HAL
open science

Taxon influence index: assessing taxon-induced incongruities in phylogenetic inference

Mahendra Mariadassou, Avner Bar-Hen, Hirohisa Kishino

To cite this version:

Mahendra Mariadassou, Avner Bar-Hen, Hirohisa Kishino. Taxon influence index: assessing taxon-induced incongruities in phylogenetic inference. Systematic Biology, 2012, 61 (2), pp.337-345. 10.1093/sysbio/syr129 . hal-02642639

HAL Id: hal-02642639

https://hal.inrae.fr/hal-02642639v1

Submitted on 28 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Syst. Biol. 61(2):337–345, 2012
© The Author(s) 2012. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syr129
Advance Access publication on January 5, 2012

Taxon Influence Index: Assessing Taxon-Induced Incongruities in Phylogenetic Inference

MAHENDRA MARIADASSOU[1,2,*], AVNER BAR-HEN[1], AND HIROHISA KISHINO[2]

[1]*Department of Mathematics and Informatics, MAP5, Université Paris Descartes, 45 rue des Saints Pères, 75270 Paris Cedex 06, France; and*
[2]*Department of Agricultural and Environmental Biology, Laboratory of Biometrics, Graduate School of Agriculture and Life Sciences, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan;*
**Correspondence to be sent to: MAP5, Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris Cedex 06, France; E-mail: mahendra.mariadassou@jouy.inra.fr.*

Received 3 November 2011; accepted 5 December 2011
Associate Editor: Cécile Ané

Abstract.—Understanding the evolutionary history of species is at the core of molecular evolution and is done using several inference methods. The critical issue is to quantify the uncertainty of the inference. The posterior probabilities in Bayesian phylogenetic inference and the bootstrap values in frequentist approaches measure the variability of the estimates due to the sampling of sites from genes and the sampling of genes from genomes. However, they do not measure the uncertainty due to taxon sampling. Taxa that experienced molecular homoplasy, recent selection, a spur of evolution, and so forth may disrupt the inference and cause incongruences in the estimated phylogeny. We define a taxon influence index to assess the influence of each taxon on the phylogeny. We found that although most taxa have a weak influence on the phylogeny, a small fraction of influential taxa strongly alter it even in clades only loosely related to them. We conclude that highly influential taxa should be given special attention and sampling them more thoroughly can lead to more dependable phylogenies. [Bootstrap support; taxon sampling; taxon influence; tree robustness.]

The rapid increase in published genomic sequence for diverse organisms offers growing opportunities to infer the phylogenetic tree of groups of taxa. As with the estimate in any other inference problem, the inferred tree is subject to errors and uncertainties. Moreover, the inferred tree may not be stable with respect to small perturbations in the alignment data. Since most applications of phylogenetics require accurate and dependable phylogenetic estimates, it is crucial to determine how confident we can and should be in the inferred phylogenetic tree. Two main sources of uncertainty lie in variation among sites, studied in the bootstrap literature, and in variation among taxa, studied in the taxon sampling literature. The aim of this article is to quantify the influence of a taxon on the phylogenetic estimates.

The example of rodents highlights the importance of good taxon sampling. Philippe (1997) and Cao et al. (1997) works on rodents show that introducing a few additional taxa in a phylogenetic study can have a strong impact on the inferred tree. In the rodent phylogeny studied in these two papers, claims of D'Erchia et al. (1996) that "the guinea pig is not a rodent" based on a 16-taxa phylogeny are seriously challenged when as few as 3 additional taxa are included in the analysis. In previous work, Lecointre et al. (1993) even argue that the number and choice of taxa included in the analysis has more impact on the inferred phylogeny than the choice of an evolutionary model. The field of taxon sampling has since been the focus of much attention (Pollock et al. 2002; Zwickl and Hillis 2002; Hillis et al. 2003; Hedtke et al. 2006).

It is largely agreed upon (Cao et al. 1994; Philippe 1997; Rannala et al. 1998; Poe and Swofford 1999; Zwickl and Hillis 2002; Poe 2003; Hedtke et al. 2006) that denser taxon sampling usually leads to more accurate phylogenies, especially for large number of taxa. Other studies (Pollock et al. 2002) also suggest that if an additional taxon is available, it is usually sound to use it in the inference before pruning it from the tree. However, the effect of an additional taxon depends on the position of this taxon in the phylogeny (Goldman 1998; Geuten et al. 2007); additional taxa that break long branches are expected to improve the stability of the tree (Heath et al. 2008), whereas adding additional long branches can hinder the stability and accuracy of the inference (Kim 1998). It is also known that adding an outgroup can disrupt the ingroup topology even for small size topologies (Holland et al. 2003; Shavit et al. 2007). Furthermore, the yeast phylogeny studied by Rokas et al. (2003) and reanalyzed by Gatesy et al. (2007) shows that *removing* problematic taxa can lead to more stable and accurate phylogenies.

To our knowledge, the first attempt to assess the stability of a tree with respect to taxon sampling is due to Lanyon (1985). His method proceeds in three steps. First and starting from a distance matrix bearing n taxa, n reduced distance matrices are obtained, each bearing $n-1$ taxa, by deleting in turn each taxon from the original matrix. Then, a tree is derived from each of the n reduced matrix. Finally, a strict consensus tree is constructed by combining the n jackknife trees. The consensus tree and the tree derived from the complete distance matrix can be compared to identify the areas of topological agreement or disagreement. Lapointe et al. (1994) extended Lanyon's procedure to accommodate for branch lengths and briefly discussed the effect of multiple taxa deletion instead of single ones. Both Lanyon (1985) and Lapointe et al. (1994) procedures are restricted to trees derived from distance data. Siddall (1995) adapted Lanyon's procedure to parsimony analyses but changed its goal. Siddall sticks with the complete tree derived from the original data and uses the jackknife trees not to construct an alternative consensus tree but to compute Jackknife Monophyly Index (JMI) values for each clade of the complete tree. Like bootstrap values, JMI values are measure the stability

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012

of a clade with respect to taxon sampling. All three procedures are designed to identify weak parts of the tree, not taxa responsible for it.

In this work, we reinvestigate the use of jackknifing to assess the stability of a tree derived from maximum likelihood (ML) analyses with respect to taxon removal. Unlike Siddall (1995), we are interested not only in the unstable clades of the complete tree but also in the taxa responsible for it. We therefore introduce Taxon Influence Index (TII), which is devoted to the detection of highly influential taxa. TII quantifies the influence of a taxon on the phylogeny estimate by excluding it from the analysis and quantifying the resulting modifications on the inferred phylogeny. We also adapt JMI values to branches for ML analysis and define them as the number of taxa that can be excluded in turn without altering the branch. We use two examples (placental mammalian and reptiles) to illustrate the utility of the method.

Material and Methods

Methods

Taxon influence index.—Let us consider an alignment of homologous sequences of s taxa. Removing each taxon in turn from the alignment, we can generate s new smaller alignments. Using any inference method, we infer T^* from the complete alignment and a smaller tree T_i from the alignment lacking taxon i. We then prune taxon i from T^* to obtain T_i^* (Fig. 1). We hereafter refer to T^* as the "whole tree," T_i as the "inferred tree," and T_i^* as the "pruned tree." The TII of taxon i is the distance between trees T_i and T_i^*: $\mathrm{TII}(i) = d(T_i, T_i^*)$.

The sequences are not realigned each time a taxon is pruned so that TII does not mix the influence of a taxon on the alignment and its influence on the phylogenetic estimates. Whenever the tree is perfectly stable with regards to taxon sampling or when the influence of taxon i over the result is small, we expect the pruned tree and the inferred tree to be very similar. In the extreme case, where a taxon is duplicated in the alignment, the two copies of this taxon should be clustered at a tip of the topology and removing any of them should not modify the topology, nor the tree, in the slightest. Although TII can be used with any of the several distances on trees, we focus here on two distances: Robinson–Foulds (RF) distance (Robinson and Foulds 1979) and branch score distance (BSD) (Kuhner and Felsenstein 1994). RF accounts only for topological differences, whereas BSD weighs the topological differences by the length of the affected branches.

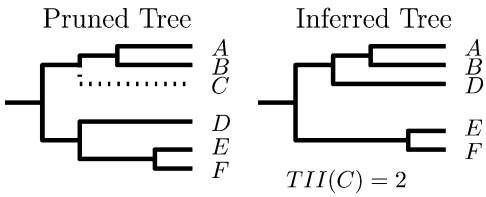

FIGURE 1. TII of taxon C. The pruned tree is obtained by pruning taxon C from the complete tree. The inferred tree is inferred directly from A, B, D–F only. The RF distance between the pruned and inferred trees is 2 so $\mathrm{TII}(C) = 2$.

Branch taxa support.—TII can detect any influential taxon but the pattern of the changes is also interesting. For example, are the branches affected when removing a taxon always the same (indicating some weakness of these particular branches) or are they well distributed across the tree (indicating for example an alignment's short length)? The study of branch stability is helpful to answer these questions; internal branches of the tree are scored for their robustness to taxon removal and more generally, changes in the taxon sample. A branch not affected by taxon sampling is robust and can reasonably be trusted, whereas a branch affected by many taxa, even those far away from it, is highly sensitive to taxon sampling and should be considered cautiously.

We define the branch taxa support $\mathrm{BTS}(b)$ of an inner branch b of T^* (the whole tree) as the number of pruned trees T_i in which it is also present or equivalently as the number of taxa that can be jackknifed without affecting that branch. Since T^* has one more leaf than T_i, it also has one more inner branch and so not all branches of T^* have a counterpart in T_i. Indeed, an inner branch connected to the terminal branch of i disappears when taxon i is pruned from the tree. Since T^* is a binary tree, branch b can be found at most in $s - 2$ (respectively $s - 1$ and s) of the T_i if it is connected to 2 terminal branches (respectively 1 and 0). Hereafter, for easier comparison with usual support values, the $\mathrm{BTS}(b)$ values are expressed as a percentage of their maximum value and range in $[0, 100]$. Others schemes such as weighing T_i with the bootstrap value of branch b in T_i may be sensible, but here, we considered only binary $0, 1$ contributions of each T_i to $\mathrm{BTS}(b)$. BTS values differ from JMI values (Siddall 1995) in the way they handle missing branches. If an inner branch cannot be found in a jackknife tree, JMI counts the branch as present and leave the maximum value unchanged for that branch, whereas BTS counts it as absent but decreases the maximum BTS value for that branch accordingly. As such, JMI values are artificially inflated for branches connected to terminal branches, although the difference is negligible for data sets with a large number of taxa. Siddall (1995) also restricts JMI to rooted trees, whereas BTS values are defined for both rooted trees and unrooted trees. The only difference between rooted and unrooted trees is that the inner branch created when rooting the tree has a perfect BTS value of 100 by construction. Finally, he advises against jackknifing an outgroup. Since outgroups can have a strong impact on the ingroup topology (Shavit et al. 2007), we think outgroup taxa should be jackknifed like other taxa.

Extension to Bayesian methods.—Although we only consider ML analysis here, TII is readily amenable to any inference method as long as it outputs a single tree. This is not the case for Bayesian methods. Even though the

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012

result of Bayesian phylogenetic inference is often summarized as a majority-rule consensus tree (MRC) (e.g., MrBayes, Ronquist and Huelsenbeck 2003) one of the strengths of Bayesian methods is the ability to account for uncertainty by providing a posterior distribution of the topology instead of a point estimate. Following Cranston and Rannala (2007) approach on agreement subtree, TII and BTS scores can be modified to use more of the posterior distribution than just the MRC tree.

As for BTS scores, the modification only consists of weighing each branch by its posterior probability $Q_i(b)$ in each of the pruned tree: $\mathrm{BTS}(b) = \sum_i Q_i(b)$. $Q_i(b)$, as a posterior probability, can take any value between 0 and 1. The ML equivalent would be to weigh each branch b by its bootstrap proportions, instead of 1 if the branch is recovered and 0 else as we did.

To compute TII, we need to measure the "distance" between posterior distributions instead of between trees. Although many such distances exist (Gibbs and Su 2002), we believe they are not well adapted to this problem; none of them includes any information about the tree structure. We instead propose the following, inspired by Cranston and Rannala (2007). We build, for each taxon i a pruned posterior distribution Q_i^* from the complete posterior distribution Q^* by pruning taxon i from the topologies of the posterior and condensing resulting identical trees. In addition, we also compute an inferred posterior distribution Q_i (from the jackknife data set lacking taxon i). The TII is then defined as the average distance between a random tree from Q_i and a random tree from Q_i,

$$\mathrm{TII}(i) = \mathbb{E}_{Q_i^* \otimes Q_i}[d(T, T')] = \sum_{T,T'} Q_i^*(T) Q_i(T') d(T, T')$$

with $d(\cdot, \cdot)$ is the same distance as before. Although with this definition the TII is not strictly a distance anymore, it is easy to compute as the average distance between the Markov chain Monte Carlo of Q^* and Q_i once they reach convergence.

Material

We examined two empirical data sets, one consisting of mitochondrial protein sequences of placental mammals and the other of mitochondrial DNA sequences of reptiles.

Sequence data.—The placental mammal data set was taken from Kitazoe et al. (2007) and consists of mitochondrial protein sequences (3658 amino acid sites in total) from 61 placental mammals, belonging to Laurasiathera, Supraprimates, Xenartha, and Afrotheria plus 7 outgroup taxa, belonging to Monotremata and Marsupialia. The gaps were excluded and the sequences were not realigned when removing a taxon. Although the original data set contains 69 taxa, two of them, labeled tenrec1 and tenrec2 are genetically so close, as shown by the phylogenies published in Kitazoe et al. (2007), that we decided to keep only tenrec1 and relabeled it tenrec (*Echinops telfairi*). Our placental mammal data set thus consists of only 68 taxa, instead of 69 in the original data. As pointed out by Kitazoe et al., these data present relatively long sequences, good taxon sampling, and very little missing data. Another advantage of mammals is that their phylogeny has been intensively studied and that many problems and hard-to-resolve clades have been identified (Prasad et al. 2008). Of particular interest is the position of the guinea pig (*Cavia porcellus*) in the order Rodentia, which has long been a heated issue among molecular phylogeneticists (Graur et al. 1991; Hasegawa and Fujiwara 1993; Cao et al. 1994, 1997; D'Erchia et al. 1996; Philippe 1997; Belfiore et al. 2008).

The reptile data set was taken from Jonniaux and Kumazawa (2008) and consists of complete mitochondrial DNA sequences (11,264 nucleotides in total) from 28 taxa, belonging for most to Squamata. The outgroup consisted of two actinopterygian fishes *Crossostoma lacustre* and *Oncorhyncus mykiss*. The gaps were excluded from the alignment and the sequences were not realigned when removing a taxon. The data set did not include any missing sites and the associated molecular phylogeny is quite well resolved. The question of interest for this data set is the branching order of 4 lizard infraorders (Gekkota, Anguimorpha, Iguania, and Scincomorpha; Evans 2003).

Phylogenetic Analysis

Although the use of TII is amenable to any phylogenetic inference method, we restricted the analysis to ML for the sake of brevity.

Evolution model.—Phylogenetic trees mammals were inferred using PhyML (Guindon and Gascuel 2003). For the placental mammals, we used the mtMam + I + Γ4 model, selected by ProtTest 1.4 (Abascal et al. 2005) as the best model no matter what the criterion (AIC, BIC). The mtMam empirical rate matrix is the one used in the best four models (mtMAM and any combination of +I and +Γ4), followed by mtREV in the next four models. The hill-climbing search was initiated at the BIONJ tree of the alignment, the default starting tree for PhyML, and 200 replicate ML bootstrap analyses were performed. For the reptiles, we used the GTR + I + Γ4 model, selected by ModelTest 3.06 (Posada and Crandall 1998) as the best for the AIC criterion. Thanks to moderate size of the data set, we used 10 random trees in addition to the default BIONJ starting tree. Again, 200 replicate ML bootstrap analyses were performed.

Analyses with PhyML were scripted using custom shell scripts. The TII and BTS scores were computed using *R* scripts (available on demand from ABH).

RESULTS

Inference Quality

We checked that the inferred tree T_i was a better ML estimate of the tree than T_i^* for the jackknife alignment. Since T_i is inferred to maximize the likelihood of the

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012

jackknife data set, whereas T_i^* maximizes the likelihood of the complete data set before being pruned, we expect the likelihood scores to be systematically higher for T_i than for T_i^*. Results from our analyses confirm it.

Placental Mammals

TII distribution and influential taxa.—TII values of the taxa are plotted in Fig. 2. We note that guinea pig has the highest TII (12), confirming previous findings of guinea pig being hard to place in the mammalian tree (Cao et al. 1994). The result is robust to model choice (with or without rate across sites (RAS) and with mtREV instead of mtMAM), with guinea pig TII always being the highest, between 12 and 14. The comparison of the pruned and inferred tree (not shown) for guinea pig reveals that removing as little as one taxon can affect the topology even in remote places; removing the guinea pig disrupts the clades of the insectivores and modifies the position of the northern tree shrew (*Tupaia belangeri*), 6 branches away from it.

Using a cutoff value of 8, which represents two standard deviations from the mean, three taxa are identified as influential (marked in bold and annotated in Fig. 3) and concentrated among Glires: guinea pig, European red squirrel (*Sciurus vulgaris*), and rabbit (*Oryctolagus cuniculus*). No matter what distance is used (RF or BSD) the same taxa stand out as influential (Fig. 2) and the TII-induced order is conserved; only 4 of the remaining 65 taxa change rank when changing the distance. But the number of influential taxa is highly dependent on the model: it varies from 4 for the mtMam + I + Γ to 10 ∼ 12 for mtMam + Γ and mtREV + Γ. Fortunately, there is an important overlap; for example, the 3 taxa influential under mtMam + I + Γ are part of the set of taxa influential under mtMam + Γ. Conversely, 20 taxa (again varying with the model from 7 in mtREV/mtMAM + Γ to 20 in mtMam + I + Γ) are extremely stable in the sense that their removal does not disturb the topology at all.

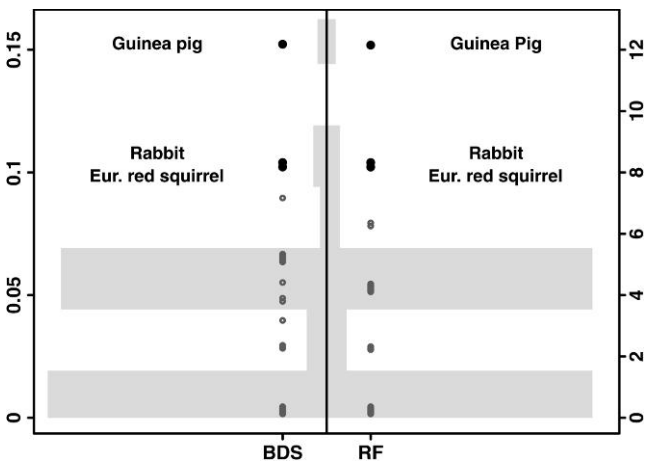

FIGURE 2. Dot-plot and histogram of TII values for BSD (left) and RF (right) distance for placental mammals. Taxa with TII higher ≥0.75 (BSD) or ≥8 (RF) are labeled with their names. Taxa with exact same location have been jittered for better legibility.

Remarkably, the stable taxa are well distributed over the tree and are either part of a clade of just two sister taxa or at the end of a long branch.

Branch taxa support.—With the exception of influential taxa and extremely stable taxa, most of the TII values are 4. This means that most inferred trees are slightly different from the corresponding pruned trees, with a difference of only two branches. We use the stability scores to check whether these differences are well distributed over the whole topology T^* or concentrated on a limited number of branches. The results are shown in Fig. 3 (inset). Interestingly, there is no correlation between BTS scores and branch lengths (0.12, $P = 0.33$) even when restricting the analysis to the branches with BTS < 100%.

Two branches with very low BTS scores belong to the Afrotheria (Fig. 3), indicating a poorly resolved clade. Indeed, even if a taxon is only weakly connected to the Afrotheria, removing it from the analysis often changes the inner configuration of the Afrotheria clade. These branches also have very low bootstrap values (11%, 54%). The same branches emerge again as sensitive to taxon sampling when removing up to the three most influential taxa, confirming that they are intrinsically hard to resolve.

A detailed comparison between BTS scores and bootstrap values is informative about their similarities and differences. First, bootstrap is more conservative than BTS: All branches with 100% bootstrap values also have 100% BTS, but some branches (20) with 100% BTS do not have 100% bootstrap values (marked in light gray in Fig. 3). Second, even though there is a significant correlation between BTS and bootstrap values (0.55, $P = 10^{-6}$), this correlation rests on most branches having both 100% bootstrap and BTS values. For the 9 branches whose both BTS and bootstrap values are lower than 100% (marked in dark gray in Fig. 3), the correlation is very low (0.25). Except for the two branches aforementioned, the bootstrap values are much smaller than their BTS equivalent: they vary between 11% and 75%, whereas all BTS scores are over 92%. This remark is consistent with (Siddall 1995), which noted that "Comparison of [BTS] and [bootstrap values] demonstrate that [BTS] is consistently and significantly greater than [bootstrap values]. Simple linear correlation [...] indicated that they are not independent".

Reptiles

TII distribution and influential taxa.—TII values of reptile taxa are plotted in Fig. 4. The most influential taxa are *Shinisaurus crocodilus* and *Coleonyx variegatus* with a TII of 8. When accounting for branch lengths, we must add *Sceloporus occidentalis* to that list. Apart from a modification at the deepermost node of the tree, all three taxa induce very localized changes, limited to the relative branching order of the scincomorphans, the anguimorphans, and the iguanians. The first remark confirms the observation previously made on mammals that the

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012

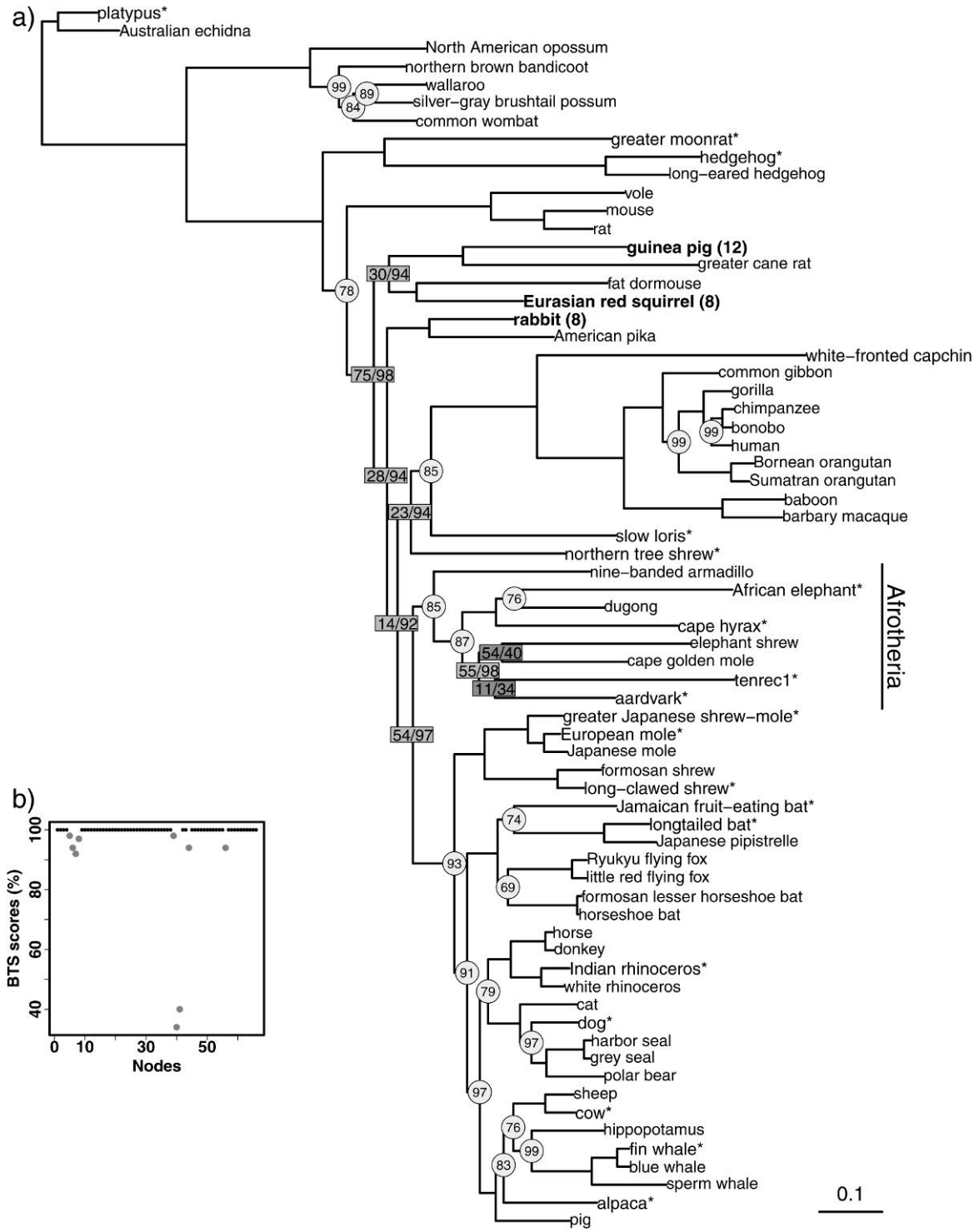

FIGURE 3. a) Placental mammals phylogeny with BTS scores and bootstrap values. 100% bootstrap and BTS values are omitted. Branches with 100% BTS scores but <100% bootstrap are annotated with their bootstrap scores in light gray circles. Branches with <100% bootstrap and BTS scores are annotated with their bootstrap (left) and BTS (right) scores in gray rectangles. The dark gray rectangles correspond to the two branches with very low BTS. Influential taxa (RF TII ≥8) are in bold and annotated by their TII. Stable taxa are annotated with ∗. Inset b) BTS scores (in %) of internal branches.

influence of a taxon need not be limited to low-level clades containing that taxon.

This is further confirmed when looking at the next most influential taxa, those with a TII of 6. Although most (6 out of 9) are located in the poorly resolved clades, three influential taxa highlight the low reliability of these clades while being located far away from them (Fig. 5)

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012

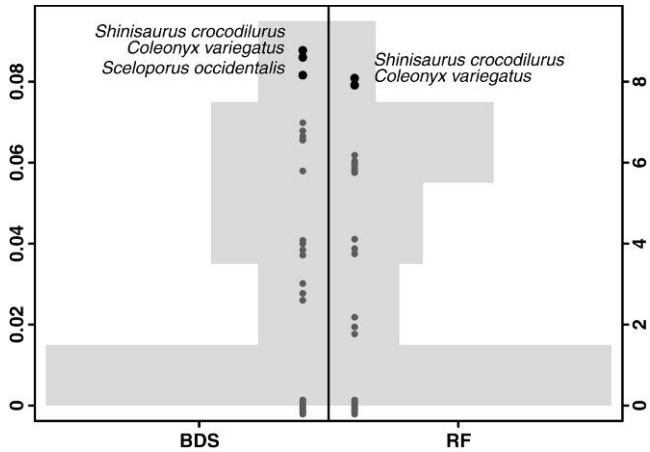

FIGURE 4. Dot-plot and histogram of TII values for BSD (left) and RF (right) distance for reptiles. Taxa with TII higher ≥0.076 (BSD) or ≥8 (RF) are labeled with their names. Taxa with exact same location have been jittered for better legibility.

Branch taxon support.—We observe the same pattern for reptiles as for mammals: no correlation between BTS and branch lengths. BTS values always higher than their bootstrap counterpart. However, correlation between BTS and bootstrap values is significant (>0.95, $P < 0.001$) no matter whether all or only weak branches are considered. We observe again that although most branches have a BTS score of 1 and are thus robust to taxon sampling, a limited number (here 3) have very low BTS values. These branches correspond to clades containing anguimorphans, scincomorphans, amphisbaenians, and iguanians, indicating a poorly resolved part of the phylogeny. These results are consistent with the uncertainty over the monophyly of Scincomorpha but at odds with the dichotomy between Iguania and Scleroglossa (Evans 2003). They are in agreement with a recent classification of Vidal and Hedges (2009), based on nuclear DNA, which rejects monophyly of both Scleroglossa (iguanians are highly nested within squamates) and Scincomorpha (Scincomorpha is redefined to include a single family and a new unranked taxon is created). They are also consistent with the phylogenies presented in Jonniaux and Kumazawa (2008, Fig. 3 and S1) and highlight the difficulty to correctly resolve the branching order of the four infraorders.

DISCUSSION

Influential Taxa and Rogue Taxa

TII is used to detect influential taxa, that is, to say taxa that strongly impact the phylogenetic estimates. TII procedure is similar to Lanyon (1985) but the focus

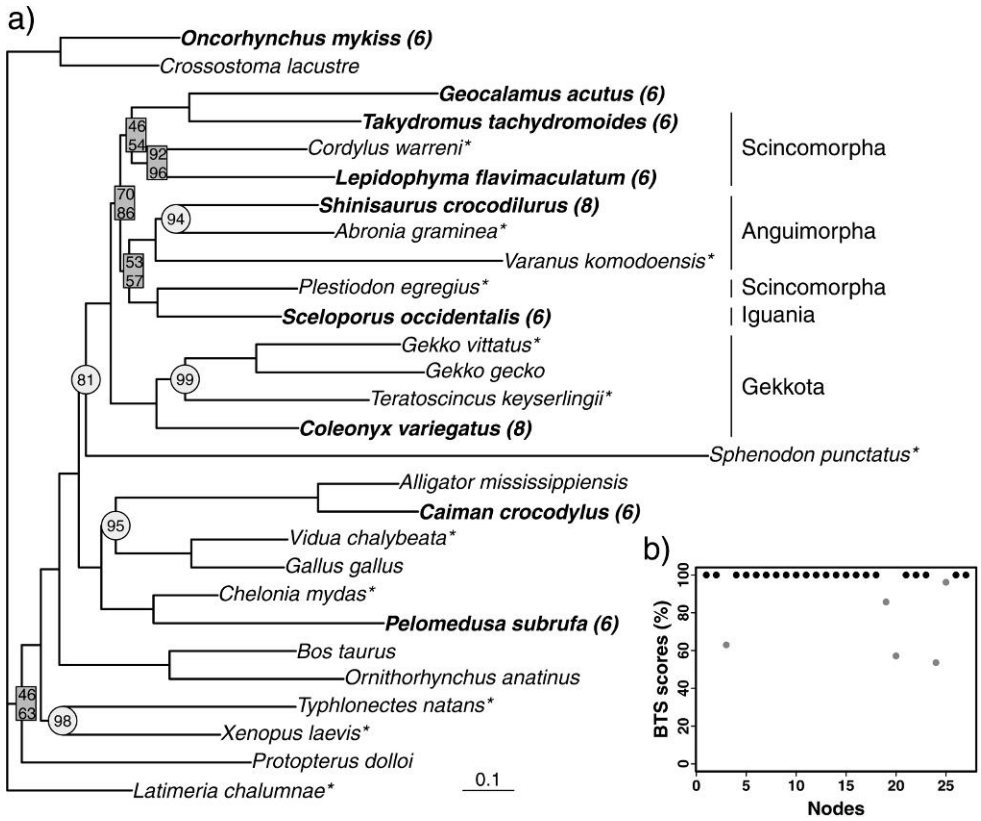

FIGURE 5. a) Reptile phylogeny with BTS scores and bootstrap values. 100% bootstrap and BTS values are omitted. Branches with 100% BTS scores but <100% bootstrap are annotated with their bootstrap scores in light gray circles. Branches with <100% bootstrap and BTS scores are annotated with their bootstrap (top) and BTS (bottom) scores in dark gray rectangles. Influential taxa (RF TII ≥ 6) are in bold and annotated by their TII. Stable taxa are annotated with ∗. Inset b) BTS scores (in %) of internal branches.

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012

is different; Lanyon focused on constructing a consensus tree whereas we are interested in detecting influential taxa and assessing the stability of the tree constructed on the complete alignment. BTS values are closer to (Siddall 1995) JMI and (Thorley and Wilkinson 1999) measure of leaf stability. However, leaf stability examines only the impact of a taxon on triplets not on the complete topology, and JMI makes unnatural choices in handling missing taxa. Rosenberg and Kumar (2001) and Pollock et al. (2002) also used the same jackknifing procedures but both authors were interested in the general impact of taxon sampling on the overall accuracy of the reconstructed tree and thus considered simulation experiments in which the true phylogeny is known, which is often not the case in practice. TII is more general as it quantifies both the influence of each taxon and the stability of each branch in relation to taxon sampling.

Finally, the issue of what to do with influential taxa remains open, as an influential taxon might not be a rogue taxon. The term "rogue" is generally restricted to taxa whose presence impedes phylogeny estimation (Wilkinson 1996; Sullivan and Swofford 1997) and the characterization of a taxon as such requires further investigations and independent lines of evidence. Indeed, a taxon that has a stabilizing beneficial effect on the phylogeny estimate is certainly influential but definitely not rogue. Siddall (1995) proposes a criterion to distinguish "critical" or stabilizing taxa from "problematic" or rogue, ones base on the number of equally parsimonious trees. Unfortunately, his criterion is inherent to parsimony analyses. We argue that influential taxa should not automatically be discarded from the analysis but rather encourage further investigations. In the reptile case study, a denser taxon sampling of reptile infraorders may improve the accuracy of the reconstructed phylogeny, but without further lines of evidence, we can only say that they are influential not beneficial nor rogue. In the mammals study case, stabilizing rodents may be more beneficial as removing the guinea pig from the analysis decreases the overall bootstrap values of the tree. Discovering why only some taxa and not others disrupt the phylogeny may help understand how and why evolution models fail us.

TII and BTS Scores

TII and BTS scores are negatively correlated: a high average TII means a low average BTS score and vice-versa. In the placental mammal phylogeny, only 20 taxa have absolutely no impact on the tree topology when pruned from the data set. This fraction is small at first sight but reflects the presence of two overall unstable clades: the first one consisting of aardvark (*Orycteropus afer*) and tenrec (*E.*) and the second one of elephant shrew (*Elephantulus* sp. *VB001*) and cape golden mole (*Chrysochloris asiatica*). These two clades account by themselves for 37 taxa with a TII of 4. The number of taxa modifying the tree elsewhere than in these two branches reduces to 11: American pika (*Ochotona collaris*), cape golden mole, dugong (*Dugong dugon*), elephant shrew, Eurasian red squirrel, fat dormouse (*Myoxus glis*), greater cane rat (*Thryonomys swinderianus*), guinea pig, mouse (*Mus musculus*), nine-banded armadillo (*Dasypus novemcinctus*), and rabbit. In the reptile phylogeny, where no unstable clade drives down BTS values to the same extent: 12 taxa (out of 28) do not change the tree at all and 7 more barely change it.

As expected, most taxa leave the tree completely or almost completely unchanged. In both case studies, half of the stable taxa belong to clades with only two taxa. This is not surprising because the two taxa of such a clade, especially if the terminal branches are very short, have very similar sequences and are almost redundant. Removing any one of them affects the inference process only marginally. For sister taxa with short terminal branch lengths, it might be worthwhile to prune the two taxa at the same time.

TII and Long Branches

When two non-adjacent taxa share many homoplasic character states along long branches, some methods (most famously parsimony) interpret such similarity as homology. The resulting tree depicts the two taxa as sister to one another, attributing the shared changes to a branch joining them; this effect is termed long-branch attraction (Felsenstein 1978). We can therefore expect taxa at the end of long terminal branches to affect the inference and have high TII.

And indeed, in the mammal phylogeny, the 11 taxa retaining positive TII after controlling for the two unstable branches are at the end of terminal branches that are significantly longer than the average terminal branch (Wilcoxon signed ranks test, increase = 81%, $P = 0.002$). However, the reverse is not true, only 8 (44%) from the 17 taxa at the end of the 25% longest terminal branch lengths are influential. This ratio never exceeds 47%, achieved for the 20% longest terminal branches. In the reptile phylogeny, influential taxa are at the end of average branches (decrease = 10%, $P > 0.4$). Influential taxa are therefore not just an artifact of long terminal branches.

Relation with Bootstrap Support

The most popular method to assess uncertainty is to compute bootstrap values. This approach has strong theoretical justification in certain circumstances (Hasegawa and Kishino 1989; Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999) but the link between bootstrap values and support for a clade is far from straightforward (Zharkikh and Li 1992; Hillis and Bull 1993). More importantly, bootstrap aggregates different sources of uncertainties and is unable to pinpoint specific sources of uncertainty, be it problematic sites (Bar-Hen et al. 2008), taxa, or branches. Finally, bootstrap is not designed to study the uncertainty induced by taxon sampling.

Most branches are highly resilient to taxon sampling and only a few poorly resolved to begin with are clearly

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012

affected by taxon sampling. Comparison of BTS scores with bootstrap values suggests that poor taxon sampling and influential species make a more localized contribution to phylogenetic variability than the broad impact of site sampling. More importantly, branches with BTS scores <100% are also among those with lowest bootstrap values. The very low bootstrap values of these branches probably arise from two correlated causes. First, the branch might just be wrong or intrinsically hard to resolve because it encompasses taxa whose positions in the tree are unclear. These taxa, by being in poorly sampled clades or by exhibiting peculiar features could equally be in several places in the tree. Therefore, only some of the equally likely topologies will contain the branch of interest. Second, there might not be a real phylogenetic signal supporting this branch, any subtle modification of the data set, be it pruning a taxon or bootstrapping sites will result in a different topology. Bootstrapping sites modifies the alignment to a greater extent than pruning a taxon and mimics the stochastic variations induced by sampling the sites. It thus captures at least two sources of variations for these branches: the first is normal sensitivity to the alignment's length, predicted by standard sampling theory. The second is excessive sensibility to the alignment induced by influential taxa: if a taxon position is unclear and essentially random within a given clade, the bootstrap topology will change depending on the resampled proportions of sites favoring one or another position of that influential taxon. This is consistent with the reptile phylogeny: low bootstrap nodes correspond to species switching from one place to another, whereas high bootstrap nodes are completely stable with respect to taxon sampling. BTS scores help isolate the two sources.

Limitations and Future Work

TII and BTS score computations require the estimation of quite a few trees; the number of trees to infer grows linearly with the number of taxa, and because of increasing complexity with a larger number of taxa, the inference time for each of them also increases. This is not specific to the proposed measures and holds for many quantities computed on trees. The total computation time increases more than linearly with the number of taxa. Computation of TII values and BTS scores is fast as it only requires comparison at the branch level. TII is thus useful for moderate data sets but not for very large ones.

By pruning only one taxon at the time, we are able to detect single taxon that exhibit peculiar evolutionary features, as corroborated by previous findings about the guinea pig (Cao et al. 1997), but we are unable to detect troublesome groups of taxa. To do so, we would need to remove two, three, or more taxa at a time. The large number of possibilities make inference of all the small trees unrealistic. The most promising paths to tackling this problem are to cluster taxa or to remove them sequentially. The first option is to "cartoon" the phylogenies by clustering taxa in groups whose inner phylogeny is well supported and choosing one representative in each group while discarding all the others to reduce the size of the topology. The second option is to remove the taxa sequentially: remove the highest TII taxon first, compute the TII again on the remaining taxa, remove the new highest TII taxon and so on until either a given number of taxa have been filtered or the highest TII does not exceed some threshold. We have however no good criterion to choose the number of taxa to filter out or the threshold.

Bootstrap and posterior probabilities are good ways to assess the uncertainty induced by site sampling but aggregate many sources of uncertainty with no way to easily ascertain which factors contribute most. They are also much more difficult to correctly interpret than thought at first (Yang 2007; Susko 2009). Furthermore, they are not designed to study the impact of taxon sampling on the inference. We show that some taxa have a large impact on the phylogenetic estimation and propose an index to identify them quantitatively with the ambition to better characterize the factors of uncertainty in phylogenetic reconstructions.

Supplementary Material

Data files associated with this study can be found at http://datadryad.org, doi:10.5061/dryad.86137tk6.

Funding

This work was supported by a grant from the Collège Doctoral Franco-Japonais (CDFJ 2007/533596E).

Acknowledgments

We are indebted to L. de Oliveira Martins, Z. Yang, P. Vandenkoornhuyse, and J. Felsenstein for useful discussions and comments on earlier versions of the manuscript. We would also like to thank J. Sullivan, C. Ane, and an anonymous reviewer for reviewing the manuscript and for their insightful comments.

References

Abascal F., Zardoya R., Posada D. 2005. Prottest: selection of best-fit models of protein evolution. Bioinformatics. 21:2104–2105.

Bar-Hen A., Mariadassou M., Poursat M.-A., Vandenkoornhuyse P. 2008. Influence function for robust phylogenetic reconstructions. Mol. Biol. Evol. 25:869–873.

Belfiore N.M., Liu L., Moritz C. 2008. Multilocus phylogenetics of a rapid radiation in the genus thomomys (rodentia: geomyidae). Syst. Biol. 57:294–310.

Cao Y., Adachi J., Yano T., Hasegawa M. 1994. Phylogenetic place of guinea pigs: no support of the rodent-polyphyly hypothesis from maximum-likelihood analyses of multiple protein sequences. Mol. Biol. Evol. 11:593–604.

Cao Y., Okada N., Hasegawa M. 1997. Phylogenetic position of guinea pigs revisited. Mol. Biol. Evol. 14:461–464.

Cranston K.A., Rannala B. 2007. Summarizing a posterior distribution of trees using agreement subtrees. Syst. Biol. 56:578–590.

D'Erchia A.M., Gissi C., Pesole G., Saccone C., Arnason U. 1996. The guinea-pig is not a rodent. Nature. 381:597–600.

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012

Evans S.E. 2003. At the feet of dinosaurs: the early history and radiations of lizards. Biol. Rev. 78:513–551.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27(4):401–410.
Gatesy J., DeSalle R., Wahlberg N. 2007. How many genes should a systematist sample? conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst. Biol. 56: 355–363.
Gibbs A.L., Su E. 2002. On choosing and bounding probability metrics. Intl. Stat. Rev. 7:419–435.
Geuten K., Massingham T., Darius P., Smets E., Goldman N. 2007. Experimental design criteria in phylogenetics: where to add taxa. Syst. Biol. 56:609–622.
Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. Biol. Sci. 265:1779–1786.
Graur D., Hide W.A., Li W.H. 1991. Is the guinea-pig a rodent? Nature. 351:649–652.
Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.
Hasegawa M., Kishino H. 1989. Confidence limits on the maximum likelihood estimate of the hominoid tree from mitochondrial DNA sequences. Evolution. 43:672–677.
Hasegawa M., Fujiwara M. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mol. Phylogenet. Evol. 2:1–5.
Heath T.A., Hedtke S.M., Hillis D.M. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239–257.
Hedtke S.M., Townsend T.M., Hillis D.M. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55:522–529.
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.
Hillis D.M., Pollock D.D., McGuire J.A., Zwickl D.J. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.
Holland B.R., Penny D., Hendy M.D. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—a simulation study. Syst. Biol. 52:229–238.
Jonniaux P., Kumazawa Y. 2008. Molecular phylogenetics and dating analyses using mitochondrial DNA sequences of eyelid geckos (Squamata: Eublepharidae). Gene. 407:105–115.
Kim J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. 47(1):43–60.
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.
Kitazoe Y., Kishino H., Waddell P.J., Nakajima N., Okabayashi T., Watabe T., Okuhara Y. 2007. Robust time estimation reconciles views of the antiquity of placental mammals. PLoS One. 2:e384.
Kuhner M.K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.
Lapointe F.J., Kirsch J.A.W., Bleiwess R. 1994. Jackknifing of weighted trees: validation of phylogenies reconstructed from distance matrices. Mol. Phylogenet. Evol. 3:256–267.
Lanyon S.M. 1985. Detecting internal inconsistencies in distance data. Syst. Zool. 34:397–403.
Lecointre G., Philippe H., L H.L.V., Guyader H.L. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phylogenet. Evol. 2:205–224.
Philippe H. 1997. Rodent monophyly: pitfalls of molecular phylogenies. J. Mol. Evol. 45:712–715.
Poe S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52: 423–428.
Poe S., Swofford D.L. 1999. Taxon sampling revisited. Nature. 398(6725):299–300.
Pollock D.D., Zwickl D.J., McGuire J.A., Hillis D.M. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.
Posada D., Crandall K.A. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.
Prasad A.B., Allard M.W., NISC Program, Green E.D. 2008. Confirming the phylogeny of mammals by use of large comparative sequence datasets. Mol. Biol. Evol. 25:1795–1808.
Rannala B., Huelsenbeck J.P., Yang Z., Nielsen R. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.
Robinson D.F., Foulds L.R. 1979. Comparison of weighted labelled trees. Lecture Notes in Mathematics, Vol. 748. Berlin: Springer. p. 119–126.
Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425(6960):798–804.
Ronquist F., Huelsenbeck J.P. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Rosenberg M. S., Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. U.S.A. 98:10751–10756.
Shavit L., Penny D., Hendy M.D., Holland B.R. 2007. The problem of rooting rapid radiations. Mol. Biol. Evol. 24:2400–2411.
Shimodaira H., Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.
Siddall M.E. 1995. Another monophily index: revisiting the jackknife. Cladistics. 11:33–56.
Sullivan J., Swofford D.L. 1997. Are guinea pigs rodents? the importance of adequate models in molecular phylogenetics. J. Mamm. Evol. 4(2):77–96.
Susko E. 2009. Bootstrap support is not first-order correct. Syst. Biol. 58:211–233.
Thorley J.L. and Wilkinson M. 1999. Testing the phylogenetic stability of early tetrapods. J. Theor. Biol. 200(3):343–344.
Vidal N., Hedges S.B. 2009. The molecular evolutionary tree of lizards, snakes and amphisbaenians. C.R. Biol. 332:129–139.
Wilkinson M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13(3):437–444.
Yang Z. 2007. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol. Biol. Evol. 24:1639–1655.
Zharkikh A., Li W.H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. i. Four taxa with a molecular clock. Mol. Biol. Evol. 9:1119–1147.
Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

Downloaded from http://sysbio.oxfordjournals.org/ by guest on February 20, 2012