Syst. Biol. 60(5):685–699, 2011

DOI:10.1093/sysbio/syr041
Advance Access publication on May 3, 2011

Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes

MARIA ANISIMOVA[1,2,3,*], MANUEL GIL[1,2], JEAN-FRANÇOIS DUFAYARD[3], CHRISTOPHE DESSIMOZ[1,2], AND OLIVIER GASCUEL[3]

[1]*Department of Computer Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland;* [2]*Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland; and* [3]*Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS—Université Montpellier 2, Montpellier, France;*
[*]*Correspondence to be sent to: Maria Anisimova, Department of Computer Science, ETH Zürich, Universitaetsstrasse 6, 8092 Zürich, Switzerland; E-mail: maria.anisimova@inf.ethz.ch.*

Received 16 March 2010; reviews returned 2 September 2010; accepted 1 March 2011
Associate Editor: Tiffani Williams

Abstract.—Phylogenetic inference and evaluating support for inferred relationships is at the core of many studies testing evolutionary hypotheses. Despite the popularity of nonparametric bootstrap frequencies and Bayesian posterior probabilities, the interpretation of these measures of tree branch support remains a source of discussion. Furthermore, both methods are computationally expensive and become prohibitive for large data sets. Recent fast approximate likelihood-based measures of branch supports (approximate likelihood ratio test [aLRT] and Shimodaira–Hasegawa [SH]-aLRT) provide a compelling alternative to these slower conventional methods, offering not only speed advantages but also excellent levels of accuracy and power. Here we propose an additional method: a Bayesian-like transformation of aLRT (aBayes). Considering both probabilistic and frequentist frameworks, we compare the performance of the three fast likelihood-based methods with the standard bootstrap (SBS), the Bayesian approach, and the recently introduced rapid bootstrap. Our simulations and real data analyses show that with moderate model violations, all tests are sufficiently accurate, but aLRT and aBayes offer the highest statistical power and are very fast. With severe model violations aLRT, aBayes and Bayesian posteriors can produce elevated false-positive rates. With data sets for which such violation can be detected, we recommend using SH-aLRT, the nonparametric version of aLRT based on a procedure similar to the Shimodaira–Hasegawa tree selection. In general, the SBS seems to be excessively conservative and is much slower than our approximate likelihood-based methods. [Accuracy; aLRT; branch support methods; evolution; model violation; phylogenetic inference; power; SH-aLRT.]

Computing and evaluating tree branch supports—measures of confidence in given branches—are indispensable parts of phylogenetic inference. In particular, support measures are crucial to validating or refuting biological hypotheses on the basis of trees (e.g., Baum et al. 2005). Parallel to the development of phylogenetic inference methods, various measures of branch support have been proposed (for review see Wrobel 2008). In the statistical paradigm, the perhaps three most desirable properties of a branch support measure are high accuracy, power, and robustness. High accuracy implies that under the true model, incorrectly inferred branches should not be statistically supported. High power implies that correctly inferred branches should have high statistical support. As for high robustness, it conveys the notion that modeling inadequacies—which are unavoidable when dealing with real biological data—do not strongly affect the accuracy of the measure. These three properties are typically validated in simulations, where the model, tree, and all other parameters are known. However, conclusions drawn from simulation studies are highly dependent on the simulation design: size and properties of the synthetic test data, the analysis model, nature and extent of model violations introduced to assess robustness, and the approach used to characterize the properties of the support measure. Thus, simulation studies need to be complemented with empirical tests based on real biological data. Although in real data, the random and the systematic errors are hard to separate, simulations with model violations may be used to explore the methods' biases, if only for very specific and very few simulation scenarios.

In this work, we focus on likelihood-based support measures that are used in the model-based tree inference by maximum-likelihood (ML) or the Bayesian approach. In recent years, likelihood-based approaches have established themselves as methods of choice. In particular, since the introduction of ML in phylogenetics (Felsenstein 1981), the heuristics for ML tree search have steadily improved in terms of efficiency and speed (Lemmon and Milinkovitch 2002; Guindon and Gascuel 2003; Hordijk and Gascuel 2005; Stamatakis and Ott 2008; Stamatakis et al. 2008; Guindon et al. 2010). Although nonparametric ML bootstrap (Efron 1979; Felsenstein 1985) and the Bayesian approach (Rannala and Yang 1996; Larget and Simon 1999; Mau et al. 1999) are popular for evaluating branch supports, their interpretation and accuracy has been often disputed (see discussion in Anisimova and Gascuel 2006). Furthermore, even with fastest tree search or Markov chain Monte Carlo (MCMC) sampling algorithms, these classical techniques become slow and even impractical with an increase of sample size and sequence length. Recently, fast approximate methods to evaluate branch supports have been proposed, which make it possible to compute

branch supports for trees with several hundred taxa in less than an hour on a standard computer. Anisimova and Gascuel (2006) developed an approximate likelihood ratio test (aLRT), a frequentist test that compares two best nearest-neighbor interchange (NNI) configurations around the branch of interest. In simulations, this test was shown to be accurate, powerful, and robust to moderate model misspecifications. Tests on large real data sets demonstrated that for a fixed type I error rate threshold α, the frequency of inconsistent supports on suboptimal trees for aLRT is $\leq \alpha$, although some dependency on the inferred tree is expected (Stamatakis et al. 2008). As with any model-based method, more serious models violations may cause the aLRT to become overconfident. Typically, the use of a more conservative test helps to reduce the excess of false positives (FPs). This is often achieved by introducing a nonparametric element within an algorithm for evaluating branch supports. For example, nonparametric bootstrap may be used to reduce Bayesian posterior branch supports, which are typically perceived as too high (Douady et al. 2003). However, this requires computationally intensive MCMC approximation of Bayesian posterior probabilities (PP) to be done multiple times ($\geq$100), making such a procedure prohibitively expensive, especially with large samples or limited computing power. One alternative is a fast nonparametric version of the aLRT (Shimodaira–Hasegawa [SH]-aLRT), which was developed and implemented in the PHYML phylogenetic inference software (Guindon et al. 2010). SH-aLRT is derived from the SH multiple tree comparison procedure (Shimodaira and Hasegawa 1999) and is fast due to the RELL technique based on the resampling of estimated log likelihoods (Kishino and Hasegawa 1989). Additionally, here we propose a new simple *à la Bayes* modification of the aLRT for rapid and accurate approximation of branch supports (aBayes). Another solution is a fast approximation to the standard nonparametric ML bootstrap (rapid bootstrap or RBS), implemented in the other popular fast phylogenetic ML inference program RAxML (Stamatakis et al. 2008). The RBS supports exhibited strong correlation with standard bootstrap (SBS) values on several large real data sets (Stamatakis et al. 2008). However, with real data, the true tree topology is typically unknown, making it hard to evaluate the accuracy of the method. Given the popularity of the bootstrap, it is of a considerable interest to investigate and compare the properties of standard and RBS procedures. Although results from simulation studies cannot be directly transferred on to real data, computer simulations play an essential role in evaluating method performance because the true simulation scenario (e.g., tree topology) is then known. Here, we present results from a computer simulation evaluating four approximate tests (aLRT, SH-aLRT, aBayes, and RBS), and make comparisons with SBS and PP supports where possible.

Clearly, values of branch support computed with different methods have different interpretation and are difficult to compare. Typically, researchers inferring phylogenies use rules of thumb (depending on the method) to guide their decision making. For example, a posterior probability value PP = 0.9 is generally considered insufficient, whereas a bootstrap value of the same magnitude serves as evidence of high support.

Here we take a pragmatic approach and use simulations to evaluate the error rates produced by each method for a given support value. Although the results of simulations should not be overgeneralized, they provide a rough idea of the methods robustness and should help in guiding the decisions of how high a support value should be to provide sufficient evidence for a branch. In addition, we complement our analyses by evaluating all methods in the traditional probabilistic sense that is compared with an estimated true probability, which has been done in numerous studies of bootstrap and Bayesian supports.

Despite the differences in method formulation, we suggest that a threshold-based evaluation of branch supports is more meaningful for all methods. Although the full Bayesian method is expected to have the probabilistic interpretation under a true model, two approximate tests (aBayes and RBS) may also approximately satisfy the probabilistic interpretation in cases where model assumptions are not seriously violated. Finally, we illustrate the behavior of different methods for estimating branch supports on real data, where contradicting results have been previously reported.

Methods and Data

Existing Methods of Branch Support

Phylogeny inference by ML requires optimization of the log-likelihood function $\ell(\mathbf{T}, \boldsymbol{t}, \theta|\mathrm{D}) = log\ \Pr(\mathrm{D}|\mathbf{T}, \boldsymbol{t}, \theta)$ over the space of possible topologies $\mathbf{T}$, branch lengths $\boldsymbol{t}$, and evolutionary model parameters θ. Argument values maximizing the likelihood of observing sequence data D are the ML estimates of the topology, its branch lengths, and model parameters; these estimates are denoted as $(T_{\mathrm{ML}}, t_{\mathrm{ML}}, \theta_{\mathrm{ML}}) = \mathrm{argmax}\ \ell(\mathbf{T}, \boldsymbol{t}, \theta|\mathrm{D})$. A tree found in a heuristic ML search is better described as the "best known" ML tree, as heuristic algorithms do not guarantee that the global ML tree is found. Two fast implementations of ML tree search heuristics employed in this study are PHYML (Guindon and Gascuel 2003; Guindon et al. 2010) and RAxML (Stamatakis et al. 2005; Stamatakis 2006).

Approximate tests implemented in PHYML compare optimized log likelihoods for three NNI configurations around the branch of interest: one optimal and two suboptimal. For suboptimal configurations, log likelihoods are optimized only over five branch lengths (those of the branch of interest and the four adjacent branches), whereas other branch lengths and all model parameters are kept at their ML estimates. Log likelihoods for the three configurations may be ordered, so that ℓ_1, ℓ_2, and ℓ_3, denote the best ML score, second best and the worst, respectively. Note that $\ell_1 = \ell(\mathrm{T}_{\mathrm{ML}}, t_{\mathrm{ML}}, \theta_{\mathrm{ML}}|\ \mathrm{D})$.

The aLRT for a branch evaluates the statistic $2(\ell_1 - \ell_2)$, that is, double the log-likelihood difference for the best known ML configuration and the second best NNI rearrangement around the branch of interest (Anisimova and Gascuel 2006). The significance of a branch support is tested based on the comparison of the Bonferroni-corrected test statistic with the $0.5\chi_0^2 + 0.5\chi_1^2$ distribution. The resulting P values may be converted into support values ranging from 0.125 to 1 (Anisimova and Gascuel 2006). Note that this test is related to the interior branch test previously studied for distance and likelihood tree inference.

The aLRT with the nonparametric SH correction (SH-aLRT) uses a routine developed in the spirit of the Shimodaira–Hasegawa (SH) algorithm for tree comparison (Guindon et al. 2010). Assuming site independence, the log-likelihood ℓ_C for configuration C is a sum of site log likelihoods over the alignment of length n. Pseudoreplicates are generated by resampling n sites with replacement. For each pseudoreplicate, log-likelihoods ℓ_C^* for configuration C may be quickly computed by summing up the original site log likelihoods for configuration C corresponding to resampled positions (RELL procedure). The condition $\ell_1^* \geq \ell_2^* \geq \ell_3^*$ may no longer hold for resampled data. Starting with the branch support count SH = 0, for each pseudoreplicate the following procedure is repeated:

(a) using the bootstrapping property $E(\ell_C^*) = \ell_C$, center pseudolikelihoods $\ell_C^0 = \ell_C^* - \ell_C$ for each C;
(b) order centered values $\{\ell_C^0\}$ so that $\ell_{\text{best}}^0 = \max_C\{\ell_C^0\} \geq \ell_{\text{next}}^0 \geq \ell_{\text{worst}}^0$;
(c) increase SH by 1 if $\ell_1 - \ell_2 \geq \ell_{\text{best}}^0 - \ell_{\text{next}}^0$.

The SH-aLRT branch support is measured by the proportion of replicates for which condition in (c) holds.

Finally, the RBS heuristic was developed within the popular package RAxML to accelerate the computation of SBS (Stamatakis et al. 2008). RAxML uses the Lazy Tree Rearrangement algorithm to search the tree space (Stamatakis et al. 2005), similar to the Subtree Pruning and Regrafting algorithm implemented in PHYML (Hordijk and Gascuel 2005). During the RBS search, a flexible model (GTR) is imposed, and an approximation of the gamma distribution is used for rate heterogeneity. This strategy facilitates rapid progress towards higher likelihood areas of the tree space.

Bayesian-like Modification of the aLRT (aBayes)

Let T_C represent the topology corresponding to one of the three NNI configurations around the branch of interest. We approximate the posterior probability of configuration C using the Bayes rule:

$$\Pr(T_c|D) = \frac{\Pr(\mathrm{D}|T_c)\Pr(T_c)}{\sum_{i=1}^{3}\Pr(\mathrm{D}|T_i)\Pr(T_i)}$$

and assuming only three possible configurations (with no rearrangements within subtrees) with a flat prior $\Pr(T_1) = \Pr(T_2) = \Pr(T_3)$. Log likelihoods of the three configurations are reused in this calculation as $\log \Pr(D|T_C) = \ell_C$. Despite these crude assumptions, made for simplicity and speed benefits, this *à la Bayes* interpretation of the aLRT branch support is an interesting showcase of an extreme approximation, as we demonstrate below its clear advantages over most branch tests. Note that this approach appears similar to using likelihood weights to puzzle quartets (Strimmer and Rambaut 2002) and an earlier suggestion of likelihood mapping for assessing the phylogenetic content of a sequence alignment for quartets (Strimmer and Von Haeseler 1997).

Speed

Our fast approximate likelihood-based methods (aBayes, aLRT, SH-aLRT) do not significantly augment the time required for the ML tree inference because all the calculations are based on fast approximations and reuse intermediate likelihood values obtained during heuristic tree search. In other words, the time required to perform ML tree inference with PHYML with one of the fast approximate methods is roughly the same as the time spent to only infer an ML tree (with no branch support estimation). In contrast, SBS with 100 replicates requires 100 times longer. RBS takes on average a third of the time of SBS (both with 100 replicates), but the exact time varies depending on the size and the properties of the data set and the model used for the inference.

Availability

SH-aLRT and aLRT are available in PHYML v. 3.0 (Guindon et al. 2010). The implementation of the aBayes method is now included in the current version of the PHYML program, which is available for download (http://www.atgc-montpellier.fr/phyml).

Evaluation of Performance in Simulations

Because the methods under scrutiny are nonhomogeneous in the statistical sense, we take a practical approach to evaluate error rates committed for a given fixed support value threshold. This strategy resembles the so-called approximate *frequentist framework* (Anisimova and Gascuel 2006). We define the approximate measures of FP and false-negative (FN) error rates for a given threshold α:

$$\text{FP rate}\ (\alpha) = \Pr\ (\text{support} \geq 1 - \alpha \mid \text{branch is NOT correct})$$

$$\text{FN rate}\ (\alpha) = \Pr\ (\text{support} < 1 - \alpha \mid \text{branch is correct})$$

For aLRT, developed in the frequentist framework, the support $\geq 1-\alpha$ when the branch test is significant at level α (and support $< 1-\alpha$ when the branch test is not significant at α). This means that an aLRT support of 0.95 corresponds roughly to a 5% FP rate. This is because for the aLRT, the threshold is unambiguously defined as a size of test (or significance level) because

it relies on the theoretical asymptotic distribution of the test statistic under the null hypothesis. For other support methods, such correspondence is not expected. Instead, the FP and FN rates are error rate estimates that one may expect for a given support value $(1-\alpha)$.

The tree topology is not fixed in our simulations, so these probabilities must be estimated as average frequencies over all branches in the tree (rather than focusing on one particular branch in the same topology over simulated replicates). Therefore, controlling this estimated overall FP rate for a tree is equivalent to controlling the false discovery rate FDR, which recently became the method of choice in controlling for multiple testing in a variety of situations (Storey 2002).

Ideally, the desirable test has a low FP rate (or high accuracy, i.e., 1− FP rate) and a high power (1− FN rate). When the FP rate is reduced, the power is also reduced, and vice versa. Often an FP error is considered to be a worse mistake than an FN error. Therefore, the foremost requirement for a method is to have an acceptable FP error rate. Accurate approaches (i.e., with low FP) are then further compared on the basis of their power (the higher the better).

The process of branch testing may also be considered as a binary classification procedure, where a branch is labeled "correct" if it has sufficiently high support and "incorrect" otherwise. With a perfect support measure, only true branches are labeled as "correct," and all branches absent from the true tree are classified as "incorrect." Such binary prediction may be assessed by the Matthews correlation coefficient: $\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$, where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Bayesian posterior clade probabilities (unlike other types of branch supports) by definition should have the *probabilistic interpretation* (Huelsenbeck et al. 2002; Huelsenbeck and Rannala 2004), so that the estimated posterior probability for a clade is very close to the true probability of this clade being correct. This nice property, however, is often affected by model violations (Huelsenbeck and Rannala 2004). Several authors interpreted SBS supports as clade probabilities (Hillis and Bull 1993; Efron et al. 1996; Murphy et al. 2001). First objections to such an interpretation (Felsenstein and Kishino 1993) were supported by evaluations in the probabilistic framework (Berry and Gascuel 1996; Anisimova and Gascuel 2006). Although the probabilistic interpretation is not expected to be strictly applicable to any of the described approximate measures, we use the probabilistic framework to contrast methods properties and to verify whether any of the support methods produce values close to probabilistic interpretation.

For simulated data, because the true tree is known, the probabilities of an inferred clade to be true (found in the true tree) can be estimated. For all inferred trees in our simulation, we first ordered all inferred bipartitions (i.e., interior branches) by their estimated branch supports. Next, going over this ordered list in sliding windows of size 100 with step 10, we estimated the probabilities of an inferred clade (defined by a bipartition) to be true as the proportions of correctly inferred bipartitions in each window. For each window, the estimated probability of an inferred clade to be true was compared with the average inferred branch support in the same window.

Simulated Data

The performance of the above-mentioned measures was compared on 1000 simulated replicates, each with 100 taxa and 600 nucleotides and based on a random tree (data from Desper and Gascuel 2004). Trees were simulated using the beta-splitting model (Aldous 1996), which generalizes the uniform distribution on phylogenies and the standard Yule–Harding branching process (Yule 1925; Harding 1971), both of which are typically used to generate a distribution of biologically relevant trees. Deviations from molecular clock were introduced to each tree (Desper and Gascuel 2004). Sequence data were generated using the K2P + covarion model, similar to Galtier (2001), where evolutionary rates vary among sites and over time. Analyses were performed under the incorrect models HKY+Γ_4 (moderate model violation) and JC+Γ_4 (serious violation). For comparison with the Bayesian approach and with the results from our previous study, we used 1500 smaller simulated data sets, each generated under HKY+Γ_4 with 12 taxa and 1000 nucleotides, and based on a distribution of phylogenies generated using the standard speciation process with deviations from the molecular clock (data from Anisimova and Gascuel 2006). The data were analyzed under both the correct model HKY + Γ_4 and the incorrect model JC + Γ_4. The Bayesian MCMC analyses were conducted with MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001) as described in Anisimova and Gascuel (2006). To address concerns that 4×10^4 generations used in our previous study may not have been sufficient (despite good convergence diagnostics), we also run longer chains (4×10^5 generations) under each model.

It has previously been noticed that bootstrap proportions as well as PP can be too high not only for incorrect but also for nonexisting (i.e., zero-length) branches (e.g., Lewis et al. 2005; Yang 2007; Guindon et al. 2010). Thus, we tested how often branch partitions were inferred with high supports on star-like data, as can be the case for viral data or samples of deep divergence confounded by selection (adaptive radiation). We simulated 100- and 12-taxa star trees with branches drawn from the exponential distribution with a mean of 0.1 expected substitutions per branch per site. All star trees were simulated under HKY + Γ_4.

Real Data

The degree of agreement and correlation between branch support measures was assessed on real data. Eight highly diverse data sets from the comparative

study of bootstrap and Bayesian posterior supports (Douady et al. 2003) were analyzed with PHYML and RAxML, and different methods were used to estimate clade supports. The choice of these data was motivated by an apparent conflict of highly supported bipartitions. Because robustness to model violations is part of our assessment, we do not perform a model selection step (like in Douady et al. 2003), but instead use the sufficiently complex HKY + Γ_4 model for all data. We acknowledge that this model provides a certain degree of violation for each data set.

For further tests, we selected eight amino acid genes from metazoa lineages that were used to reconstruct the relative position of bilaterian animals (Philippe et al. 2005). Conflicting inferences from such multilocus protein data were previously reported (Lartillot et al. 2007). All bilaterian protein sequences were analyzed under WAG + Γ_4. Finally, we performed a detailed analysis of the *rbcL* plant data set containing 500 taxa and 1428 nucleotides (previously analyzed in Guindon and Gascuel 2003). The main features of all real data sets analyzed here are listed in Table 1.

Results and Discussion

The Effect of Model Misspecifications on the Accuracy of Branch Supports

For simulated data, we compared the performance of various branch support measures under mild or strong model violations, in both the frequentist and probabilistic frameworks. Analyses assuming the HKY + Γ model represent mild model violation (detected by the Goldman–Cox test; Goldman 1993), whereas JC + Γ is a more serious model violation (e.g., compared with the HKY + Γ model, JC + Γ has much higher Akaike information criterion [AIC] score; Akaike 1973). However, the degree of deviation from the true model varies among simulated replicates, with the differences of AIC scores under the two models measuring the severity of using JC instead of HKY for a particular data set. We observed weak but significant positive correlation between the distance from the more flexible to the simple model and the FP rate ($r^2 = 0.1$, P value = 0.002; Suppl. Fig. S1, available from http://www.sysbio.oxfordjournals.org/).

Table 1. Simulated and real data sets used for the comparison of branch support methods

Data set description	Data type	No. taxa	Sequence length	Gaps or missing (%)	Phylogenetic signal[a]
Simulated data					
(A) 1000 replicates from Desper and Gascuel (2004)	DNA	100	600	None	Distribution with mean = 0.26
(B) 1500 replicates from (Anisimova and Gascuel 2006)	DNA	12	1000	None	Distribution with mean = 0.48
(C) 1000 replicates simulated on large random star trees; simulation model HKY + Γ	DNA	100	600	None	0
(D) 1000 replicates simulated on random star trees; simulation model HKY + Γ	DNA	12	1000	None	0
Real data from Douady et al. (2003)					
(1) Orchids, nuclear ribosomal *ITS*	DNA	23	682	7.93	0.26
(2) Mammals, nuclear protein-coding *vWF*	DNA	13	1161	0.23	0.50
(3) Insects 1, nuclear protein-coding EF1α	DNA	14	2033	6.15	0.18
(4) Insects 2, mitochondrial (12*S*−16*S* rRNA, *COI*, *COII*)	DNA	14	2249	0.06	0.17
(5) Sharks 1, mitochondrial (12*S*−16*S* rRNA)	DNA	23	1880	3.24	0.19
(6) Sharks 2, mitochondrial (12*S*−16*S* rRNA)	DNA	21	1963	2.58	0.24
(7) Snakes, mitochondrial (12*S*−16*S* rRNA)	DNA	23	1545	6.46	0.32
(8) 3 domains of life, *HMGR*	AA	15	258	19.10	0.56
Real Metazoan proteins from Lartillot et al. (2007)					
(9) *sap40*	AA	49	190	3.71	0.29
(10) *rpl5*	AA	47	249	5.07	0.27
(11) *vata*	AA	39	598	24.57	0.32
(12) *yif1p*	AA	26	145	15.12	0.31
(13) *glcn*	AA	25	215	16.71	0.33
(14) *rpo-A*	AA	27	713	34.78	0.34
(15) *rpo-B*	AA	33	1145	31.94	0.31
(16) *nsf2-F*	AA	26	414	13.41	0.30
Real data from test set used in Guindon and Gascuel (2003)					
(17) Protein-coding *rbcL*	DNA	500	1398	2.25	0.28

[a]The phylogenetic signal is the proportion of the total tree length that is taken up by internal branches (Phillips et al. 2001).

In the frequentist framework, all methods appeared accurate with moderate model violations (analysis with HKY + Γ). We evaluated the performance of methods by estimating the error rates for several fixed thresholds of support values, paying particular attention to error rates achieved at high threshold values such as 0.99, 0.95, and 0.9 (Fig. 1). Recall that for aLRT a support threshold of 0.95 corresponds to 5% significance level. This means that among all branches reconstructed with an aLRT support of ≥ 0.95, the FP rate should not exceed 5%. Although this requirement may not hold for other methods, low levels of FP rates respective to the support threshold is clearly a desirable property for any test. Moreover, plotting the FP rates for all methods as a function of the support threshold gives the user a valuable information on what can be expected for certain values of supports for a given method, and how the same absolute value of support compares with supports obtained by other methods (in terms of FP error rates).

In 100-taxa data sets, for a threshold of 0.95 FP rate remained below 5% for all methods with the exception of aLRT, which was slightly above—with 6.3% (Table 2 and Fig. 1a). Such behavior of the aLRT under moderate model misspecifications is consistent with our earlier report (Anisimova and Gascuel 2006). More conservative tests typically suffer from loss of power. Indeed, the least conservative methods aLRT and aBayes had clearly a higher power than the other measures (Table 2 and Fig. 1a). Consequently, with moderate model violations, aLRT and aBayes appear to be the preferred methods.

Under more serious model violations (analysis with JC+Γ), FP rate of aLRT increased to 10% at the 5% significance level. All other tests, however, remained accurate (Table 2 and Fig. 1b). An elevated FP error rate for aLRT may be due to its explicit use of the theoretical distribution under the null hypothesis, which is known to be sensitive to model misspecifications. Although some increase of FP rate was also observed for other methods, error rates remained low, showing better resistance to model misspecifications.

As a consequence of the accuracy power trade-off, the power of the methods also slightly increased (Table 2 and Fig. 1b). Note that RBS could not be performed under either HKY + Γ or JC + Γ because RAxML enforces GTR + Γ as part of its heuristic algorithm. When compared with other methods performed under HKY + Γ, RBS has good accuracy and relatively good power (Fig 1a). Such competitive performance may be due to using a more flexible model (GTR + Γ) to navigate to more promising areas of tree space. Although RBS compares favorably with both SBS and SH-aLRT in

TABLE 2. FP error rate (FP rate) and power of branch support methods for simulated data set (A in Table 1) for a threshold of 0.95

Analysis model	Support method	FP rate (%)	Power (%)
HKY + Γ			
	aLRT	6.3	79
	aBayes	2.8	71
	SH-aLRT	0.2	36
	SBS	0.3	48
	RBS	0.2	35
JC + Γ			
	aLRT	10	81
	aBayes	5	74
	SH-aLRT	0.3	38
	SBS	0.3	35

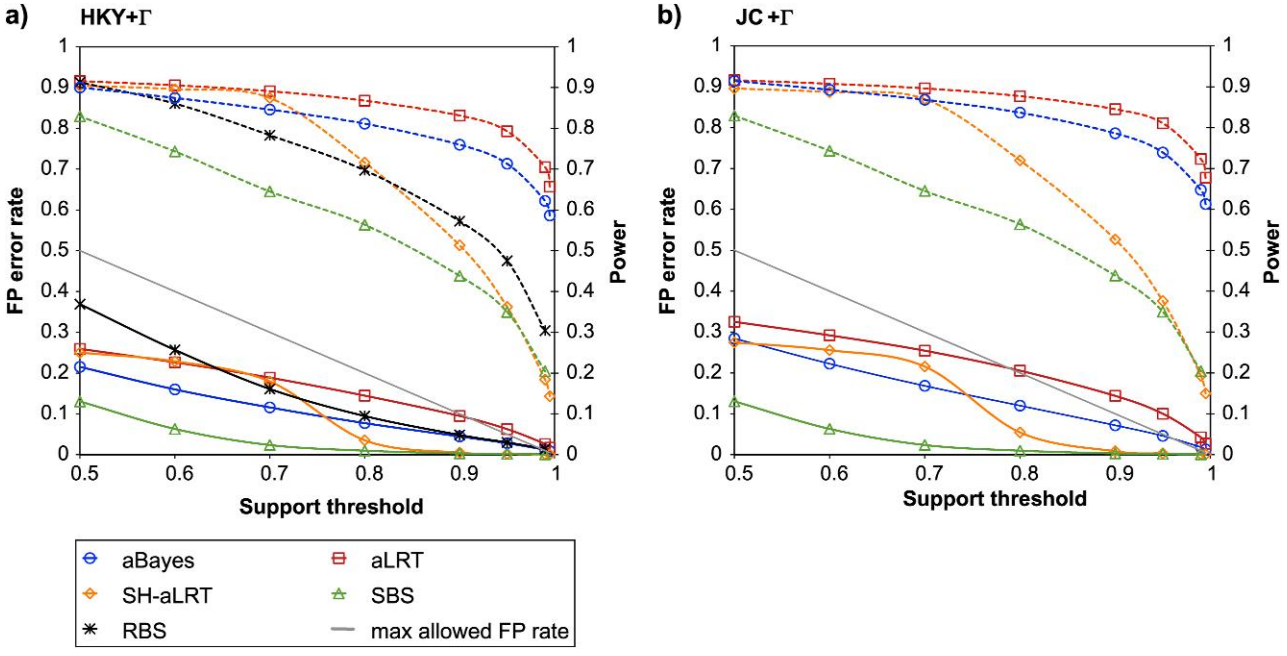


FIGURE 1. FP error rate (continuous lines) and power (dotted lines) of branch support methods. Data are simulated with 100 taxa, 600 nucleotides under the covarion model and analyzed using incorrect models: (a) HKY + Γ and (b) JC + Γ.

terms of accuracy-power balance (Fig. 1a), it tends to behave differently to SBS. On average, RBS produces visibly higher branch supports than SBS in simulations (and on real data; see below). For a threshold of >0.9, RBS and aBayes produce very similar rates of FPs, but aBayes achieves much higher power (Fig. 1a), and requires only a fraction of computing time compared with RBS.

Based on the MCC values, aLRT and aBayes ensure the best binary prediction of correct/incorrect branches, outperforming the nonparametric approaches (SH-aLRT, RBS, and SBS) by a large margin (Suppl. Table S1). Although the branch support estimation problem is not a typical classification problem, the trend seen in MCC values highly correlates with the power of these tests. Indeed, the power of nonparametric tests is much lower than that achieved by aLRT and aBayes.

The above simulations demonstrate aBayes as a clear winner judging by its speed and the power-accuracy balance it achieves, even with serious model violations. But even for aBayes, we observed a certain tendency for the FP rate to increase with more serious model violations. Thus, if severe model violations are anticipated, the nonparametric SH correction appears more suitable for testing branch support than the generally more powerful aBayes. Indeed, the SH-aLRT is sufficiently conservative to avoid high FP rates and at the same time is very fast computationally, eliminating the need to use the SBS procedure. Indeed, with a threshold of > 0.9 both SBS and SH-aLRT are very conservative and have very similar accuracy, whereas SH-aLRT has better power for a range of thresholds <0.95.

Difficulties with Probabilistic Interpretation

The traditional approach to evaluating the accuracy of branch support is to compare the estimated supports with the estimated probability of a clade to be true. Computer simulations are used to verify whether or not such probabilistic interpretation is appropriate. The probabilistic interpretation of estimated supports is desirable as it is more intuitive, and therefore has been a popular evaluation strategy. However, in reality, such property is hard to achieve even when probabilistic interpretation is expected to hold, as it is the case for the Bayesian PP inferred under the true evolutionary model and correct prior distributions (Huelsenbeck and Rannala 2004; Yang and Rannala 2005; Kolaczkowski and Thornton 2007; Yang 2007). None of the other support estimation methods allows the probabilistic interpretation, including PP supports estimated under the incorrect model (or incorrectly specified priors). Despite this, studies where the probabilistic interpretation is attached to branch supports abound (e.g., Murphy et al. 2001). Here we show that such interpretation would be inappropriate in most cases (including PP) and suggest that the frequentist approach should be preferred. Such an approach is a decision rule (as described above), which controls the rate of FP error. Below we illustrate this by simulation.

To include computationally expensive PP supports in our evaluation, we used smaller simulated data sets of 12 taxa (Table 1, Data set B). Trees inferred using the Bayesian approach and ML were compared with the known true trees, so that each inferred clade (i.e., split) was classified to be true or false. To test the suitability of the probabilistic interpretation for branch supports, true probabilities of a split to be correct were plotted against the average inferred support values (calculated in sliding windows, as described above; Fig. 2). PP values estimated from short and long MCMC chains (see Methods section) were very similar under both the true and the incorrect model. We thus report only the results obtained from long chains here.

When the model was true, the PP indeed closely reflected the true probabilities for a split to be correct (Fig. 2a), consistent with our earlier report as well as other studies (Huelsenbeck and Rannala 2004; Anisimova and Gascuel 2006). With stronger model violation (assuming JC + Γ), the PP were often much higher than the estimated true probability (Fig. 2a). The tendency to overestimate branch supports was often observed in empirical studies. Two strategies were proposed to remedy this. Lewis et al. (2005) introduced a nonzero prior for trees with polytomies. Alternatively, using a branch length prior with higher weighting for near-zero values has been suggested (Yang et al. 2005; Yang 2007). Here we opted to make a comparison of the original Bayesian implementation (MrBayes) with the modified implementation using the two-exponential mixture as the branch length prior (mb2E; http://abacus.gene.ucl.ac.uk/software.html). With the mixed prior, we reanalyzed the 12-taxa replicates and, indeed, observed on average lower posterior supports for clades compared with the original MrBayes implementation. However, these were far from the true probabilities even when the analysis model was true (Fig. 2a). No other branch supports were close to true probability values. The trends we observed are however informative because they are consistent with the evaluation under the frequentist framework.

We also used large data sets of 100 taxa (as above) to depict different support values in the probabilistic framework (Fig. 2b). Under the true model, aBayes supports were lower than true probabilities for 12 taxa (Fig. 2a), but slightly higher than true probabilities for 100 taxa sets (Fig. 2b). With model violations, aBayes supports became higher than true probabilities for both 12- and 100-taxa data sets. In comparison with aBayes under the same model, PP supports were on average visibly higher than aBayes supports, but also closer to the true probabilities (Fig. 2a). Consistent with the frequentist evaluation, the more conservative SBS tended to be much lower than the true probability (Fig. 2b). Although high SH-aLRT supports (> 0.85) were very similar to bootstrap values, lower SH-aLRT supports (< 0.85) were higher than bootstrap values. On the other hand, aLRT supports were always higher than estimated true probabilities, even with the true model.

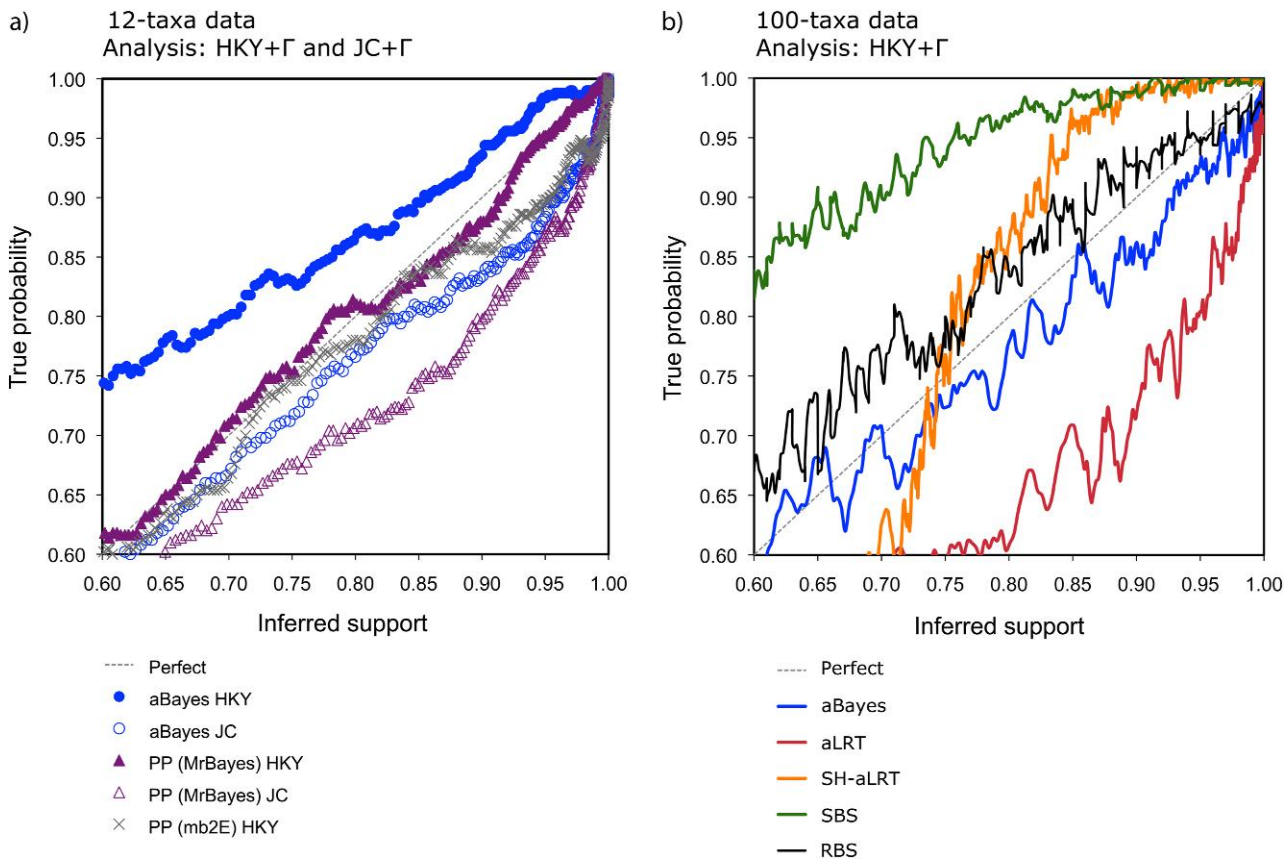


FIGURE 2. Probabilistic interpretation is rarely achieved. Inferred average support of a clade is plotted against the true probability under the true (HKY + Γ) and the incorrect (JC + Γ) models.

For 100 taxa, aBayes and RBS provided the closest estimates for the true probability, with aBayes being slightly higher and RBS slightly lower than the true probabilities on average (Fig. 2b).

Using the wrong model exaggerated these trends (Fig. 2a; data not shown for 100 taxa), making the estimated supports even further from the estimated probabilities regardless of the method. Again, the behavior of RBS could not be tested under the wrong model, as during the heuristic ML search RAxML enforces GTR + Γ and does not perform searches under simpler models.

The above clearly demonstrates that the probabilistic interpretation of support values is unachievable and is indeed misleading in most cases. Under the correct model, the Bayesian approach is superior. With model misspecifications, aLRT and aBayes outperform the Bayesian approach in both frameworks (Figs. 1 and 2). Even under the true substitution model, it would come as no surprise if the Bayesian approach is vulnerable to violations of other assumptions, such as those concerning the tree shape: the distribution of branch lengths, the branching pattern, or presence of recombination. For example, the choice of branch length prior was demonstrated to have a strong effect on the Bayesian estimation (Yang and Rannala 2005), and different exponential priors were proposed to be used in order to correct for often elevated posteriors and to resolve the so called "star-tree paradox" (Yang 2007). In our simulated 12-taxa data sets, internal and external branch lengths were sampled from the same distribution and trees were strictly bifurcating. Thus, the use of mixed priors for the analysis of these data also represents a violation of assumptions. Surprisingly, the effect of prior misspecification was almost as strong as that observed for the original MrBayes under the wrong substitution model (compare lines for MrBayes under JC + Γ and mb2E under HKY + Γ in Fig. 2a).

We further tested the robustness of branch support methods on data evolved on star trees. For example, trees reconstructed from measurably evolving populations tend to lack resolution at most nodes supporting a star-like model of evolution (Drummond et al. 2003). Viral and bacterial data often have high levels of mutation and recombination weakening tree signal. For the simulated 12-taxa star-like data sets (Table 1, Data set D), we observed only 0.3% of PP $\geq$ 0.95. This error was still acceptably low for aBayes (2%), but higher with aLRT (7.6%), slightly trespassing the traditionally acceptable level of 5%. We expect that these error rates may rise for larger trees and with model misspecifications. For example, using the correct model HKY + Γ

for the analysis of 100-taxa data sets simulated on star trees, aLRT generated 20.9% of branch supports ≥ 0.95, whereas aBayes had 9.9% of equally high supports. When simulated data were analyzed with JC + Γ (incorrect model), these frequencies slightly rose: to 23.5% and 13.2%, respectively. In contrast, SH-aLRT, RBS, and SBS almost never produced high branch supports on star trees, even when the model was violated: frequency of high supports for inferred bipartitions was always $<$ 0.01%. Star-like evolution should be easily detectable, for example, using simple LRT comparing a binary inferred tree versus a star tree (we provide one such implementation and its description on our website: http://people.inf.ethz.ch/anmaria/tree-likeness).

Evaluation of Branch Supports on Real Data

General comparison of branch supports.—The general trends observed in our simulations were confirmed on real data. We plotted branch supports obtained with different methods against SBS. Though the dispersion was considerable, the global pattern was easily seen from the arrangement of correlation lines (Suppl. Fig. S2). As in simulations, SBS produced the lowest supports, whereas PP, aLRT, and aBayes values were on average more optimistic (Suppl. Fig. S2). Most of the time SH-aLRT and RBS were higher than SBS, but visibly lower than other supports. For protein data and for 500-taxa *rbcL* data set, we observed that SH-aLRT supports were often higher than RBS (Suppl. Figs. S2b,c), whereas the reverse was true for nuclear data (Suppl. Fig. S2a). Note that this effect may be due to the fact that our nuclear data sets contained on average fewer taxa than protein data. Recall that in our simulations the performance of RBS was worse for smaller data sets. For protein and *rbcL* data sets, the behavior of RBS was the closest to SBS, as it was intended in the RBS approximation algorithm (Stamatakis et al. 2008). However, the correlations we observed were not nearly as high as those reported for RAxML test data sets (R^2 ranging from 0.85 to 0.98): for protein data sets, the linear correlation of SBS and RBS had slope $S = 0.76$ and $R^2 = 0.66$, whereas for *rbcL* data these were $S = 0.92$ and $R^2 = 0.69$.

Comparison of aBayes supports and Bayesian posteriors (PP).—Among the three most optimistic measures, aBayes can be considered as a crude approximation of PP values. Thus, we studied the correlations of these two supports for bipartitions inferred from real data sets (as listed in Table 1 except for *rbcL* data). Trees and PP values were estimated with MrBayes. Then aBayes supports were calculated on these inferred topologies with our modified version of PHYML3.0.

The strength of correlation between aBayes and PP varied from data set to data set. But in general, for nucleotide data, the correspondence of aBayes and PP values was very good and highly significant (especially in their high range), with slope $S = 0.98$, intercept $I = 0.01$ and $R^2 = 0.69$ (Fig. 3a). Out of the total common 104 splits, aBayes and PP agreed for 91.3% of them (both supports $\geq$0.95, or both supports $<$0.95), and disagreed for 9 splits (one support $\geqslant$0.95 and the other support $<$0.95). If a lower threshold of 0.9 was used, disagreements were observed only for five splits (with aBayes supporting three of them and PP the other two).

For protein data sets, the correlation was statistically weaker, and depended on the model used for the analysis (WAG + Γ, LG + Γ+F). Under WAG + Γ,

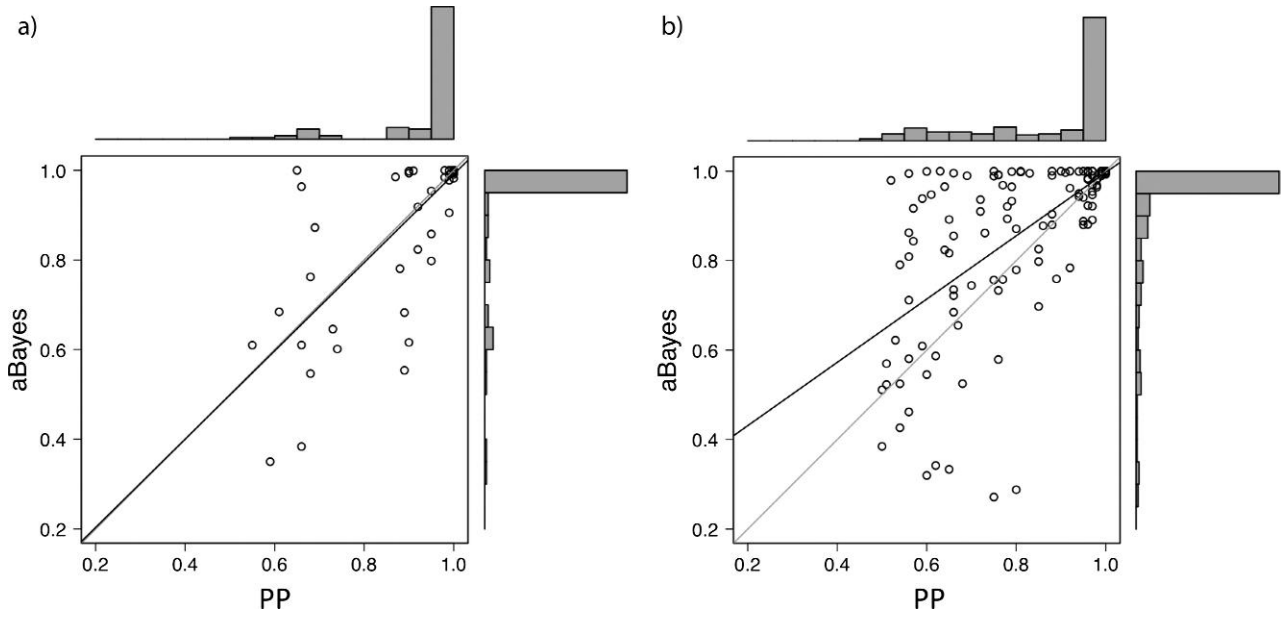


FIGURE 3. Bayesian PP compared with aBayes supports, and their distributions in real data: (a) DNA data 1–8 from Table 1, analyzed assuming HKY + Γ; (b) AA data 9–16 from Table 1, analyzed assuming WAG + Γ.

there was a light tendency of aBayes to be on average higher than corresponding PP (Fig. 3b): $S = 0.71$, $I = 0.29$, $R^2 = 0.49$. Out of the total 211 common splits, aBayes and PP agreed on 85% of inferred splits (supporting 58% of splits and both doubtful about 27% of splits). PP was in disagreement with aBayes for the remaining 15% of splits: 11% were supported only by aBayes and 4% only by PP. It is possible that aBayes supports are on average higher than PP due to serious model violations known to exist for these data (Lartillot et al. 2007): for example, under LG + Γ+ F, we observed a slight tendency for aBayes supports to be lower than PP (data not shown). However, we postulate that an inappropriately specified model may cause not only the increase of branch supports, but also a decrease. One such example was shown above for simulated data, where PP supports from the modified Bayesian approach (mb2E) decreased compared with PP values obtained with the original MrBayes implementation.

To further investigate the relationship between aBayes and PP, we reestimated trees for each of 1500 simulated 12-taxa data using the Bayesian approach, with both the correct model HKY + Γ and incorrect JC + Γ. As with real data, we then used these inferred trees to estimate the corresponding aBayes branch supports (with correct or incorrect models, respectively). When simulated data were analyzed with the true model HKY + Γ, 97% of tree partitions were inferred correctly with the Bayesian approach. The agreement between aBayes and PP supports (both ≥ 0.95, or both <0.95) was observed for 97% of all correctly inferred partitions. Moreover, 88% of inferred partitions received high supports with both methods, whereas 9% were poorly supported by both methods. Disagreements were observed only for 3% of correctly inferred partitions, with most such splits having low or insufficiently high aBayes but high PP support. The correlation between aBayes and PP was high and significant (Fig. 4a,b). In agreement with our results presented above, aBayes supports show a tendency of being slightly lower than PP values. For incorrectly inferred branches (3%), the agreement between aBayes and PP was also very good: 90% of the incorrectly inferred partitions received insufficient support with both methods (<0.95), whereas only 6.5% of incorrect branches were highly supported by both methods. The disagreement was observed in the remaining 3.4% of cases, again mostly due to lower aBayes supports.

When the simplistic JC + Γ model was used in the analysis, the agreement between aBayes and PP values was once again observed for most of all correctly inferred partitions (97.5%), with slightly lower correlation but similar to the simulation where the true model was used (Fig. 4c, d). However, for incorrectly inferred branches (4.3%), the distribution of branch supports shifted significantly, with more incorrect branches having higher support compared with the analyses under the true model (compare Fig. 4a,b with Fig. 4c,d). In more detail, 80.2% of the incorrectly inferred partitions received insufficient support with both methods, whereas 12.5% of incorrect branches were highly supported by both methods (6% more compared with the analysis under the true model). The disagreement was observed in the remaining 7.3% of cases, again mostly due to lower aBayes supports (4% more compared with the analysis under the true model).

Note that theoretically, there is no reason for aBayes supports to be biased to higher values compared with PP. Recall that aBayes support of a partition corresponding to configuration C is $\tilde{p}_c = \frac{\Pr(D|T_c)}{\Sigma_{i=1}^{3}\Pr(D|T_i)}$, where trees T_i are derived from the ML topology and are not rearranged within. In a proper Bayesian calculation, the posterior probability of a partition is calculated over all topological possibilities: $p_c = \frac{\Pr(D|T_c)+\Pr(D|\bar{T}_c)}{\Sigma_{i=1}^{3}\Pr(D|T_i)+\Pr(D|\bar{T}_{1,2,3})}$, where $\bar{T}_C$ refers to a set of topologies with branch lengths that do not conserve tree T_C but support the same partition of leaves as configuration C. Similarly, $\bar{T}_{1,2,3}$ is a set of topologies with branch lengths that are not consistent with either of the three configurations around the branch of interest. In addition, each probability term is calculated as an integral over the branch length distribution, unlike with aBayes approach where only ML estimates are used. Mathematically aBayes support may be not only higher but also lower or equal to PP value.

Examples from real data.—Above we observed that the disagreements between aBayes and PP are infrequent but exist, and it is not clear whether they are due to excessive caution of one method or to the overconfidence of the other. To explore this on real data, here we present two contrasting examples of suboptimal metazoan phylogenies from the set of proteins previously considered by Lartillot et al. (2007). Although it may be undesirable to study suboptimal trees, scientists are regularly confronted with such without knowing. Thus, observations from incorrectly reconstructed trees can be illuminating. In our example of the metazoan phylogeny (Philippe et al. 2005; Lartillot et al. 2007), the true tree is unknown, but the assumed species relationship for major animal phyla is (((((N, A), P), D), C), F)) (Fig. 5c). However, note that gene trees may often differ from the species tree (Galtier and Daubin 2008; Degnan and Rosenberg 2009). Most strikingly, for more than five taxa, the most probable reconstructed gene tree is distinct from the species tree (Degnan and Rosenberg 2006).

Here we aim to investigate the robustness of the branch supports estimated for real data, especially for small samples analyzed under misspecified models where the inferred tree is highly likely to contain many incorrect splits. First, we considered the protein alignment of the ribosomal *sap40* gene with 190 sites and 49 sequences. For this small sample but a large number of divergent lineages, we expect the reconstruction to be very unstable. Indeed, trees reconstructed by PHYML, MrBayes, and RAxML under WAG + Γ show the lowest degree of agreement compared with all other data sets in Table 1 (with only ~70% common splits). We compared branch supports for the suboptimal tree

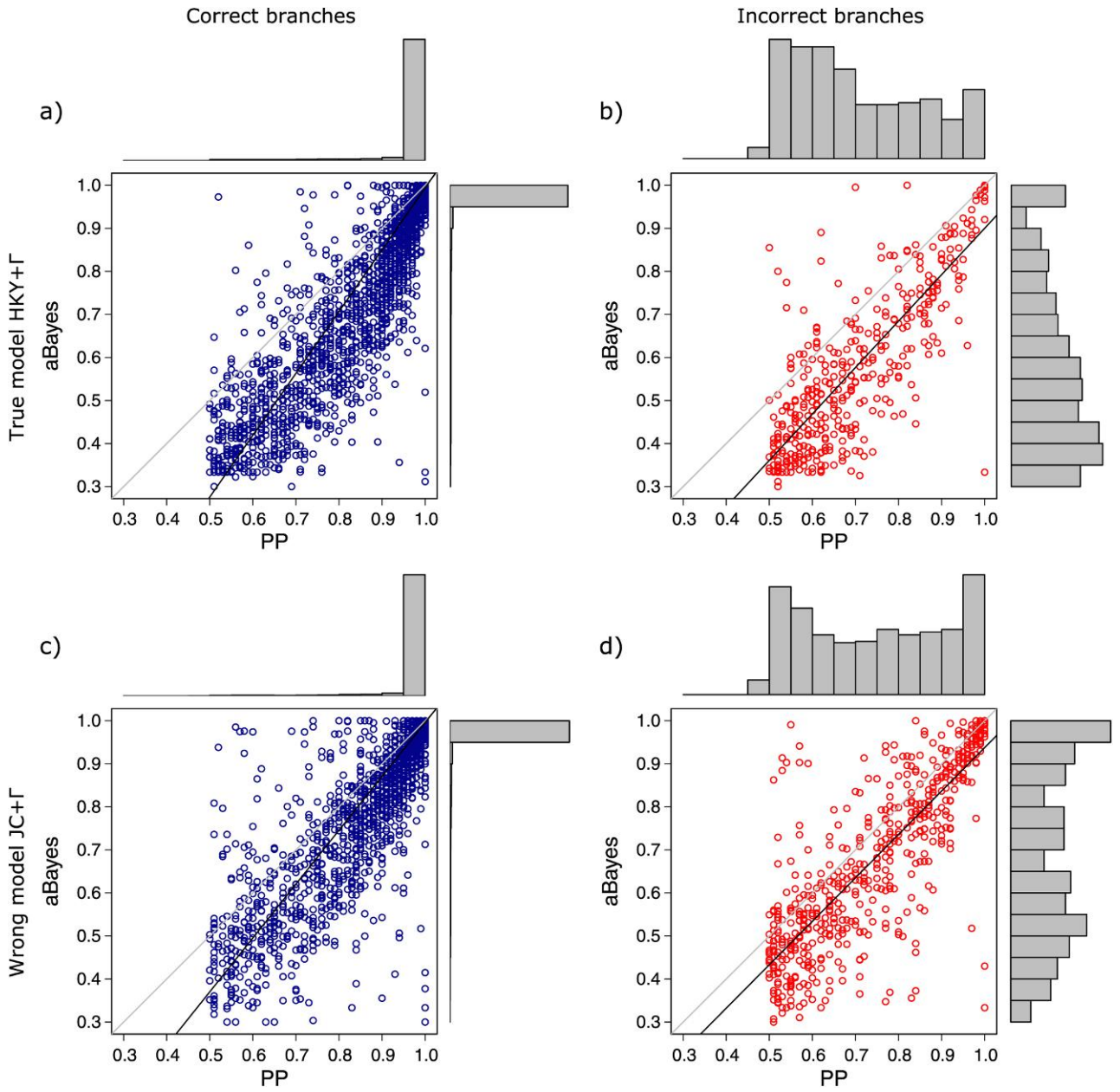
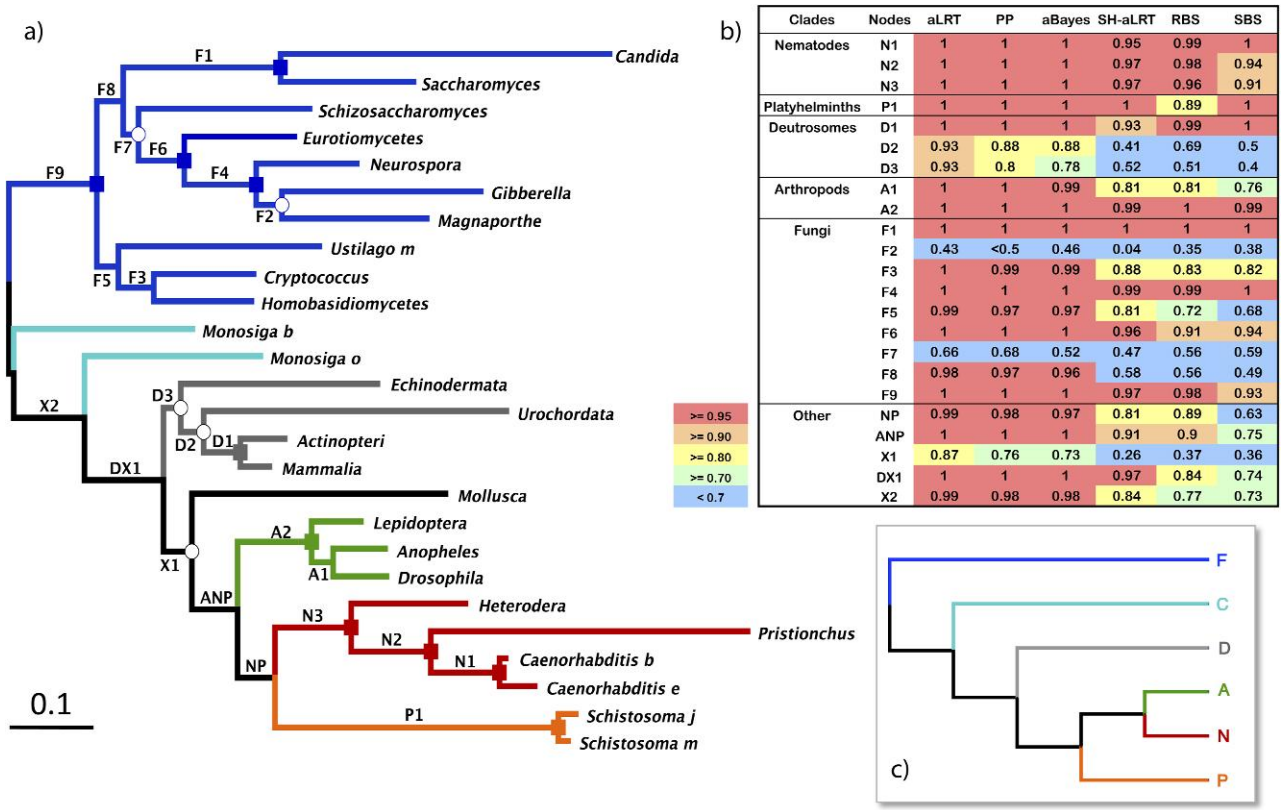


FIGURE 4. Bayesian PP compared with aBayes supports and their distributions in simulations: (a) for correctly inferred branches under HKY + Γ; (b) for incorrectly inferred branches under HKY + Γ; (c) for correctly inferred branches under JC + Γ; (d) for incorrectly inferred branches under JC + Γ. This figure is available in black and white in print and in colour at *Systematic Biology* online.

reconstructed by PHYML. Nodes unambiguously supported by all methods were always found close to the tips of the tree, where relationships were easier to reconstruct (Suppl. Fig. S3). Deeper nodes were more problematic. One obvious fault in the reconstruction was branch PX, grouping Mnemiopsis (Ctenophora) together with Platyhelminthes (Platyzoa), most likely due to the long-branch attraction artifact affecting the reconstruction of deep-rooted phylogenies (Felsenstein 2004). According to current beliefs, Ctenophora is an outgroup of Bilateria, and so Mnemiopsis should branch off split X2. This means that tree partitions X1 and APX1–APX3 are also reconstructed wrongly. Given the short alignment, the tree reconstruction for 49 very diverse species is particularly hard, especially considering existing model violations. The distribution of support values contains a large proportion of low supports—a sign that optimal reconstruction conditions were not met. In such conditions, we recommend opting for the more conservative branch support methods. Indeed, in our example SH-aLRT, RBS, and SBS do not support the incorrect branches X1, APX1–APX3, and PX. On the



Clades	Nodes	aLRT	PP	aBayes	SH-aLRT	RBS	SBS
Nematodes	N1	1	1	1	0.95	0.99	1
	N2	1	1	1	0.97	0.98	0.94
	N3	1	1	1	0.97	0.96	0.91
Platyhelminths	P1	1	1	1	1	0.89	1
Deutrosomes	D1	1	1	1	0.93	0.99	1
	D2	0.93	0.88	0.88	0.41	0.69	0.5
	D3	0.93	0.8	0.78	0.52	0.51	0.4
Arthropods	A1	1	1	0.99	0.81	0.81	0.76
	A2	1	1	1	0.99	1	0.99
Fungi	F1	1	1	1	1	1	1
	F2	0.43	<0.5	0.46	0.04	0.35	0.38
	F3	1	0.99	0.99	0.88	0.83	0.82
	F4	1	1	1	0.99	0.99	1
	F5	0.99	0.97	0.97	0.81	0.72	0.68
	F6	1	1	1	0.96	0.91	0.94
	F7	0.66	0.68	0.52	0.47	0.56	0.59
	F8	0.98	0.97	0.96	0.58	0.56	0.49
	F9	1	1	1	0.97	0.98	0.93
Other	NP	0.99	0.98	0.97	0.81	0.89	0.63
	ANP	1	1	1	0.91	0.9	0.75
	X1	0.87	0.76	0.73	0.26	0.37	0.36
	DX1	1	1	1	0.97	0.84	0.74
	X2	0.99	0.98	0.98	0.84	0.77	0.73

FIGURE 5. Comparison of branch support measures on the *nsf2-F* gene: (a) Metazoan phylogeny reconstructed for the *nsf2-F* gene with ML using PHYML; (b) estimated branch supports corresponding to reconstructed branches, and (c) the hypothesized species tree (Guindon and Gascuel 2003; Lartillot et al. 2007).

other hand, more conservative methods fail to support many correct branches (e.g., N9, D2, F4, NX, X2, X3), where other tests show high supports.

In contrast, the phylogenetic reconstruction was much more stable for the vesicular fusion protein *nsf2-F*, with most branch supports in agreement. The *nsf2-F* alignment is more than twice as long and contains roughly half the number of lineages compared with the *sap40* alignment. All programs reconstructed the same tree, which was largely matching the species tree, except for the sister relationship between nematodes and arthropods (Fig. 5a). Instead, Platyhelminths were clustered with Nematodes, which was observed in other studies and was well supported even with bootstrap when multiple genes were used. This may be again attributed to the long-branch attraction artifact in metazoan genes (Philippe et al. 2005; Lartillot et al. 2007). This relationship is supported by aBayes, aLRT, and PP (>0.95) but has insufficient SH-aLRT, RBS, and SBS supports (Fig. 5a,b).

Although the tree signal was overall much better for *nsf2-F*, bootstrap supports (RBS and SBS) and SH-aLRT lack power (low supports for monophyly of Diptera (A1), Basidiomycota (F5), Ascomycota (F8)). Higher power of SH-aLRT compared with RBS and SBS can be demonstrated by its support of the partition separating Fungi and Choanoflagellida from bilaterian animals (DX1). Note that a more relaxed threshold is typically chosen for bootstrap supports (e.g., 0.7–0.8). Yet, the accuracy of a branch support method depends on the data size and evolutionary complexity. We thus warn against arbitrary thresholds. The choice of a threshold and a support method should be guided by the properties of the data set. Using an adequate model to accommodate data specifics is highly desirable, though possibly elusive.

CONCLUSIONS AND RECOMMENDATIONS

Branch supports evaluated in our study fall into two categories: parametric (aLRT, aBayes, PP) and nonparametric (SH-aLRT, RBS, SBS). Simulations presented above warn against using aLRT with serious model violations. Other tests appear more robust to model violations: aBayes exhibits the best power, whereas nonparametric tests are much less powerful but more conservative. Although the distributions of aBayes and PP supports tend to be very similar, aBayes method is more conservative and more robust to model violations. Considering this, the *à la Bayes* interpretation of the aLRT branch support is an interesting showcase of an extreme approximation, as we demonstrate its clear advantages.

Based on our results, it is tempting to think that the distribution of estimated branch supports has some capacity to help with educated guesses about the correctness of the reconstructed phylogeny. For finite data samples, the distribution of branch supports of a tree may give an overall impression of the quality of a tree: distributions steeply peaking at higher support values may provide a sign of robust inference, whereas high percentage of low supports is an indication of a poor tree signal. Kolaczkowski and Thornton (2007) also noticed that the shape of the distribution of PP supports may depend on the pattern of branch lengths in the true tree. We suggest that the extent of model violations, the true tree and the information content of a sample also play role. Moreover, with long sequences and strong systematic bias (due to model violations), any of tested branch support methods will become very confident in supporting wrong bipartitions, as can be seen from phylogenetic analyses of concatenated genes (Jeffroy et al. 2006). Increased taxon sampling helps to avoid artifacts related to the long-branch attraction (Pick et al. 2010).

More conservative tests (SH-aLRT, RBS, SBS) may be useful if the model is known to have serious violations. SH-aLRT is extremely fast and has the highest power among nonparametric tests, while remaining as conservative as SBS. These characteristics make SH-aLRT very attractive. RBS often produces higher supports than SBS and therefore presents a different measure that should be interpreted with caution. Because RAxML does not allow evaluation under simplistic models, we did not test how this heuristic behaves with strong model violations. Our simulations show that RBS requires roughly a third of the computational time of the SBS but is less conservative and more powerful. Although the computation of RBS is now possible for hundreds and thousands of sequences, the running time for RBS is considerably longer than for aLRT-based tests. In real data, SH-aLRT and RBS produced closer estimates, whereas aBayes was closer to aLRT values, but less liberal.

The probabilistic interpretation is likely never achieved by any test (except for Bayesian posteriors under the true model and priors). Thus, none of the branch support measures should be interpreted as a probability. Instead, we recommend the frequentist interpretation, controlling the rate of FPs.

Even though our analyses cover a wide range of topologies, divergences, and models, the conclusions should not be overgeneralized: the properties of the support measures investigated here may vary with divergence, tree shape, and model fit. Before selecting a particular branch support method, we recommend testing the data set for signs of model violations. ABayes is recommended when no serious model violations are suspected, whereas SH-aLRT may be more robust if key assumptions are not met. Violations of model assumptions may be known based on biological knowledge or detected through model selection procedures (for review see Posada and Buckley 2004). Sometimes, other signs of model inadequacy can be observed, such as strong deviations of observed likelihoods (or other statistics and patterns) from the expected under the assumed model.

Real data are inherently difficult to interpret, highlighting the difficulties we face when validating methods in simulations. Equally, testing only on real data is subjected to prejudices. A recently proposed framework for more objective testing on real data is a promising step (Dessimoz and Gil 2010). Using appropriate models should improve the data fit, although this may become computationally intensive for large data sets. Intuitively, with the addition of taxa, the amount of data required for accurate model estimation should grow rapidly for a fixed average pairwise divergence (but see Erdos et al. 1999 for an interesting theoretical result on a related question). Optimizing model complexity, fit, and sample size becomes the key to more accurate phylogenetic inference.

Overall, the promising results of this study lay a solid foundation for further work on new approximate likelihood-based branch supports. Fast well-performing methods like aBayes and SH-aLRT may be especially advantageous in metagenomics studies of microbial communities, where the number of taxons in a sample is in the thousands. Moreover, the idea of relying on approximate likelihoods may provide a good starting point for evaluating branch supports in multiloci analyses, where the objective is to reconstruct a species tree, rather than a gene tree. Evaluating branches inferred by supertree and supermatrix methods is currently emerging as a new research direction in phylogenetic reconstruction (Delsuc et al. 2005; Joly and Bruneau 2009; Ropiquet et al. 2009).

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found at http://www.sysbio.oxfordjournals.org/.

Funding

M.A., M.G., and C.D. were supported by the Swiss Federal Institute of Technology (ETH Zürich). At the time of the final revision, M.A. was also receiving funding from the Swiss National Science Foundation (31003A_127325). J.-F.D. and O.G. were supported by MitoSys and PlasmoExplore ANR projects.

Acknowledgments

We would like to thank Ronald DeBry, Tiffani Williams, Bryan Kolaczkowski and one anonymous reviewer for their constructive suggestions, which helped to improve the manuscript.

References

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov B.N., Csaki F., editors. Second International Symposium on Information Theory. Budapest (Hungary): Akademiai Kiado. p. 267–281.

Aldous D. 1996. Probability distributions of cladograms. In: Aldous D., Pemantle R., editors. Random discrete structures. New York: Springer-Verlag. p. 1–18.

Anisimova M., Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. 55:539–552.

Baum D.A., Smith S.D., Donovan S.S. 2005. Evolution. The tree-thinking challenge. Science. 310:979–980.

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

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Delsuc F., Brinkmann, H. Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Desper R., Gascuel O. 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.

Dessimoz C., Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11:R37.

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

Drummond A.J., Pybus O.G., Rambaut A., Forsberg R., Rodrigo A.G. 2003. Measurably evolving populations. Trends Ecol. Evol. 18: 481–488.

Efron B. 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7:1–26.

Efron B., Halloran E., Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees [corrected and republished article originally printed in Proc. Natl Acad. Sci. U.S.A. 1996, 93:7085–7090]. Proc. Natl. Acad. Sci. U.S.A. 93:13429–13434.

Erdos P.L., Steel M.A., Szekely L.A. Warnow T.J. 1999. A few logs suffice to build (almost) all trees. (i). Random Struct. Algor. 14:153–184.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:783–791.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

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

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866–873.

Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363: 4023–4029.

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

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W. et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. Syst. Biol.

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

Harding, E. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Probab. 3:44–77.

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

Hordijk W., Gascuel O. 2005. Improving the efficiency of spr moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics. 21:4338–4347.

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

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

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

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Joly S., Bruneau A. 2009. Measuring branch support in species trees obtained by gene tree parsimony. Syst. Biol. 58:100–113.

Kishino H., Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29: 170–179.

Kolaczkowski B., Thornton J.W. 2007. Effects of branch length uncertainty on Bayesian posterior probabilities for phylogenetic hypotheses. Mol. Biol. Evol. 24:2108–2118.

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

Lartillot N., Brinkmann H., Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7(Suppl 1):S4.

Lemmon A.R., Milinkovitch M.C. 2002. The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation. Proc. Natl Acad. Sci. U.S.A. 99:10516–10521.

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

Mau B., Newton M.A., Larget B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics. 55:1–12.

Murphy W.J., Eizirik E., O'brien S.J., Madsen O., Scally M. et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science. 294:2348–2351.

Philippe H., Lartillot N., Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. Mol. Biol. Evol. 22: 1246–1253.

Phillips M.J., Lin Y.H., Harrison G.L., Penny D. 2001. Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. Proc. Biol. Sci. 268: 1533–1538.

Pick K.S., Philippe H., Schreiber F., Erpenbeck D., Jackson D.J. et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol 27:1983–1987.

Posada D., Buckley T.R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53: 793–808.

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

Ropiquet A., Li B., Hassanin A. 2009. Supertri: a new approach based on branch support analyses of multiple independent data sets for assessing reliability of phylogenetic inferences. C. R. Biol. 332:832–847.

Shimodaira H., Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Stamatakis A. 2006. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

Stamatakis A., Hoover P., Rougemont, J. 2008. A rapid bootstrap algorithm for the raxml web servers. Syst. Biol. 57:758–771.

Stamatakis A., Ludwig T., Meier H. 2005. Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 21:456–463.

Stamatakis A., Ott M. 2008. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Phil. Trans. R. Soc. Lond. B Biol. Sci.

Storey J.D. 2002. A direct approach to false discovery rates. J. R. Stat. Soc B. 64:479–498.

Strimmer K., Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. Biol. Sci. 269:137–142.

Strimmer K., Von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl Acad. Sci. U.S.A. 94:6815–6819.

Wrobel B. 2008. Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods. J. Appl. Genet. 49:49–67.

Yang Z. 2007. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol. Biol. Evol. 24:1639–1655.

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

Yang Z., Wong W.S., Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107–1118.

Yule G. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis. Phil. Trans. R. Soc. Lond. B Biol. Sci. 213:21–87.