Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability

Christophe J. Douady, Frédéric Delsuc,† Yan Boucher,* W. Ford Doolittle,* and Emmanuel J. P. Douzery†*

*Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada; and †Laboratoire de Paléontologie, Paléobiologie et Phylogénie, Institut des Sciences de l'Evolution, Université Montpellier II, Montpellier, France

Owing to the exponential growth of genome databases, phylogenetic trees are now widely used to test a variety of evolutionary hypotheses. Nevertheless, computation time burden limits the application of methods such as maximum likelihood nonparametric bootstrap to assess reliability of evolutionary trees. As an alternative, the much faster Bayesian inference of phylogeny, which expresses branch support as posterior probabilities, has been introduced. However, marked discrepancies exist between nonparametric bootstrap proportions and Bayesian posterior probabilities, leading to difficulties in the interpretation of sometimes strongly conflicting results. As an attempt to reconcile these two indices of node reliability, we apply the nonparametric bootstrap resampling procedure to the Bayesian approach. The correlation between posterior probabilities, bootstrap maximum likelihood percentages, and bootstrapped posterior probabilities was studied for eight highly diverse empirical data sets and were also investigated using experimental simulation. Our results show that the relation between posterior probabilities and bootstrapped maximum likelihood percentages is highly variable but that very strong correlations always exist when Bayesian node support is estimated on bootstrapped character matrices. Moreover, simulations corroborate empirical observations in suggesting that, being more conservative, the bootstrap approach might be less prone to strongly supporting a false phylogenetic hypothesis. Thus, apparent conflicts in topology recovered by the Bayesian approach were reduced after bootstrapping. Both posterior probabilities and bootstrap supports are of great interest to phylogeny as potential upper and lower bounds of node reliability, but they are surely not interchangeable and cannot be directly compared.

Introduction

Testing evolutionary hypotheses in a phylogenetic context becomes more reliable as reconstruction methods based on more realistic models of molecular evolution are available. However, computing time burden limits the application of model-based methods such as maximum likelihood (ML) when many taxa and/or assessment of reliability via standard—nonparametric—bootstrap methods are involved (Felsenstein 1985). Time savings thus account in part for the increasing popularity of Bayesian inference methods (e.g., Karol et al. 2001; Lutzoni, Pagel, and Reeb 2001; Murphy et al. 2001), as implemented in programs like MrBayes (Huelsenbeck and Ronquist 2001). These methods promise computational tractability with large data sets and complex evolutionary models (Larget and Simon 1999; Huelsenbeck et al. 2001).

Bayesian inference of phylogeny combines the prior probability of a phylogeny with the tree likelihood to produce a posterior probability distribution on trees (Huelsenbeck et al. 2001). The best estimate of the phylogeny can be selected as the tree with the highest posterior probability (i.e., the MAximum Posterior probability [MAP] tree) (Rannala and Yang 1996). Topologies and branch lengths are not treated as parameters—as in ML methods (Felsenstein 1981)—but as random variables. Because posterior probabilities cannot be obtained analytically, they are approximated by numerical methods known as Markov chain Monte Carlo (MCMC) or Metropolis coupled MCMC (MCMCMC). These chains are designed to explore the posterior probability surface by integration over the space of model parameters. Trees are sampled at fixed intervals and the posterior probability of a given tree is approximated by the proportion of time that the chains visited it (Yang and Rannala 1997). A consensus tree can be obtained from these sampled trees, and Bayesian posterior probabilities (PP) of individual clades, as expressed by the consensus indices, may be viewed as clade credibility values. Thus, Bayesian analysis of the initial matrix of taxa and characters produces both a MAP tree and estimates of uncertainty of its nodes, directly assessing substitution model, branch length, and topological variables, as well as clade reliability values, all in a reasonable computation time.

Reliability of nodes in phylogenetic trees is classically evaluated in two ways. First, from the initial matrix of characters, a strength of grouping value is measured, that is, the least decrease in log-likelihood associated with the breaking of the clade defined by that node (e.g., Meireles et al. 1999). The statistical significance of this decrease can be estimated with nonparametric or parametric tests (e.g., Goldman, Anderson, and Rodrigo 2000). With Bayesian methods, reliability of MAP tree nodes derives directly from corresponding posterior probabilities. In the second way, the initial character matrix is redrawn with replacement, and bootstrap percentages (BP) are calculated, for example under the ML criterion (BP_{ML}), and interpreted as a measure of experiment repeatability (Felsenstein 1985) or phylogenetic accuracy (Felsenstein and Kishino 1993).

The Bayesian approach is presumed to perform roughly as bootstrapped ML (Huelsenbeck et al. 2001) but runs much faster (Larget and Simon 1999; Huelsenbeck et al. 2001). However, Bayesian phylogenetics has its currently unsolved problems, and "perhaps the most vexing mystery is the observed discrepancy between Bayesian posterior probabilities and nonparametric

Key words: Bayesian, bootstrap, Markov chain Monte Carlo, maximum likelihood, phylogeny, posterior probability.

E-mail: cdouady@dal.ca.

Mol. Biol. Evol. 20(2):248–254. 2003
DOI: 10.1093/molbev/msg042

bootstrap support values" (Huelsenbeck et al. 2002). Recent analyses have aimed at comparing Bayesian and ML supports by studying the correlation between PP and BP_{ML} (Leaché and Reeder 2002; Whittingham et al. 2002). A compilation of literature values (Karol et al. 2001; Murphy et al. 2001; Buckley et al. 2002; Leaché and Reeder 2002; Whittingham et al. 2002; Wilcox et al. 2002) reveals that plotting PP as a function of BP_{ML} can show significant correlation ($P < 0.02$), but that the strength of this correlation is highly variable and sometime very low (correlation coefficient r^2 between 0.33 and 0.99; median at 0.73). Moreover, the slope (S) of the regression line (S between 0.29 and 1.08; median at 0.79) indicates that BP_{ML} values are generally lower than PP values. This trend has already been noticed by Rannala and Yang (1996) in their pioneering work, where PP values appeared systematically higher than resampling estimated log-likelihood (RELL) bootstrap support values.

As more phylogenetic results relying strictly on Bayesian analyses are published (Arkhipova and Morrison 2001; Henze et al. 2001; Lutzoni, Pagel and Reeb 2001), a better understanding of the relation between PP and BP_{ML} becomes essential. Wilcox et al. (2002) explored this relation by performing simulations on their original data set. They conclude that, under the condition of their study, PP and BP are both overconservative measures of phylogenetic accuracy but that Bayesian support values provide closer estimates of the true probabilities of recovering clades. Thus they advocate the preferential use of PP rather than BP (Wilcox et al. 2002). However, cases where conflicting hypotheses are supported by high posterior probabilities have been reported (Buckley et al. 2002; Douady et al. in press). This suggests that at least in certain cases, PP put overconfidence on a given phylogenetic hypothesis, and drawing conclusions from this sole measure of support might be misleading.

To better understand the relationship between PP and BP, we applied standard (i.e., nonparametric) bootstrap resampling procedures to the Bayesian approach, studying the correlation between PP, BP_{ML}, and BP_{Bay}—that is, posterior probabilities estimated after bootstrapping of the data—for eight empirical data sets spanning different kinds of characters, types of sequences, genomic compartments, and taxonomic groups. Even when the correlation between PP and BP_{ML} was weak ($r^2 < 0.52$), it became very strong ($r^2 > 0.96$) when Bayesian posterior probabilities are computed on bootstrapped data matrices. Moreover, albeit less clearly, simulation seems to confirm this trend. These simulations also tend to predict that PP overcome BP support for both true and false nodes. We discuss the effect of the bootstrapped approach in the case of apparent conflicts between data sets and consider its practical implications for measuring phylogenetic reliability.

Material and Methods

Maximum Likelihood and Standard Bayesian Analyses

Eight highly diverse empirical data sets were chosen (see details in table 1), including two pairs showing conflict (i.e., PP strongly supporting mutually exclusive nodes): mitochondrial versus nuclear markers for 14 cicadas (Buckley et al. 2002) and mitochondrial rRNA markers for 20 chondrichthyans and either one or three outgroup taxa (Douady et al. in press). The model of sequence evolution that best fits each DNA data set and the corresponding GTR substitution rate parameters, shape of the four-categories gamma distribution (Γ_4) and fraction of invariable sites (INV) were estimated by Modeltest 3.06 (Posada and Crandall 1998), and then used in PAUP* 4b10 (Swofford 2002) to compute ML bootstrap percentages (BP_{ML}) after 100 pseudoreplications with Neighbor-Joining starting trees and Tree Bisection-Reconnection branch swapping. For the amino acid data set, BP_{ML} were obtained using PROML version 3.6a2.1 of the PHYLIP package (Felsenstein 2001) with a JTT substitution matrix provided by E. Tillier (personal communication) combined to a Γ_4 + INV model, and parameters optimized by Puzzle 4.0.2 (Strimmer and von Haeseler 1996).

Bayesian posterior probabilities (PP) were computed under the same ML models with MrBayes 2.01 (Huelsenbeck and Ronquist 2001) by running four chains for 100,000 MCMCMC generations using the program default priors on model parameters. For all analyses, 1,000 trees were sampled from the posterior probability distribution (one every 100 generations) and a conservative 50% of the trees (500) was systematically discarded as "burn-in" to ensure that the chains have reached stationarity.

Bootstrapped Bayesian Analyses

We generated 100 bootstrap pseudo-replicates for each of the eight data sets using the program SEQBOOT 3.6a2.1 (Felsenstein 2001). For each pseudoreplicate, Bayesian posterior probabilities were estimated as previously described (i.e., the tree sampling and burn-in value were fixed as for the standard Bayesian approach). Bootstrapped Bayesian support was computed for each node into three ways: (1) the bootstrapped posterior probabilities (BP_{Bay}) obtained from the consensus of the $500 \times 100 = 50{,}000$ trees generated from the 100 bootstrapped pseudoreplicates, (2) the Bayesian bootstrap percentages obtained from the consensus of the 100 MAP trees (i.e., a "consensus of consensus" procedure), and (3) the average of each nodes PP for the 100 MAP trees. Given the tedious aspect of preparing files for bootstrapped Bayesian analyses, a Perl script was custom made and is available upon request.

Simulation Studies

We also explored the relation between PP and BP using a simulation design. Monte Carlo simulation of 100 data sets of 1,000 characters for seven taxa each was performed using SEQ-GEN 1.2.5 (Rambaut and Grassly 1997), under a model topology and associated branch lengths taken from the armadillo subset of VWF xenarthran data. The K2P model of nucleotide substitution (Kimura 1980) was chosen with a transition:transversion ratio of 2.00 and a Γ_8 distribution with $\alpha = 1.00$. BP_{ML} and PP supports were obtained for these 100 simulated data sets following the same procedure as described above. For computing time reasons (i.e., running

Table 1
Linear Correlation Between ML Bootstrap Percentages (BP_{ML}) and Bayesian Support[a] for Eight Highly Diverse Empirical Data Sets

	BP_{ML} (X axis)					
	PP (Y axis)			BP_{Bay} (Y axis)		
Data	r^2	S	B	r^2	S	B
Orchids, ITS[b]	0.85	0.59	44.09	0.99	1.22	−21.47
Mammals, VWF[c]	0.93	0.74	27.29	0.99	1.07	−8.13
Insects, EF1α[d]	0.75	0.36	64.33	0.99	1.07	−6.97
Insects, mitochondrial[e]	0.75	0.93	12.05	0.99	1.10	−9.95
3 Domains, HMGR[f]	0.73	0.59	43.89	0.98	0.98	1.77
Sharks, 12S to 16S (23 taxa)[g]	0.52	0.18	83.48	0.96	0.95	2.81
Sharks, 12S to 16S (21 taxa)[h]	0.49	0.38	64.70	0.99	0.98	1.19
Snakes, 12S to 16S[i]	0.27	0.25	73.37	0.95	0.93	4.85
Combination of 6 data sets[j]	0.54	0.47	55.64	0.96	1.01	−1.85
Combination of 8 data sets	0.54	0.45	57.40	0.97	1.01	−1.38

NOTE.—S and B are, respectively, the slope and the intercept of the linear correlation $Y = S \times BP_{ML} + B$.

[a] Without (PP) and with bootstrap (BP_{Bay}).

[b] Subset of nuclear ribosomal ITS (682 aligned nucleotides, nt) for 23 Diseae orchids, including 10 Satyriinae, 12 Disiinae, and one Brownleeinae species (Douzery et al. 1999); highest likelihood tree: (*Brownleea*,(((*Disa uniflora*,(*Disa racemosa*,*Disa pillansii*,(*Disa cardinalis*,*Disa tripetaloides*))),(*Disa glandulosa*,*Disa longicornis*)),((*Monadenia*,(*Disa chrysostachya*,*Herschelia*)),(*Disa rosea*,*Disa sagittalis*))),(((*Satyrium membranaceum*,(*Satyrium humile*,(*Satyrium stenopetalum*,(*Satyrium acuminatum*,(*Satyrium carneum*,*Satyrium ligulatum*))))),(*Satyrium nepalense*,*Satyrium odorum*)),(*Satyrium bicallosum*,*Satyrium rhynchanthum*))).

[c] Nuclear protein coding gene *vWF* (1,161 nt) for 13 xenarthran mammals (Delsuc et al. 2002); highest likelihood tree: (((*Dasypus novemcinctus*,***Dasypus kappleri***),((***Euphractus sexcinctus***,(***Chaetophractus villosus***,***Zaedyus pichiy***)),(***Tolypeutes matacus***,(***Priodontes maximus***,***Cabassous unicinctus***)))),(*Cyclopes didactylus*,(*Tamandua tetradactyla*,*Myrmecophaga tridactyla*)),(*Choloepus didactylus*,*Bradypus tridactylus*)). Seven armadillo taxa are in bold and were used to provide the model tree for simulations.

[d] EF1α protein coding gene (2,033 nt) of Buckley et al. (2002) for 14 cicada insects; highest likelihood tree: (*Diemeniana frenchi*,*Diemeniana tillyardi*,(((*Amphipsalta cingulata*,*Notopsalta sericea*),(*Cicadetta celis*,*Cicadetta puer*)),(*Pauropsalta johanae*,(*Myersalna depicta*,((*Maoricicada cassiope*,*Maoricicada hamiltoni*),((*Kikihia scutellaris*,*Kikihia cauta*),(*Rhodopsalta cruentata*,*Rhodopsalta leptomera*)))))).

[e] Mitochondrial (12S–16S ribosomal RNA [rRNA] + COI + COII) markers (2,249 nt) of Buckley et al. (2002) for 14 cicadas; highest likelihood tree: (*Diemeniana frenchi*,*Diemeniana tillyardi*,(((*Amphipsalta cingulata*,*Notopsalta sericea*),(*Cicadetta celis*,*Cicadetta puer*)),(*Pauropsalta johanae*,(*Myersalna depicta*,((*Kikihia scutellaris*,*Kikihia cauta*),((*Maoricicada cassiope*,*Maoricicada hamiltoni*),(*Rhodopsalta cruentata*,*Rhodopsalta leptomera*)))))).

[f] 3-Hydroxy-3-methylglutaryl coenzyme A reductase (HMGR, 258 amino acids) for 15 taxa representing all three domains of life (Eukarya-Bacteria-Archea; Boucher et al. 2001); highest likelihood tree: (((*Archaeoglobus profundus*,(*Archaeoglobus fulgidus*,(*Streptococcus pyogenes*,*Pseudomonas mevalonii*))),((*Saccharomyces cerevisiae*,*Homo sapiens*),(*Arabidopsis thaliana*,*Zea mays*))),((*Methanothermobacter thermautotrophicus*,(*Vibrio cholerae*,*Haloferax volcanii*)), (((*Pyrococcus abyssi*,*Pyrococcus horikoshii*), *Streptomyces aeriouvifer*), *Aeropyrum pernix*))).

[g] Shark mitochondrial 12S–16S rRNA for 23 taxa (1,880 nt; Douady et al. in press); highest likelihood tree: (*Petromyzon marinus*,(*Polymixia japonica*,((((((*Centrophorus granulosus*,*Squalus acanthias*),(*Squatina californica*,*Pristiophorus nudipinnis*)),((*Heterodontus francisci*,*Ginglymostoma cirratum*),(((((*Isurus oxyrinchus*,*Isurus paucus*), *Lamna nasus*),*Carcharodon carcharias*),(*Carcharias taurus*,*Alopias vulpinus*)),((*Carcharhinus porosus*,*Mustelus manazo*),*Scyliorhinus canicula*)))),(*Hexanchus griseus*,*Heptranchias perlo*)),(*Raja radiata*,*Urobatis jamaicensis*)),*Hydrolagus colliei*)),*Siren intermedia*).

[h] Shark mitochondrial 12S–16S rRNA for 21 taxa (1,963 nt, Douady et al. in press); highest likelihood tree: (*Polymixia japonica*, (((((((*Centrophorus granulosus*,*Squalus acanthias*),(*Squatina californica*,*Pristiophorus nudipinnis*)),(*Hexanchus griseus*,*Heptranchias perlo*)),((((((*Isurus oxyrinchus*,*Isurus paucus*),*Lamna nasus*),*Carcharodon carcharias*),*Carcharias taurus*),*Alopias vulpinus*),((*Carcharhinus porosus*,*Mustelus manazo*),*Scyliorhinus canicula*))),*Heterodontus francisci*),*Ginglymostoma cirratum*),(*Raja radiata*,*Urobatis jamaicensis*)),*Hydrolagus barbouri*).

[i] Snake mitochondrial 12S–16S rRNA for 23 taxa (1,545 nt; Wilcox et al. 2002); highest likelihood tree: ((*Leptotyphlops dulcis*,(*Typhlops jamaicensis*,*Typhlops ruber*)),(*Anilius scytale*,((*Trachyboa boulengeri*,(*Tropidophis greenwayi*,(*Tropidophis pardalis*,(*Tropidophis feicki*,*Tropidophis melanurus*)))),((*Xenopeltis unicolor*,(*Morelia boeleni*,*Loxocemus bicolor*)),((*Cylindrophis ruffus*,(*Uropeltis melanogaster*,*Rhinophis philippinus*)),(((*Ungaliophis continentalis*,*Exiliboa placata*),*Eryx conicus*),(*Boa constrictor*,(*Acrochordus javanicus*,(*Pituophis lineaticolis*,(*Crotalus polysticus*,*Azemiops feae*)))))))))).

[j] Six strictly independent data sets (sharks 12S–16S [21 taxa] and insects EF1α data sets excluded).

2,500 times MrBayes), BP_{Bay} were only computed for the 25 data sets showing the greatest contrast between BP_{ML} and PP.

Results and Discussion

Standard and Bootstrapped Bayesian Posterior Probabilities Versus Maximum Likelihood Bootstrap

For all eight data sets, the scatter plots of PP and BP_{Bay} versus BP_{ML} were very similar. In all cases, PP versus BP_{ML} are characterized by a moderate dispersion but a flattened slope (S column in table 1: 0.18–0.93), while BP_{Bay} versus BP_{ML} have very little unexplained variation, slopes of correlation lines appearing much steeper and being always very close to 1 (0.93–1.22). Figure 1 illustrates this trend for three individual data sets, showing that the results are independent of the nature of the data analyzed: nucleotide versus amino acid characters, nuclear versus mitochondrial compartments, protein coding versus noncoding markers and different taxonomic groups and levels (fig. 1*A*–*C*). Because empirical observations suffer from the difficulty of drawing general conclusions from a limited number of observations, we combined the six strictly independent data sets and confirmed our observations (fig. 1*D* and table 1). Therefore, we are confident that, in empirical data sets, PP and BP_{ML} will prove only moderately correlated ($r^2 = 0.27$–0.93; $S = 0.18$–0.93; $P < 0.02$), whereas the BP_{Bay} and BP_{ML} are strongly correlated ($r^2 = 0.95$–0.99; $S = 0.93$–1.22; $P < 10^{-6}$).

We tested several of the assumptions leading to the strong correlation between BP_{Bay} and BP_{ML}. First, the

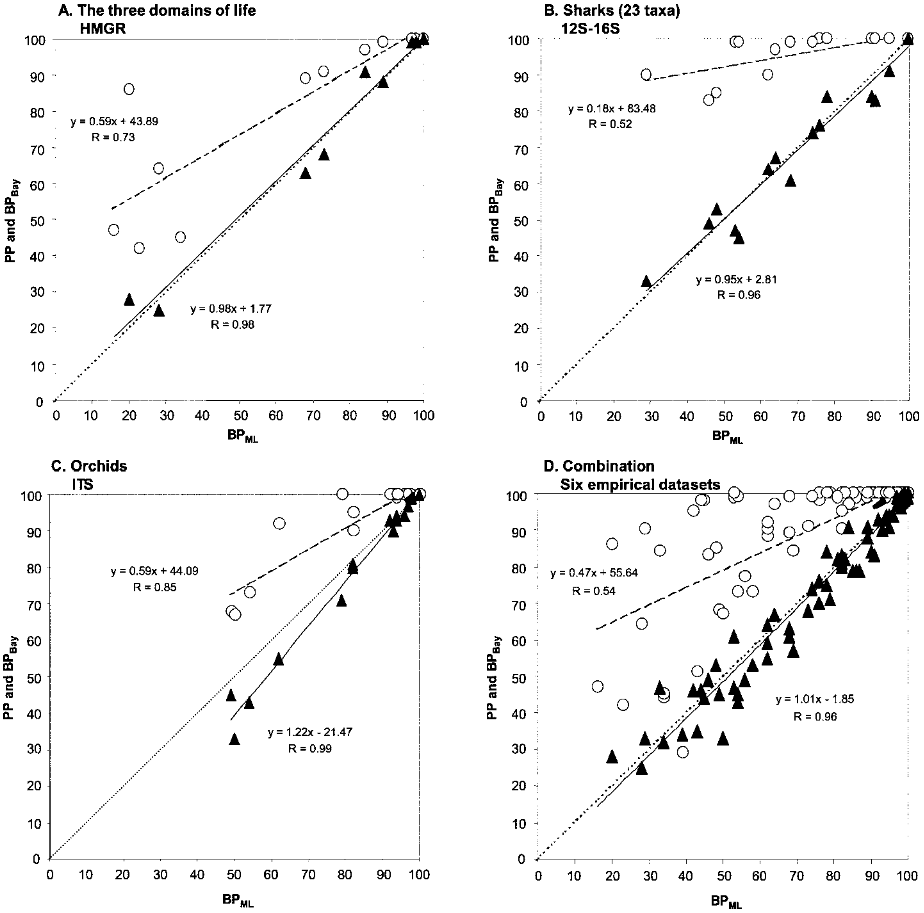

FIG. 1.—Linear correlation between maximum likelihood bootstrap percentages (BP$_{ML}$) and Bayesian posterior probabilities (PP; circles) or bootstrapped Bayesian posterior probabilities (BP$_{Bay}$; triangles) for empirical data sets. The dotted line represents a slope of 1—with equality of BP$_{ML}$ and PP or BP$_{Bay}$—while dashed and plain lines represent PP = f(BP$_{ML}$) and BP$_{Bay}$ = f(BP$_{ML}$) regression lines, respectively. All axes represent node support as percentages. See table 1 for further information regarding data sets.

possibility that the quality of the correlations observed could depend upon random error occurring between independent runs seems to be discarded by the minimal PP variance observed on MCMCMC repeatability plots (Huelsenbeck et al. 2001). It is thus unlikely that the low correlation between PP and BP$_{ML}$ reflects a problem of repeatability between independent runs. Second, we *a priori* removed 50% of the sampled trees as MCMCMC "burn-in." This was done to ensure that all trees sampled before stationarity were discarded, without actually checking Bayesian results of each individual bootstrap pseudoreplicate. To check for potential biases at this stage, we recomputed BP$_{Bay}$, keeping 90% of all sampled trees (i.e., removing 100 instead of 500 trees for each pseudoreplicate). Results indicate that bias is quite unlikely, as the level of BP$_{Bay}$ variation is very low (e.g., 1% for the ITS data set). Therefore, "burn-in" threshold seems to be of modest importance as long as it is kept realistic, probably because of the rapid convergence towards stationarity of our data. Third, we looked at the effect of making an overall consensus (i.e., consensus of all 50,000 trees sampled over all 100 pseudoreplicates and after a 50% burn-in) versus making the consensus of the 100 MAP trees or the average of the PP. Compilation of node supports—for example, for both ITS and Buckley et al. (2002) nuclear data sets—yields high correlations ($r^2 > 0.95$) between BP$_{ML}$ and BP$_{Bay}$, "MAP trees consensus," or "PP average." However, it seems that BP$_{Bay}$ and "PP average" node supports are closer to Bayesian philosophy, whereas "MAP tree consensus" values are closer to the ML bootstrap approach. Indeed, in the two first cases, the complete collection of trees is considered, while in the last case, a single optimal tree is kept to represent each pseudoreplicate. Given the likely loss of information during the consensus iteration, it seems that using an overall consensus was a better option to calculate bootstrapped Bayesian node support.

Such a correlation between BP$_{ML}$ and BP$_{Bay}$ seems expectable since the use of uniform priors in the Bayesian analyses involves that the posterior probability density is strongly dependent upon the likelihood function. However, this correlation is not trivial either, because the ML and the MAP trees obtained from each bootstrap pseudoreplicate are not always identical. For example, in the case of the 21 shark and xenarthran data sets, ML and MAP trees are different in 38% and 27% of the replicates, respectively. Therefore, the very high quality correlation between BP$_{Bay}$ and BP$_{ML}$ ($r^2 > 0.95$) cannot be expected *a priori.*

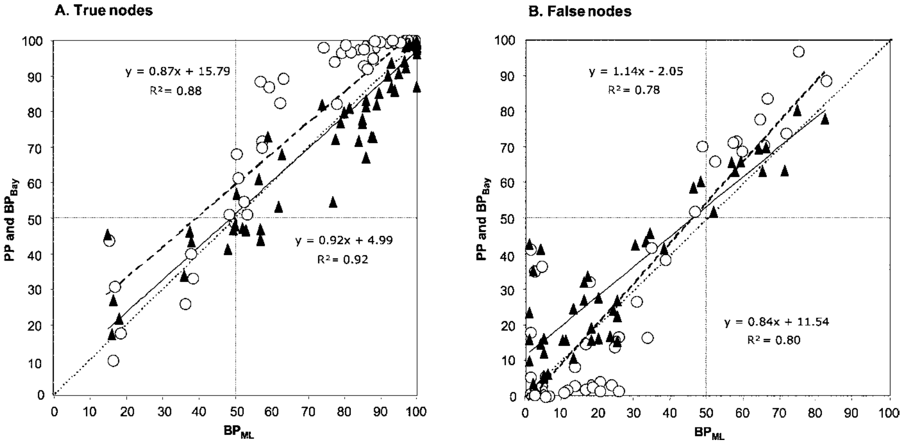

FIG. 2.—Linear correlation between maximum likelihood bootstrap percentages (BP_{ML}) and Bayesian posterior probabilities (PP; circles) or bootstrapped Bayesian posterior probabilities (BP_{Bay}; triangles) in 25 simulated data sets. "True nodes" are nodes that were present in the model topology used to simulate the data sets and "false nodes" are nodes that were not in the model topology. The dotted line represents a slope of 1—with equality of BP_{ML} and PP or BP_{Bay}—while dashed and plain lines represent PP = f(BP_{ML}) and BP_{Bay} = f(BP_{ML}) regression lines, respectively. All axes represent node support as percentages. The phylogram used for the simulation was (A:0.043143,((B:0.027559,(C:0.018247,D:0.024211):0.003601):0.011055,(E:0.005704,(F:0.010024,G:0.006528):0.000708):0.021913):0.003809) with branch lengths issued from the xenarthran data set, and taxa A–G corresponding to *Dasypus kappleri*, *Tolypeutes matacus*, *Cabassous unicinctus*, *Priodontes maximus*, *Euphractus sexcinctus*, *Chaetophractus villosus*, and *Zaedyus pichiy*, respectively.

A Simulation Study to Compare Maximum Likelihood and Bayesian Node Reliability

Nonparametric bootstrapping may be an overconservative estimator of node reliability (Hillis and Bull 1993; Wilcox et al. 2002; but see Felsenstein and Kishino 1993; Efron, Halloran, and Holmes 1996), but it remains the most commonly used way to characterize it. From the statistical point of view, posterior probabilities have the advantage to be of straightforward interpretation as they represent the probability that the corresponding clade is true, given the model, the priors, and the data (Huelsenbeck et al. 2002). However, as we showed, they are not tightly correlated with ML bootstrap percentages. Thus, these estimators seem rather different, as PP needs to be calculated on bootstrapped data to behave like BP_{ML} supports. Recently, Wilcox et al. (2002), based on a simulation study, concluded that PP and BP are both overconservative measures of node support but that PP provided closer estimates of the true probabilities of recovering clades.

Results from our simulations seem to confirm the fact that PP is less conservative than BP. Indeed, when considering true nodes—those that were present in the model topology—PP are generally higher than BP_{ML} and BP_{Bay} (fig. 2*A*, upper right quarter). However, PP is also higher when looking at strong support for false nodes—those that were absent of the model tree (fig. 2*B*, upper right quarter). Below 50% of PP and BP (fig. 2*B*, lower left quarter), that is, for values that are usually not interpreted for phylogenetic inference, there is a large dispersion of points with a trend of low BP to overestimate accuracy, as noted by Hillis and Bull (1993). As a whole these simulation results imply that, at least in certain cases, high PP falsely interpret signal and may end up strongly supporting incorrect phylogenetic relationships. Thus, the more conservative BP_{ML} and BP_{Bay} seem less subject to the behavior of strongly supporting a node when it is actually false.

Bootstrap Effect on Apparent Topological Conflicts

Bayesian analyses on bootstrapped data were able to eliminate apparent topological conflicts (fig. 3). Two nodes opposed by PP = 0.93/0.94 (*Rhodopsalta*, sister to either *Maoricicada* or *Kikihia*, depending on the choice of mitochondrial or nuclear markers [Buckley et al. 2002]) and 0.99/0.98 (relative position of Hexanchiformes in sharks' interordinal tree, depending on the choice of the outgroup [Douady et al. in press]), then, respectively, received BP_{Bay} = 59/65 and 47/57 after bootstrap resampling. Evidently, some conflicts diagnosed by PP could be biologically explained by differences between gene trees and species trees introduced by horizontal transfer, lineage sorting, and gene duplication and extinction (review in Maddison 1997). In particular, hybridization between taxa might alternatively account for the conflict observed between mitochondrial and nuclear genes in cicadas. However, in the case of sharks, the conflict arose when taxa are added to the outgroup (for the same gene). It appears more than likely that this spurious conflict was the result of the overestimation of node support based on PP and that conclusions based solely on this estimator would have been positively misleading.

The existence of strong conflicts in empirical data using standard Bayesian inference seems to argue that this approach may be sensitive to small model misspecifications as theoretically anticipated by Waddell, Kishino, and Ota (2001), subsequently shown by Buckley et al. (2002) and Buckley (2002), and acknowledged by Huelsenbeck et al. (2002). The question of the potential impact of model adequacy on the Bayesian approach may actually be

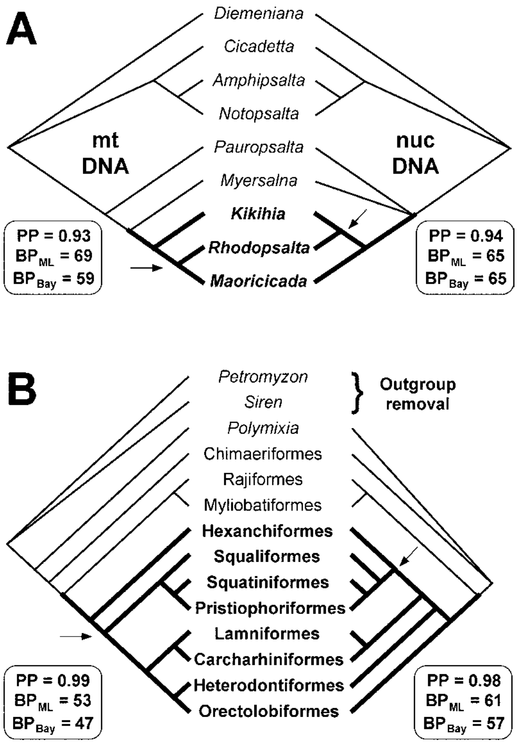

FIG. 3.—Illustration of the effect of the bootstrapped Bayesian approach in two cases of apparently conflicting nodes based on Bayesian posterior probabilities. (*A*) Conflict between maximum likelihood phylograms for mitochondrial (mtDNA) and nuclear (nucDNA) genes of Pacific *Cicada* genera (original data from Buckley et al. 2002). (*B*) Conflict about the position of Hexanchiformes within Elasmobranches between maximum likelihood phylograms of the chondrichthyan 12S and 16S rRNA data sets differing only by the number of outgroup taxa (i.e., rooting with and without *Petromyzon* and *Siren* [original data from Douady et al. in press]). PP: Bayesian posterior probability; BP$_{ML}$: maximum likelihood bootstrap percentage; BP$_{Bay}$: Bayesian bootstrapped posterior probability. The arrows indicate apparently conflicting nodes for which PP, BP$_{ML}$, and BP$_{Bay}$ are given. Thick branches connect taxa (bold names) involved in the apparent topological conflicts.

viewed as a stimulating purpose to encourage the future development of more realistic models of sequence evolution (Huelsenbeck et al. 2002). Obviously, additional studies on these issues would provide better understandings of the origin and nature of the observed differences between maximum likelihood and Bayesian node supports.

Conclusions

Drawing general conclusions from empirical studies could be problematic because we do not know how representative our example data sets are of phylogenetic problems. However, using "real" data sets does have the advantage of avoiding the simplifying assumptions inherent in simulating DNA data under an idealized model (Buckley 2002; Buckley and Cunningham 2002). Furthermore, in our case, analyses based on both empirical and simulated data seem to corroborate each other in suggesting that, being more conservative, BP$_{ML}$ and BP$_{Bay}$ might be less prone to strongly supporting a false phylogenetic hypothesis. These observations reinforce concerns regarding PP sensitivity to model misspecifications.

Nevertheless, Bayesian inference—with and without bootstrap—remains a very efficient way to simultaneously estimate substitution model parameters, branch lengths, and topology under complex models of evolutionary change (Huelsenbeck 2002). If we take the chondrichthyan 12S–16S data sets with 23 taxa as an example (fig. 3*B*), a standard Bayesian search—or one Bayesian bootstrap replicate—runs roughly 80 times faster on a 1.80 GHz Pentium 4 than a single PAUP* replicate of BP$_{ML}$ with simultaneous estimation of all parameters. Bayesian search on bootstrap data is much faster than ML if the user wants parameters to be estimated as the search goes, and it gives very similar results (fig. 1). However, in the wide majority of cases, an ML (or BP$_{ML}$) search with simultaneous estimation of the parameters is not necessary, and *a priori* approximations allow the identification of the optimal trees and bootstrap supports. The Bayesian approach also provides a unique way to analyze amino acid data with simultaneous parameters estimation, whereas this option is only available for DNA in popular phylogenetic packages such as PAUP or PHYLIP.

Both PP and bootstrap supports are of great interest to phylogeny as potential upper and lower bound of node support, but they are surely not interchangeable and cannot be directly compared. In that context, users may prefer computing PP and BP$_{Bay}$ or BP$_{ML}$ to better explore the range of node support estimates, especially when potential conflicts between data sets are explored.

Acknowledgments

As *Molecular Biology and Evolution* is no longer publishing Letters to the Editor, this paper has been reformatted into an Original Article. Chris Simon and Thomas R. Buckley kindly made their cicada data sets available. David M. Hillis and Thomas P. Wilcox provided precious enlightenment about their simulations on the snake data set. We are grateful to Alastair G. B. Simpson, the associate editor Mark A. Ragan, and two anonymous reviewers for insightful comments. This work was supported by a grant from the Canadian Institute for Health Research (MOP-42470), a Killam Postdoctoral Fellowship to C.J.D., a doctoral MENRT grant (contract 99075) to F.D., a doctoral research award from the Canadian Institutes of Health Research to Y.B., and by the TMR Network "Mammalian phylogeny" (contract FMRX-CT98-022) of the European Community, the "Génopole Montpellier Languedoc-Roussillon," and the "Action Bioinformatique inter-EPST" of the CNRS to E.J.P.D. This is the contribution ISEM 2002-048 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554-CNRS).

Note Added in Proof: Suzuki, Glazko, and Nei recently showed by simulation that posterior probabilities in Bayesian analysis can be excessively liberal, whereas bootstrap probabilities in Neighbor-Joining and maximum likelihood analyses are generally slightly conservative (2000, Proc. Natl. Acad. Sci. USA **99**:16138–16143).

Literature Cited

Arkhipova, I. R., and H. G. Morrison. 2001. Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead. Proc. Natl. Acad. Sci. USA **98**:14497–502.

Boucher, Y., H. Huber, S. L'haridon, K. O. Stetter, and W. F. Doolittle. 2001. Bacterial origin for the isoprenoid biosynthesis enzyme HMG-CoA reductase of the archaeal orders Thermoplasmatales and Archaeoglobales. Mol. Biol. Evol. **18**:1378–88.

Buckley, T. R. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst. Biol. **51**:509–523.

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.

Buckley, T. R., and C. W. Cunningham. 2002. The effect of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol. Biol. Evol. **19**:394–405.

Delsuc, F., M. Scally, O. Madsen, M. J. Stanhope, W. W. De Jong, F. M. Catzeflis, M. S. Springer, and E. J. P. Douzery. 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Mol. Biol. Evol. **19**:1656–1671.

Douady, C. J., M. Dosay, M. S. Shivji, and M. J. Stanhope. Molecular phylogenetic evidence refuting the hypothesis of Batoidea (rays and skates) as derived sharks. Mol. Phylogenet. Evol. (in press).

Douzery, E. J., A. M. Pridgeon, P. Kores, H. P. Linder, H. Kurzweil, and M. W. Chase. 1999. Molecular phylogenetics of Diseae (Orchidaceae): a contribution from nuclear ribosomal ITS sequences. Am. J. Bot. **86**:887–899.

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

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

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

———. 2001. PHYLIP (PHYLogeny Inference Package). Version 3.6a2.1. Department of Genome Sciences. University of Washington. Seattle.

Felsenstein, J., and H. Kishino. 1993. Is there something wrong with the bootstrap on phylogenies—a reply. Syst. Biol. **42**:193–200.

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49**:652–670.

Henze, K., D. S. Horner, S. Suguri, D. V. Moore, L. B. Sanchez, M. Muller, and T. M. Embley. 2001. Unique phylogenetic relationships of glucokinase and glucosephosphate isomerase of the amitochondriate eukaryotes *Giardia intestinalis*, *Spironucleus barkhanus*, and *Trichomonas vaginalis*. Gene **281**:123–131.

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. P. 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. **19**:698–707.

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.

Karol, K. G., R. M. Mccourt, M. T. Cimino, and C. F. Delwiche. 2001. The closest living relatives of land plants. Science **294**:2351–2353.

Kimura, M. 1980. A simple method for estimation evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Biol. **16**:111–120.

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.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. Nature **411**:937–940.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. **46**:523–536.

Meireles, C. M., J. Czelusniak, M. P. Schneider, J. A. Muniz, M. C. Brigido, H. S. Ferreira, and M. Goodman. 1999. Molecular phylogeny of ateline New World monkeys (Platyrrhini, Atelinae) based on gamma-globin gene sequences: evidence that *Brachyteles* is the sister group of *Lagothrix*. Mol. Phylogenet. Evol. **12**:10–30.

Murphy, W. J., E. Eizirik, S. J. O'brien et al. (11 co-authors). 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science **294**:2348–2351.

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

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.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer Associates, Sunderland, Mass.

Waddell, P. J., H. Kishino, and R. Ota. 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Informatics Series. **12**:141–155.

Whittingham, L. A., B. Slikas, D. W. Winkler, and F. H. Sheldon. 2002. Phylogeny of the tree swallow genus, *Tachycineta* (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Mol. Phylogenet. Evol. **22**:430–441.

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., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol. Biol. Evol. **14**:717–724.

Mark Ragan, Associate Editor

Accepted October 10, 2002