations (e.g., Waddell et al., 2002) of nonparametric bootstrap supports (such as RELL) and Bayesian posterior probabilities (BIC and BIC-J).

The aLRT was developed within the ML tree estimation software PHYML (Guindon and Gascuel, 2003; http://www.lirmm.fr/atgc/phyml) and an efficient implementation will be available as an option in the next release version of PHYML upon publication.

REFERENCES

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Berry, V., and O. Gascuel. 1996. On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. Mol. Biol. Evol. 13:999–1011.

Buckley, T. R., P. Arensburger, C. Simon, and G. K. Chambers. 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. Syst. Biol. 51:4–18.

Chernoff, H. 1954. On the distribution of the likelihood ratio. Ann. Math. Stat. 25:573–578.

Cowles, M. K. and B. P. Carlin. 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review. J. Am. Stat. Assoc. 91:883–904.

Cummings, M. P., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

Desper, R., and O. Gascuel. 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. 21:587–598.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. Douzery. 2003a. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

Douady, C. J., M. Dosay, M. S. Shivji, and M. J. Stanhope. 2003b. Molecular phylogenetic evidence refuting the hypothesis of Batoidea (rays and skates) as derived sharks. Mol. Phylogenet. Evol. 25:215–221.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, UK.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:13429–13434.

Efron, B., and R. Tibshirani. 1993. An introduction to the bootstrap. Chapman and Hall, New York.

Erixon, P., B. Svennblad, T. Britton, and B. Oxelman. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52:665–673.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 1985. Confidence intervals on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sanderland, Massachusetts.

Felsenstein, J., and H. Kishino. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. 42:193–200.

Galtier, N. 2004. Sampling properties of the bootstrap support in molecular phylogeny: Influence of nonindependence among sites. Syst. Biol. 53:38–46.

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.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo. Stat. Sci. 7:473–483.

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.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Huelsenbeck, J., and B. Rannala. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Larget, B. and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Leaché, A. D., and T. W. Reeder. 2002. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): A comparison of parsimony, likelihood, and Bayesian approaches. Syst. Biol. 51:44–68.

Lewis, P. O., M. T. Holder, and K. E. Holsinger. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.

Li, S. 1996. Phylogenetic reconstruction using Markov chain Monte Carlo. Ph.D. dissertation. The Ohio State University, Columbus.

Mau, B., M. A. Newton, and B. Larget. 1997. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Mol. Biol. Evol. 14:717–724.

Miller, R. G. 1981. Simultaneous statistical inference. Springer-Verlag, New York.

Ota, R., P. J. 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.

Posada, D., and K. A. Crandall. 1998. ModelTest: Testing the model of DNA substitution. Bioinformatics. 14:817–818.

Posada, D., and K. A. Crandall. 2001. Selecting models of nucleotide substitution: An application to human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. 18:897–906.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Sanderson, M. J. 1995. Objections to bootstrapping phylogenies: A critique. Syst. Biol. 44:299–320.

Self, S. G., 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.

Siddall, M. E. 2002. Measures of support. Pages 80–101 *in* Techniques in molecular systematics and evolution (R. DeSalle, G. Giribet, and W. Wheeler, eds.). Birkhäuser Verlag, Basel.

Simmons, M. P., K. M. Pickett, and M. Miya. 2004. How meaningful are Bayesian support values? Mol. Biol. Evol. 21:188–199.

Stuart, A., J. K. Ord, and S. Arnold. 1999. Kendall's advanced theory of statistics. Oxford University Press, New York.

Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50:723–729.

Suzuki, Y., G. Glazko, and N. Masatoshi. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99:16138–16143.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.) Sinauer Associates, Sunderland, Massachusetts.

Tateno, Y., N. Takezaki, and M. Nei. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony

methods when substitution rate varies with site. Mol. Biol. Evol. 11:261–277.

Taylor, D. J. and W. H. Piel. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol. Biol. Evol. 21:1534–1537.

Waddell, P. J., H. Kishino, and R. Ota. 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. Genome Inform. Ser. Workshop Genome Inform. 13:82–92.

Wilcox, T. P., D. J. Zwickl, T. A. Heath, and D. M. Hillis. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol. Phylogenet. Evol. 25:361–371.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

Yang, Z., and B. Rannala. 2005. Branch-length prior influences bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.

Zharkikh, A., and W. H. Li. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. Four taxa without a molecular clock. J. Mol. Evol. 35:356–366.