Systematic Biology Advance Access published May 19, 2009

DOI:10.1093/sysbio/syp012

Paleontology, Genomics, and Combined-Data Phylogenetics: Can Molecular Data Improve Phylogeny Estimation for Fossil Taxa?

JOHN J. WIENS*

Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA;
**Correspondence to be sent to: Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA; E-mail: wiensj@life.bio.sunysb.edu.*

Abstract.—The genomics revolution offers great promise for resolving the phylogeny of living taxa, but does it offer any benefits for reconstructing relationships among extinct (fossil) taxa? Superficially, the answer would seem to be "no," given that molecular data cannot be obtained for most fossil taxa. However, because fossil taxa often interdigitate among living taxa on the Tree of Life, molecular data may indirectly enhance phylogenetic accuracy for fossil taxa in the context of a combined analysis of morphological and molecular data for living and fossil taxa. Here, I use simulations to assess accuracy for fossil taxa in a mixed analysis of living and fossil taxa, before and after addition of molecular data to the living taxa. The results show conditions where the accuracy for fossil taxa is greatly increased by adding molecular data, sometimes by as much as 100%. In other cases, the increase is negligible, such as when fossil taxa greatly outnumber living taxa in the analysis. However, there were few cases where accuracy was significantly decreased by the addition of the molecular data, suggesting that this practice may range from highly beneficial to mostly harmless. Overall, the results suggest that improvements in molecular phylogenetics can potentially benefit phylogeny reconstruction for fossil taxa. [Accuracy; fossils; genomics; morphology; phylogeny.]

The genomics revolution is currently transforming the field of phylogenetics and efforts to reconstruct the Tree of Life. Using tools from genomics, it is now possible to address phylogenetic questions with a staggering number of informative characters from multiple, unlinked loci (e.g., Rokas et al. 2003; Takezaki et al. 2004; Philippe et al. 2005; Hallstrom et al. 2007). Although some phylogenetic problems may remain persistent due to very short times between splitting events (e.g., Rokas and Carroll 2006; Wiens et al. 2008), there seems to be great potential to resolve the Tree of Life with phylogenomic approaches.

But what about fossil taxa? Is there any way that the new wealth of molecular data can improve phylogeny estimation for extinct taxa? On the surface, the answer would seem to be "no." Apart from very recent fossil taxa (from which DNA data can sometimes be obtained) or other exceptional cases (e.g., Organ et al. 2008), the phylogenetic placement of fossil taxa is based entirely on morphological data. This is unfortunate because morphological data sets typically suffer from a limited number of characters relative to molecular data sets, and the number of characters is a key factor in estimating the correct phylogeny (review in Hillis and Wiens 2000). Furthermore, there may be problems of nonindependence among characters (e.g., Emerson and Hastings 1998; O'Keefe and Wagner 2001), which can strongly mislead analyses based on morphology alone (e.g., due to developmental processes that affect entire suites of characters, such as paedomorphosis [Wiens, Bonett, et al. 2005] or peramorphosis [Smith et al. 2007]). Thus, relationships among most extinct taxa are based entirely on morphological data, and these hypotheses may sometimes be precarious due to limited numbers of characters and potential nonindependence among these characters. Yet, there is no obvious way that the new wealth of unlinked molecular characters can directly benefit phylogenetic analyses for most fossil taxa.

However, molecular data might potentially improve phylogenetic accuracy for fossil taxa in the context of combined analyses of morphological and molecular characters for living and fossil taxa. The common practice in paleontologically based phylogenetic studies is to analyze morphological data alone, even when the analysis includes extant taxa (for two recent, high-profile examples, see Wible et al. 2007; Friedman 2008; for a summary, see Cobbett et al. 2007). A combined analysis of morphological and molecular data for living taxa should generally be more accurate than an analysis of morphology alone, given the increased number and independence of characters. If fossil taxa are included, then the combined analysis might lead to higher accuracy for the fossil taxa as well. This may be especially likely if the morphological data are insufficient to fully resolve relationships of the living and fossil taxa. Also, when molecular data improve the placement of a living taxon, this may "drag" closely related fossil taxa into more accurate positions as well.

Many researchers might hesitate to add molecular data to improve phylogeny estimation for fossil taxa because the resulting data matrix would likely be dominated by missing data. For example, if the molecular data set contained 2000 characters (a relatively small number) and the morphological data set contained 100 (a relatively large number), then any fossil taxa in the combined data matrix would contain at least 95% missing data cells. Large numbers of missing data cells have traditionally been considered problematic for phylogenetic analysis (e.g., Rowe 1988; Donoghue et al. 1989; Huelsenbeck 1991; Novacek 1992; Wiens and Reeder 1995; Wilkinson 1995; Grande and Bemis 1998; Ebach and Ahyong 2001, but see Anderson 2001; Kearney 2002). However, recent simulation and empirical studies

suggest that highly incomplete taxa can be accurately placed in phylogenetic analyses, if the overall number of characters is large and the characters that are present are reasonably accurate, despite vast numbers and high proportions of missing data cells (e.g., Wiens 2003; Driskell et al. 2004; Philippe et al. 2004; Wiens, Fetzner, et al. 2005; Manos et al. 2007; Wiens and Moen 2008).

Even if one accepts that extensive missing data are not necessarily problematic, there may still only be a limited set of circumstances under which molecular data will improve phylogeny estimation for fossil taxa. For example, if most taxa in the combined data matrix are fossils, then it is hard to imagine that adding molecular data for a limited number of living taxa will dramatically increase phylogenetic accuracy for most species. Furthermore, the DNA data must estimate the correct tree or at least must not be entirely misleading. Similarly, the morphological data must not be entirely uninformative or strongly misleading either; even in the combined analysis, the accurate placement of the fossils is still ultimately dependent on the morphological data.

Several previous studies have combined fossil and molecular data in phylogenetic analyses (e.g., Eernisse and Kluge 1993; Shaffer et al. 1997; Jordan and Hill 1999; O'Leary 1999; Sun et al. 2002; Gatesy et al. 2003; Asher et al. 2005; Xiang et al. 2005; Hermsen et al. 2006; Rothwell and Nixon 2006; Magallón 2007; Manos et al. 2007; O'Leary and Gatesy 2008). Among these studies, several found that the addition of molecular data changed the position of at least some fossil taxa (e.g., the extinct crocodilian genus *Borealosuchus* is monophyletic when analyzed using morphological data alone but becomes paraphyletic when molecular data are added; Gatesy et al. 2003). Unfortunately, even if the placement of fossil taxa differs after inclusion of the molecular data, there may be little basis for determining whether phylogeny estimation has moved closer to the true phylogeny for the organisms in question (i.e., without knowing what the true phylogeny is).

In this study, I use simulations to address whether phylogeny estimation for fossil taxa might potentially be improved by adding molecular data to a combined analysis of living and fossil taxa, and under what conditions this may (or may not) occur. Computer simulations provide a context where the true phylogeny is known. Therefore, they offer a means to compare the accuracy of different approaches with phylogeny reconstruction (i.e., how well each approach estimates the true phylogeny). Admittedly, computer simulations require many simplifying assumptions and may be inappropriate to address some types of questions (e.g., do morphological data yield accurate phylogenies for mammalian fossil taxa?). However, they may be useful for addressing more general questions, such as whether adding one set of characters can improve accuracy for taxa that entirely lack data from those characters.

Previous simulation and empirical studies have suggested that adding sets of characters with data for only some taxa can sometimes improve the overall accuracy of the entire tree (e.g., Wiens 1998a; Wiens, Fetzner, et al. 2005). However, these studies did not address whether relationships among the less complete taxa were actually improved and were not designed to mimic the combination of molecular and fossil data. Many previous studies have also discussed whether adding fossil taxa improves the estimated relationships among living taxa (e.g., Gauthier et al. 1988; Donoghue et al. 1989; Huelsenbeck 1991; Eernisse and Kluge 1993; Wiens 2005; Rothwell and Nixon 2006); here, I ask instead whether adding molecular data to living taxa can improve estimated relationships among fossil taxa.

Materials and Methods

The basic design of the simulations was as follows. DNA and morphological characters were simulated on the same 16-taxon phylogeny. Certain taxa were designated to be fossils (morphology only). A matrix of morphological data was generated for all the 16 taxa. A matrix of combined morphological and molecular data for all the 16 taxa was also generated, but this combined matrix contained only missing data cells for the molecular characters for the fossil taxa. These data matrices were then analyzed using parsimony and Bayesian methods. The morphological data were analyzed alone, and accuracy was estimated for relationships among the fossil taxa (i.e., the tree was pruned to include only the fossil taxa, and the similarity of the estimated tree to the known tree for the fossil taxa alone was assessed). The combined matrix was then analyzed, and again accuracy was assessed for the fossil taxa alone. This basic procedure was then repeated hundreds of times and for different simulated conditions, such as different numbers of characters, branch lengths, and tree shapes. The main question of the study is whether accuracy for the fossil taxa is higher before or after the addition of the molecular data to the living taxa.

A 16-taxon phylogeny was simulated using programs written by the author in C. In many of the simulations, the tree was unrooted and was either entirely asymmetric (Fig. 1*a*) or symmetric (Fig. 1*b*), to test the robustness of the results to different tree shapes using the most extreme shapes possible. The same set of branch lengths was assumed for both DNA and morphological characters (assuming that lengths in both data sets are primarily influenced by the amount of time between splitting events). The first set of simulations assumed equal lengths for all branches throughout the tree, but with different lengths used in different simulations.

The morphological data sets consisted of either 20 or 100 characters, representing relatively low and high numbers for a morphological data set for 16 species. In 7 plant and vertebrate studies reviewed in Table 2, the number of morphological characters per taxon ranges from 1.750 to 8.944 (below 5.5 in 6 of the 7 studies), with a mean of 3.991, which is similar to the range and midpoint used in the simulations (1.25–6.25, midpoint = 3.75). The simulated morphological characters were all binary, given that most morphological characters in most morphological data sets appear to be binary.

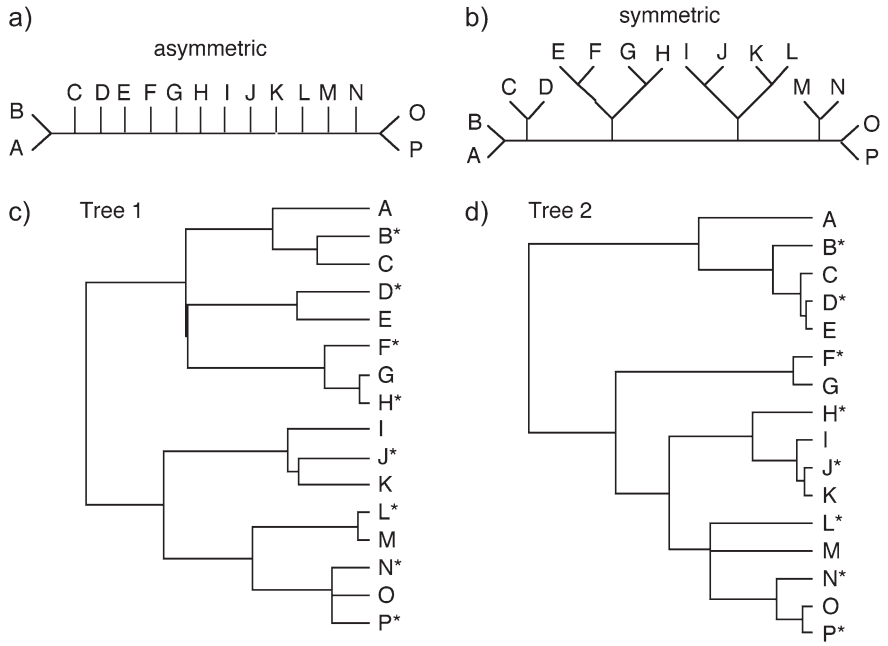

FIGURE 1. Trees used in simulations. (*a*) and (*b*) are the fully asymmetric and symmetric topologies used in the baseline simulations. The branch lengths for these two trees are arbitrary; they were either equal or varied randomly. For (*c*) and (*d*), the topology and branch lengths were randomly generated using a Yule model of speciation, and asterisks indicate the morphology-only (fossil) taxa.

The molecular data sets consisted of 2000 DNA sequence characters. Sequences were evolved assuming a 3:1 transition:transversion ratio and initial base frequencies of A = 37%, G = 12%, C = 24%, and T = 27% (parameter values based on mammalian sequences as reported by Zwickl and Hillis 2002). Although a more complex and realistic model could have been used, this added complexity would be irrelevant to this study (the accuracy of parsimony or Bayesian analysis with DNA sequence data is not the issue here). The most important property of the DNA sequence data is that it can accurately resolve relationships among the living taxa with a large number of characters. Many molecular data sets have >2000 characters, but previous studies suggest that the DNA data consistently resolve the entire tree correctly under the conditions examined here using parsimony, likelihood, and Bayesian methods (Wiens 2003; Wiens and Moen 2008). Furthermore, under conditions where the combined analysis had the lowest accuracy in this study, doubling the number of characters had little discernible impact (results not shown).

Branch lengths used were 0.01, 0.05, 0.10, and 0.20, where a branch length is defined here as the probability of a character-state change occurring along that branch. Although analysis of the DNA data is potentially accurate for all these lengths under the conditions analyzed here, these different lengths strongly affected the accuracy of the morphological data. For binary data, a length of 0.20 represents high levels of homoplasy (the consistency index is roughly 0.26, meaning that each character changes an average of 4 times across the tree). This branch length is less problematic for the DNA sequence data because, given multiple changes in the same character, DNA characters can often evolve to a different state (no homoplasy), whereas for binary characters, multiple changes must involve homoplasy. Conversely, short branch lengths are also potentially problematic for the morphological data because the combination of the limited rate of change and limited number of characters leads to a paucity of informative character changes (i.e., for 16 taxa and a length of 0.01 for each branch, only about 15% of the characters are parsimony informative).

For a given matrix, 4, 8, or 12 taxa were chosen to be fossils. In each simulation, these fossil taxa were evenly dispersed among the living taxa. Thus, when there were 4 fossil taxa, these species were A, F, K, and P (Fig. 1); for the 8 fossil taxa, these taxa were B, D, F, H, J, L, N, and O; and for 12 fossil taxa they included all taxa but A, E, L, and P. Alternately, these taxa could have either been placed randomly or clustered together. If the fossil taxa were all clustered into one clade and the living taxa in another, one would not expect the molecular data to be able to improve phylogeny estimation for them; this seems so obvious as to not be worth testing quantitatively. Alternately, if taxa were placed randomly, we would expect the results to be generally similar to those from even placement (on average). I focused exclusively on even spacing to represent the situation where the addition of molecular data is at least potentially useful for the fossil taxa.

Phylogenetic analyses were initially conducted using parsimony, as this is the method that is most widely used for reconstructing relationships among fossil taxa. However, some analyses were also conducted using Bayesian analysis, given that recent versions of MrBayes (Huelsenbeck and Ronquist 2001) allow a likelihood

model for morphological data (Lewis 2001) to be implemented and combined with analyses of DNA sequence data. Parsimony analyses were conducted using PAUP* version 4.0b10 (Swofford 2002).

The morphological data were first analyzed alone, and then, accuracy was assessed after pruning the tree to include only the fossil taxa. The combined molecular and morphological data were then analyzed and again accuracy was assessed for the pruned tree, including only the fossil taxa. Parsimony analyses consisted of a heuristic search with tree–bisection–reconnection branch swapping and 20 random-taxon-addition sequence replicates. Accuracy was assessed based on the number of nodes shared between the estimated and the true phylogenies, using a single shortest tree from each parsimony search. When averaged across replicates, a single tree from each replicate should approximate the average accuracy from comparing each shortest tree with the true phylogeny. Analyses using parsimony used 200 replicates for a given set of simulated conditions.

Bayesian analyses were conducted using MrBayes version 3.04 (Huelsenbeck and Ronquist 2001). In general, default options for Bayesian analysis were used. The DNA sequence data were analyzed using the Hasegawa–Kishino–Yano model (Hasegawa et al. 1985; accommodating unequal base frequencies and unequal transition and transversion rates), and morphological data were analyzed using the model of Lewis (2001). One hundred replicates were analyzed for each set of conditions. Each Bayesian analysis used 40 000 generations, and the first 10 000 generations were discarded as burn-in. Although this may seem like an unusually small number of generations relative to most empirical studies (which typically use several million), similar analyses using a larger number of generations show that this number is adequate (Wiens and Moen 2008). Furthermore, the results show near-perfect accuracy for many of the Bayesian analyses, indicating that there are generally few random errors (if any) generated by a failure to reach stationarity. Following standard practice, the estimated Bayesian phylogeny was based on a majority-rule consensus of the post-burn-in trees. PAUP* was used to generate these consensus trees and to compare the estimated Bayesian phylogenies with the true phylogeny.

The initial set of analyses used equal branch lengths, which is not necessarily realistic. Two additional sets of analyses were therefore conducted. First, I used randomly generated branch lengths on the fully asymmetric and symmetric unrooted topologies, with mean branch lengths of 0.01 (range 0–0.02, with lengths drawn from a uniform distribution), 0.05 (0–0.10), 0.10 (0–0.20), and 0.20 (0–0.40), that could be easily compared with the other results. A different length was selected for each branch in each replicate, but again the same length was used for both the molecular and the morphological characters.

Second, I simulated the data on two rooted topologies (Fig. 1*c*,*d*) with ultrametric branch lengths (i.e., the sum of the branch lengths from the root to the terminals is the same for all taxa). These topologies were randomly generated using a Yule (pure birth) model of speciation with Mesquite, version 1.05 (Maddison and Maddison 2004). This model generated both a topology and relative branch lengths. The length of each branch in each topology was then rescaled, so that a given set of simulations was conducted on each topology using mean branch lengths of 0.01, 0.05, 0.10, or 0.20. In this case, the different branch lengths are equivalent to different overall temporal scales for the phylogeny, from relatively recent (0.01) to more ancient (0.20). In theory, the analyses could have been conducted on hundreds of randomly simulated topologies rather than just 2, but such randomization would have made it very difficult to test the effects of combining data from living and fossil taxa (i.e., the main focus of the study), and the effects of different tree shapes and branch lengths are addressed in the other simulations.

An important property of fossil taxa is that they may retain more ancestral states than living taxa (e.g., Gauthier et al. 1988; Donoghue et al. 1989; Huelsenbeck 1991). Most simulations in this study treated the fossil taxa as equivalent to living, morphology-only taxa. However, a set of analyses were conducted on the rooted topologies in which the fossil taxa retained all the character states of their immediate ancestral node, to assess whether this impacted the results. Of course, in the real world, fossil taxa would presumably retain only some fraction of these ancestral states, but this extreme scenario was intended to offer the strongest contrast with the other simulations.

Another important property of fossil taxa is that they may be missing many characters due to preservational artifacts. In the previous analyses, I assumed that the fossil taxa were complete for all morphological characters. But in reality, fossil taxa may lack data for certain types of morphological characters that can be scored only in living taxa (e.g., soft anatomy). Furthermore, a given fossil taxon may be known from only a few incompletely preserved specimens, and so each taxon may be missing a more or less random subset of the morphological characters that could have been preserved. A limited set of simulations was conducted to assess the effects of randomly placed missing data in the fossil taxa, using the baseline simulation conditions (asymmetric and symmetric topologies with fixed branch lengths, 100 morphological characters, 8 fossil taxa). The morphological data for the fossil taxa were arbitrarily made 50% incomplete. Thus, for each taxon in each replicate, 50% of the morphological characters were randomly selected and replaced with missing data cells. Although preservation of characters in real fossil taxa presumably is not completely random, this simulation should represent a "worst-case scenario" for the distribution of missing data among characters, given that random missing data cells seem to lower accuracy more than having the same set of characters missing across all incomplete taxa (e.g., Wiens 2003).

Most simulation results are presented graphically. Standard errors of accuracy were too small to be readily visible in the figures, and so only the mean values are

shown. Similarly, even relatively small differences in mean accuracy between approaches appeared to be statistically significant, and such tests are not explicitly presented.

RESULTS

Overall, the results show that adding molecular data increases accuracy for the fossil (morphology only) taxa under a wide variety of conditions for both parsimony and Bayesian analysis. Results for different levels of completeness (see Fig. 2 for one set of results; complete results are given in Supplementary Appendix 1, http://www.sysbio.oxfordjournals.org/) show that the benefits for adding molecular data to the fossil taxa are strongest when 50% or 75% of the taxa in the matrix are living. If only 25% of the taxa have molecular data, the overall effects on accuracy for the fossil taxa are generally negligible. All further simulations were conducted for the case where 50% of the taxa are living and 50% are fossils.

The baseline results (equal branch lengths) for parsimony (Fig. 3) show substantial increases in accuracy for adding the molecular data when the number of characters is low (for intermediate to high rates and both symmetric and asymmetric trees), when there are many characters but branch lengths are long, and more generally for the symmetric tree topology (which shows lower accuracy for the morphological data than the asymmetric tree, given the same number of characters and same branch lengths). Results are very similar for Bayesian analysis under the same conditions (Fig. 4). The results are also similar when the branch lengths are allowed to vary randomly (within a set range) for both parsimony and Bayesian analysis (Supplementary Appendix 2).

For the first simulated topology under the Yule model (Fig. 5), the results are again similar, showing some benefit to adding the molecular characters under comparable conditions (e.g., few characters, long branch lengths). The results are generally similar for the second simulated topology (Fig. 6), but there is a cost in accuracy for adding the molecular data in the Bayesian analysis when branch lengths are very long, 100 morphological characters are sampled, and fossil taxa are equivalent to living taxa (in terms of retaining ancestral states). For the first topology, there is very little difference in the results when the fossil taxa retain all the character states of their immediate ancestors (Fig. 5). However, for the second simulated topology, the accuracy is higher at higher rates of change when the fossils retain their ancestral states, and adding the molecular data improves accuracy for all branch lengths (Fig. 6).

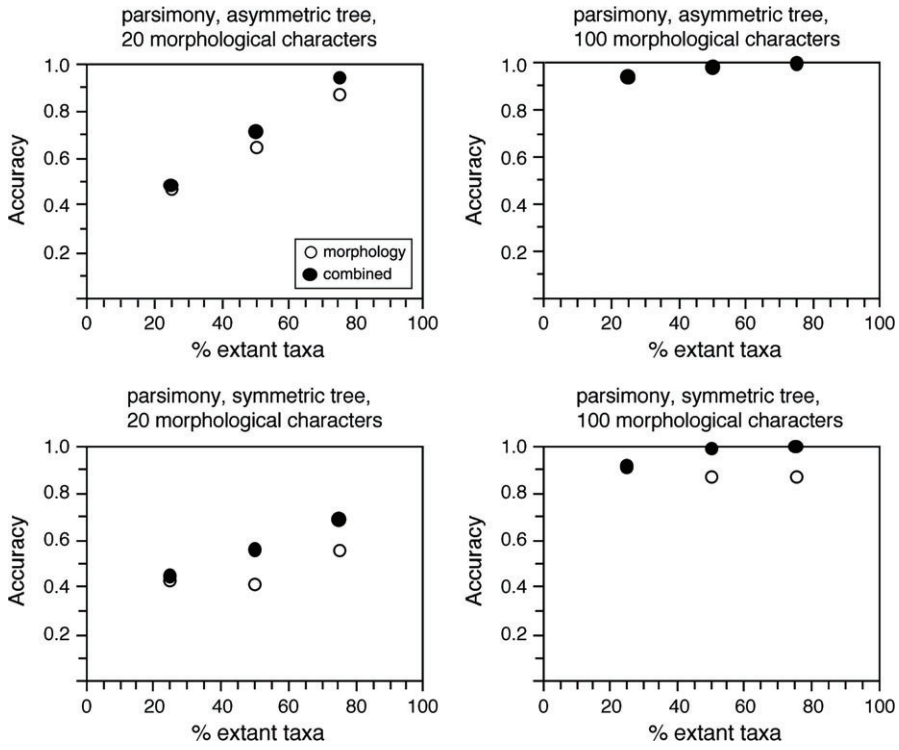

FIGURE 2. Results from simulations showing the accuracy of phylogeny estimation for fossil taxa in a combined analysis of living and extinct taxa, both with (filled circles) and without (open circles) the addition of molecular data to the living taxa. These results show the effects of varying the number of extant taxa (with molecular data) relative to the fossil taxa (morphology only), with 4 (25%), 8 (50%), or 12 (75%) of the 16 taxa extant. In these simulations, the branch length is 0.05 (intermediate low) for both the molecular and the morphological data sets. Results are based on parsimony. There are 2000 DNA sequence characters in the combined data set. Each data point represents the average accuracy from 200 replicates.

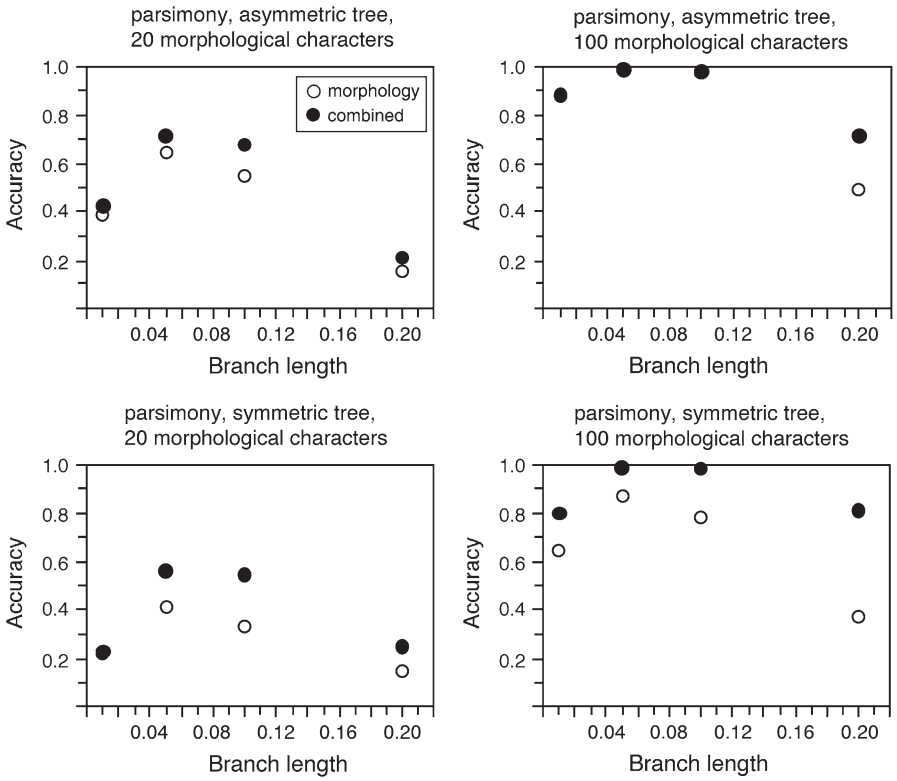

FIGURE 3. Results from simulations showing the accuracy of phylogeny estimation for fossil taxa in a combined analysis of living and extinct taxa, both with (filled circles) and without (open circles) the addition of molecular data to the living taxa. These results show the effects of different branch lengths on parsimony analysis when 50% of the taxa are extant and 50% fossils. Each data point represents the average accuracy from 200 replicates.

The presence of random missing data in the morphological characters for the fossil taxa reduced accuracy relative to analyses when the fossil taxa were more complete, as might be expected (Table 1). However, the presence of random missing data in the fossil taxa did not prevent the combined analysis from increasing accuracy for the fossil taxa (Table 1).

DISCUSSION

In this paper, I ask whether the addition of molecular data can potentially improve phylogeny estimation for fossil taxa. In theory, it seems that if fossil taxa interdigitate among living taxa on a phylogeny, then adding molecular data to the living taxa might improve phylogeny estimation for the fossil taxa, given that morphological data for both living and fossil taxa are included. To test this hypothesis, I simulated molecular and morphological data for both living (DNA + morphology) and fossil (morphology only) taxa. The results support the idea that molecular data can potentially improve phylogenetic accuracy for fossil taxa, with at least some increase under a variety of conditions for both parsimony and Bayesian analysis. In some cases, these increases can be quite dramatic (e.g., a roughly 100% increase; Fig. 3 and Supplementary Appendix 2). Perhaps just as importantly, I found few conditions where this practice led to a substantive decrease in accuracy (with one exception, discussed below). Based on these simulations then, there is potentially much to gain but generally little to lose from combining data from molecules and fossils to improve our understanding of the phylogeny of extinct taxa.

These simulations also suggest the specific conditions where increases in accuracy seem most likely. First, the fossil taxa should interdigitate among the living taxa and should not be too numerous. As an extreme example, if one combines the fossil and living taxa and each are in separate clades, the molecular data will have little or no opportunity to improve estimation for the fossil taxa. Similarly, if the living and fossil taxa interdigitate but the fossil taxa are far more numerous, the overall accuracy of the tree (and the accuracy for the fossil taxa) may not be heavily influenced by the living taxa. Indeed, in simulations where only 25% of the taxa were living, the gains in accuracy for the fossil taxa were negligible (Fig. 2). However, there were often substantial improvements when 50% of the taxa were fossils (Figs 2–6 and Supplementary Appendix 2).

Second, the morphological data must be informative, but not too informative. The benefits of the combined analysis depend on there being enough variation in the morphology to help place the fossil taxa. When there were few informative characters (e.g., when there were

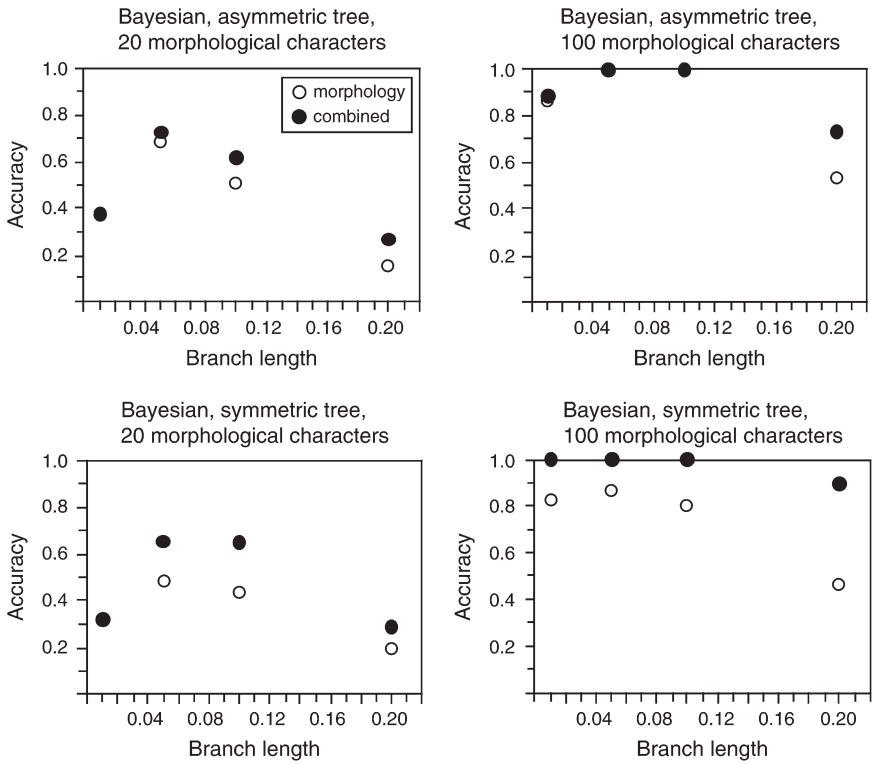

FIGURE 4. Results from simulations showing the accuracy of phylogeny estimation for fossil taxa in a combined analysis of living and extinct taxa, both with (filled circles) and without (open circles) the addition of molecular data to the living taxa. These results show the effects of different branch lengths on Bayesian analysis when 50% of the taxa are extant and 50% fossils. Each data point represents the average accuracy from 100 replicates.

only 20 characters and a branch length of 0.01), there was typically little improvement from adding the molecular data. Conversely, when the accuracy for the morphological data is very high, there is little that the molecular data can improve upon. In the real world, it seems likely that most morphological data sets for fossil taxa are neither completely uninformative (i.e., many nodes are resolved rather than being polytomies) nor are they likely to be entirely accurate (e.g., typically only some nodes are strongly supported by bootstrapping).

One set of results (Fig. 6) showed that the addition of molecular data led to significantly lower accuracy for the fossil taxa (accuracy for combined analysis = 57%) than analysis of the morphology alone (accuracy = 72%; based on a *t*-test, this difference is highly significant with $P < 0.0001$). This occurred for only one topology, and then only when the branches were very long (0.20), 100 characters were sampled, and the fossil taxa retained no more of their ancestral states than living taxa, and then only for Bayesian analysis (Fig. 6). Although it is reassuring that this occurs under such a restricted set of conditions, it is disconcerting that it occurs at all and that the ultimate cause is not obvious. The proximate cause of this pattern seems to be that fossil taxon L, with the longest terminal branch of any fossil taxon, is almost always misplaced toward the base of the tree (below clade F + G) in combined Bayesian analyses but not as frequently in analyses of the morphological data alone. However, it is unclear why this should occur more often when molecular data are added, or why the problem is worse for Bayesian analyses than for parsimony. Interestingly, even though one might expect Bayesian analysis to be less sensitive to long-branch attraction than parsimony, previous results show that this is not necessarily true when analyzing binary data with very long branches, as analyzed here (e.g., Fig. 2*c*,*d* of Wiens 2005). Overall, this incongruous result shows that some caution is warranted when adding molecular data to Bayesian analyses of fossil taxa, even though the other results of this study (including Bayesian analyses with long, unequal branch lengths; Fig. 6) suggest that this practice is generally helpful or at least innocuous.

Major Assumptions

The results of this study generally seem promising for combining molecular and fossil data, but they are based on many simplifying assumptions. First, I treated the fossil taxa as evenly dispersed across the phylogeny of living taxa. If the fossil taxa were randomly distributed, then the improvements documented here would presumably be lessened somewhat in cases where the fossil taxa were more clumped than evenly dispersed among the living taxa. However, clumping of the fossil or living taxa should have no adverse effects on accuracy when the data are combined.

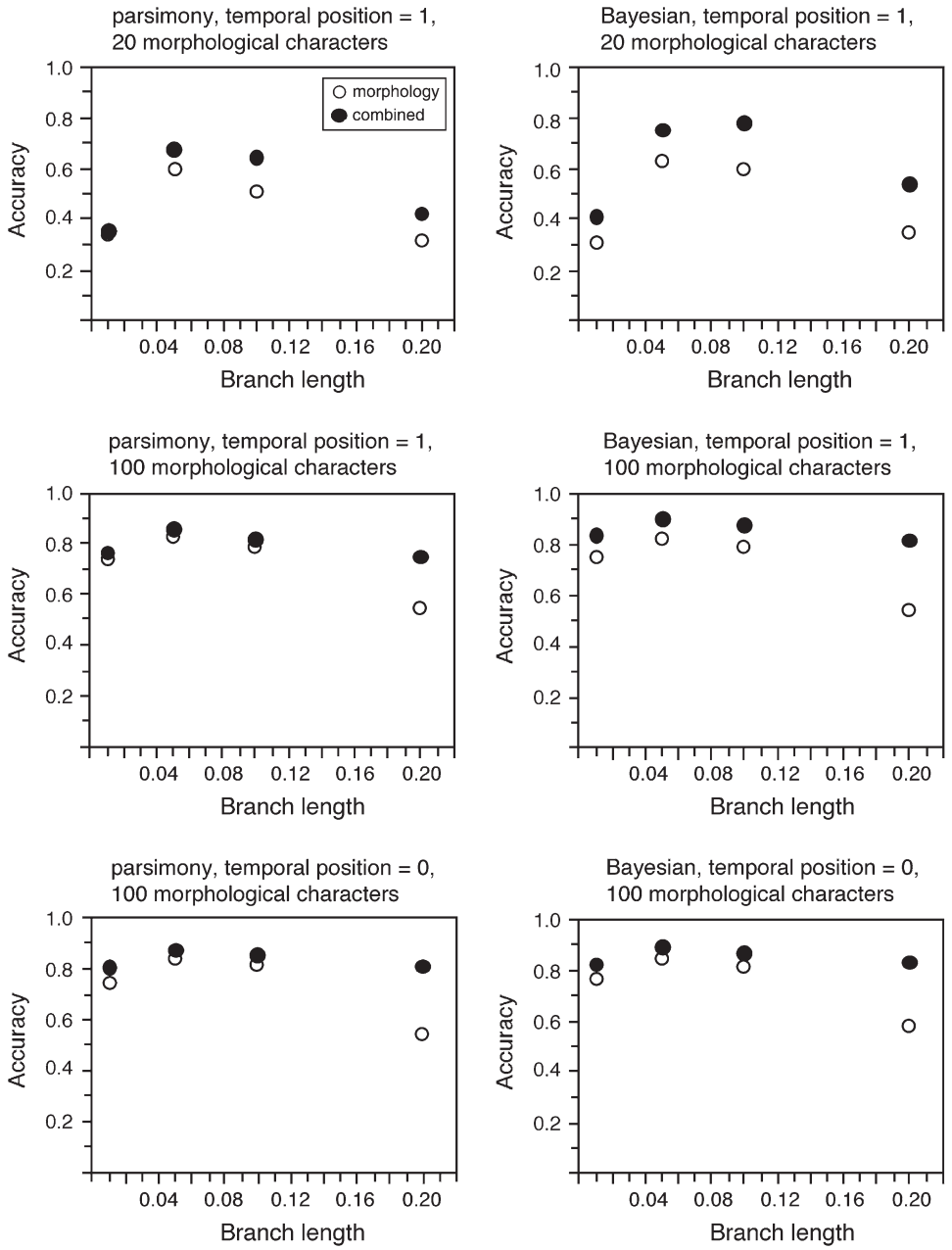

FIGURE 5. Results from simulations showing the accuracy of phylogeny estimation for fossil taxa in a combined analysis of living and extinct taxa, both with (filled circles) and without (open circles) the addition of molecular data to the living taxa. These results show the effects of different branch lengths on parsimony and Bayesian analyses when 50% of the taxa are extant and 50% fossils. One of the simulated topologies (Tree 1; Fig. 1*c*) was used, and the different branch lengths on the *x*-axis represent the mean of all internal and terminal branch lengths for the tree. A temporal position of 1 indicates that the fossil taxa were equivalent to living taxa, whereas a temporal position of 0 indicates that the fossil taxa retain all the character states of their immediate ancestors.

Second, I assumed a simple model of evolution for both the molecular and the morphological data, and that estimation of the molecular tree was straightforward. The purpose of this study was to test if an accurate molecular tree can help improve accuracy for fossil taxa. But the accuracy of many molecular trees is still in doubt (at least in part), even for trees based on large, multilocus data sets. For example, for relatively short but deep branches, there may be problems of long-branch attraction that extend across many genes (e.g., Rokas et al. 2005; Rokas and Carroll 2006). There may also be extensive discordance among gene trees due to incomplete lineage sorting on short branches throughout the tree, leading to weak support in the combined analysis of the genes (e.g., Rokas and Carroll 2006; Wiens et al. 2008).

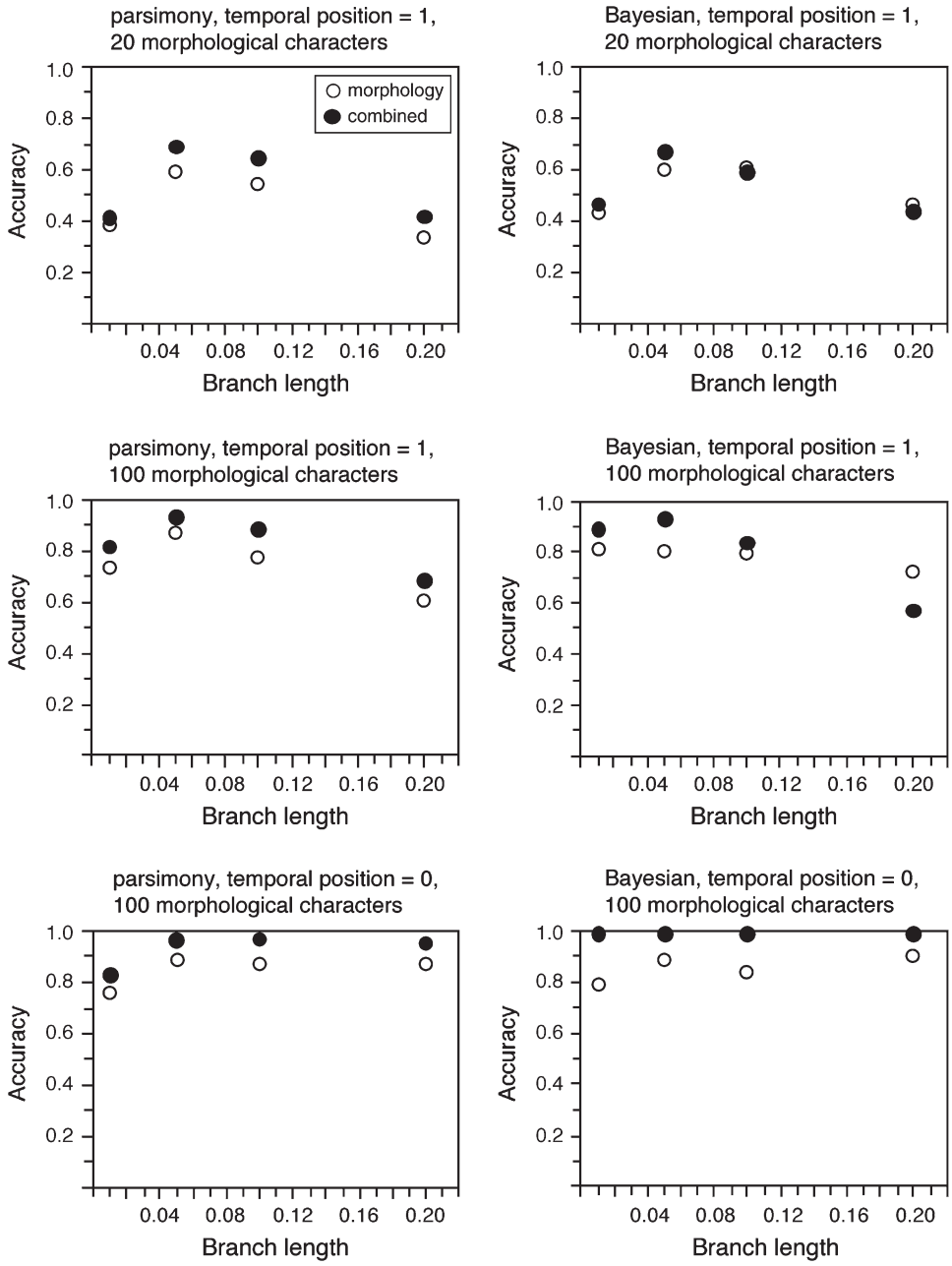

FIGURE 6. Results from simulations showing the accuracy of phylogeny estimation for fossil taxa in a combined analysis of living and extinct taxa, both with (filled circles) and without (open circles) the addition of molecular data to the living taxa. These results show the effects of different branch lengths on parsimony and Bayesian analyses when 50% of the taxa are extant and 50% fossils. One of the simulated topologies (Tree 2; Fig. 1*d*) was used, and the different branch lengths on the *x*-axis represent the mean of all internal and terminal branch lengths for the tree. A temporal position of 1 indicates that the fossil taxa were equivalent to living taxa, whereas a temporal position of 0 indicates that the fossil taxa retain all the character states of their immediate ancestors.

For the morphological data, there may be many factors that may influence accuracy beyond the conditions simulated here (i.e., limited number of characters, high rates of homoplasy, and random missing data), such as correlated character evolution. An important caveat that should be made about these results is that if there is a strongly misleading signal in the morphological data that affects 1 or more of the fossil taxa, the molecular data may do little to directly improve the situation.

Comparisons with Empirical Studies

How do the simulation results compare with those from empirical studies? Here, I review results from 7 empirical studies that have combined molecular and morphological data to address the placement of fossil taxa (Table 2). These studies were chosen because they include trees from the morphological data alone and from the combined molecular and morphological data (including fossils). Many potential studies had to be

TABLE 1. Accuracy of parsimony analysis for fossil taxa in an analysis of living and extinct taxa, when analyzed using morphological data alone (morphology only) or with molecular data added to the living taxa (combined data), contrasting analyses in which the fossil taxa are complete for the morphological characters or have 50% of their characters randomly replaced with missing data cells

	Morphology complete		Morphology missing 50%	
	Morphology	Combined data	Morphology	Combined data
Asymmetric tree				
Length = 0.01	0.880	0.886	0.594	0.581
Length = 0.05	0.988	0.986	0.882	0.904
Length = 0.10	0.976	0.980	0.845	0.883
Length = 0.20	0.493	0.713	0.280	0.432
Symmetric tree				
Length = 0.01	0.645	0.795	0.363	0.393
Length = 0.05	0.870	0.989	0.784	0.884
Length = 0.10	0.782	0.986	0.669	0.878
Length = 0.20	0.373	0.813	0.248	0.454

Note: Results are for 100 morphological characters and 8 of the 16 taxa as fossils. Each value is the average of 200 replicates.

excluded from this summary, mostly because they did not present a comparable separate analysis of the morphological data (e.g., Jordan and Hill 1999; Sun et al. 2002; Hermsen et al. 2006), did not present a tree from the combined analysis including fossils (e.g., Xiang et al. 2005), or had noncomparable taxon sampling in the separate and combined analyses (e.g., Eernisse and Kluge 1993; Magallón 2007).

Of course, it is not possible to address whether the addition of molecular data improves phylogenetic accuracy for fossil taxa in an empirical study. Nevertheless, it is possible to document whether the phylogenetic placement of fossil taxa in the tree from the combined data differs from that in the tree from morphology alone, and in what way.

In 4 of the 7 studies (Shaffer et al. 1997; Asher and Hofreiter 2006; Rothwell and Nixon 2006; Manos et al. 2007), the morphological data alone are unable to fully resolve the relationships of the fossil taxa, and the addition of the molecular data leads to a more fully resolved consensus tree that at least places the fossil taxa more precisely, if not more accurately. This is the basic scenario that one might intuitively expect and is also indirectly supported by the simulations (e.g., accuracy is often increased for the fossils in the combined analyses when the number of morphological characters is limited).

In contrast, the other 3 studies (Gatesy et al. 2003; Asher et al. 2005; O'Leary and Gatesy 2008) actually show lower resolution in the trees from the combined data relative to morphology alone (Table 2). Although this does not necessarily mean that these trees are less accurate, they are more ambiguous. These 3 studies differ from the other 4 in that the authors found extensive and strongly supported conflicts between trees from molecular and morphological data. In fact, 2 of these studies include classic cases of data set conflict (i.e., the placement of cetacean mammals and gavialid crocodilians). The causes of these conflicts remain mysterious, which makes it difficult to determine which tree is correct. Interestingly, these 3 studies also include more fossil taxa than living taxa (Table 2), conditions where simulations suggest that molecular data are unlikely to substantially improve accuracy. Finally, these 3 studies also provide examples where the placement of (at least some) fossil taxa changed considerably with the addition of molecular data, beyond mere differences in resolution.

In summary, these 7 studies demonstrate that in empirical studies, molecular data can both improve resolution for fossil taxa and substantially change their phylogenetic placement. They also seem to support the idea that adding molecular data will improve phylogeny estimation for fossil taxa when the number of fossil taxa is limited relative to the number of living taxa and when the relationships of the fossil taxa are initially unresolved (increasing either resolution in empirical studies or accuracy in simulation studies). These studies also support the prediction that if there are extensive, strongly supported conflicts between the molecular and the

TABLE 2. Summary of 7 empirical studies that compared trees from phylogenetic analyses of living and fossil taxa, before and after the addition of molecular data to the living taxa

Study	Taxon	Taxa (fossil/living)	Characters (morphology/DNA)	Resolved nodes (morphology/combined)
Shaffer et al. (1997)	Turtles	7/23	115/1 300	23/26
Gatesy et al. (2003)	Crocodilians	54/14	164/2 940	53/48
Asher et al. (2005)	Mammals (rodents, lagomorphs)	39/29	228/5 701	61/56
Asher and Hofreiter (2006)	Mammals (tenrecs)	3/20	120/855	12/20
Rothwell and Nixon (2006)	Plants (higher level)	26/30	136/5 072	26/45
Manos et al. (2007)	Plants (Juglandaceae)	5/27	56/2 006	12/33
O'Leary and Gatesy (2008)	Mammals (Cetartiodactyla)	43/28	635/40 928	51/39

Note: All results are from parsimony analysis (only 1 study included a Bayesian analysis of morphology). "Resolved nodes" refers to the number of dichotomous nodes in a strict consensus of the shortest trees from a given analysis.

morphological data (and the fossil taxa outnumber the living taxa), then it may be more difficult for the molecular data to improve estimation for the fossil taxa.

Implications for Molecular Dating Analyses

Combined analyses of molecular and morphological data for living and fossil taxa may also be useful for researchers who use molecular data to determine ages of clades (i.e., fossil-calibrated molecular clock analyses). Typically, researchers determine the ages of extant clades by incorporating the estimated age for the oldest fossil taxon that is assumed to belong to each clade (e.g., Won and Renner 2006; Hugall et al. 2007; Roelants et al. 2007), where this assumption is often based on a previous phylogenetic analysis of morphological data that places the fossil taxon in that extant clade. An alternate approach is to undertake combined analyses, such that the molecular data can help inform the position of the relevant fossils (e.g., Manos et al. 2007). The results presented here suggest that such combined analyses may lead to improved phylogenetic placement for the fossil taxa, which may then in turn improve estimation of the divergence dates.

Other Applications of Molecular Data to Phylogenetic Analysis of Fossils

I began this paper by asking how phylogenomic data might improve phylogenetic analyses of fossil taxa. Although I focused on the efficacy of combined analyses of molecular and morphological data for living and fossil taxa, there are other ways that molecular data could directly or indirectly improve phylogeny estimation for fossil taxa. First, many of the same benefits noted here for combining molecular and fossil data might potentially be obtained by simply enforcing topological constraints in the analysis of fossil taxa, such that relationships among living taxa that are well established by molecular data are fixed in the analysis, without actually including the molecular data in the same matrix as the morphology (e.g., Doyle 2006; for comparison with related approaches, see Manos et al. 2007). Many of the same advantages and disadvantages may pertain to both this approach and the combined approach addressed here (e.g., both will depend on the fossil taxa interdigitating among living taxa). A major advantage of the constraint approach is that it might require less effort than assembling the molecular data and integrating them into a combined analysis. However, if relationships among the living taxa are not fully established by the molecular data, the constraints may lead to overestimating confidence in the relationships among both living and fossil taxa. Similarly, combining living and fossil taxa in the same matrix might actually improve relationships among living taxa as well, despite the disparity in the relative numbers of characters. This idea is supported by simulations (Wiens 2005) and addressed (if indirectly) in many empirical studies (e.g., Rothwell and Nixon 2006; Manos et al. 2007).

Integrated analyses of molecular and morphological data can also be used to improve the methodology of morphology-based phylogenetics, and these improvements can then be incorporated into paleontological studies. For example, if molecular data strongly establish relationships among a set of living taxa, then one can compare how well different methods for analyzing the morphological data perform at reconstructing these quasi-known relationships (e.g., different methods for coding polymorphic morphological characters; Wiens 1998b). Although such analyses are impossible for fossil taxa, methods that perform well in morphological analyses of extant taxa should also perform well for fossil taxa. Similarly, well-supported molecular phylogenies can reveal cases where morphology-based phylogenetics gives strongly misleading results, and critical analysis of the morphology can then offer insights into the processes that cause this to occur and how they might be ameliorated (e.g., Wiens, Bonett, et al. 2005).

Conclusions and Prospects

The results of this study suggest that the new flood of phylogenomic data has the potential to improve accuracy for fossil taxa, in the context of combined analyses of molecular and morphological data for living and fossil taxa. Of course, such combined analyses will not be a panacea for all problems in the phylogenetic analysis of fossil taxa, and even in these simulations, major increases in accuracy occur only under a finite set of conditions. Furthermore, there is no guarantee that this approach will always be effective in the real world, especially when there are strong conflicts between the molecular and the morphological data. However, the simulations do establish that such increases are theoretically possible, and empirical studies suggest that conditions where this seems likely to occur are common (i.e., when morphology alone does not resolve relationships among fossil taxa, but combined analysis does). Furthermore, and perhaps just as importantly, there were no simulated conditions where this approach consistently led to a significant decrease in accuracy for both parsimony and Bayesian analyses.

Supplementary Material

Supplementary material can be found at http://www.oxfordjournals.org/our_journals/sysbio/.

Funding

The U.S. National Science Foundation (EF 0334923).

Acknowledgments

A preliminary version of this work was first presented at a symposium entitled "Bringing together the living and dead: integrating extant and fossil biodiversity in evolutionary studies," sponsored by the Botanical Society of America in August 2006 in Chico, California.

I thank the organizers (Nathalie Nagalingum and Hervé Sauquet) for inviting me to participate and J. Anderson, S. Renner, J. Sullivan, and an anonymous reviewer for helpful comments on the manuscript.

REFERENCES

Anderson J.S. 2001. The phylogenetic trunk: maximal inclusion of taxa with missing data in an analysis of the Lepospondyli (Vertebrata, Tetrapoda). Syst. Biol. 50:170–193.

Asher R.J., Hofreiter M. 2006. Tenrec phylogeny and the noninvasive extraction of nuclear DNA. Syst. Biol. 55:181–194.

Asher R.J., Meng J., Wible J.R., McKenna M.C., Rougier G.W., Dashzeveg D., Novacek M.J. 2005. Stem Lagomorpha and the antiquity of Glires. Science. 303:1091–1094.

Cobbett A., Wilkinson M., Wills M.A. 2007. Fossils impact as hard as living taxa in parsimony analyses of morphology. Syst. Biol. 56: 753–766.

Donoghue M.J., Doyle J.A., Gauthier J., Kluge A.G., Rowe T. 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20:431–460.

Doyle J.A. 2006. Seed ferns and the origin of angiosperms. J. Torrey Bot. Soc. 133:169–209.

Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the Tree of Life from large sequence databases. Science. 306:1172–1174.

Ebach M.C., Ahyong S.T. 2001. Phylogeny of the trilobite subgenus *Acanthopyge* (*Lobopyge*). Cladistics. 17:1–10.

Eernisse D.J., Kluge A.G. 1993. Taxonomic congruence versus total evidence, and the phylogeny of amniotes inferred from fossils, molecules and morphology. Mol. Biol. Evol. 10:1170–1195.

Emerson S.A., Hastings P.A. 1998. Morphological correlations in evolution: consequences for phylogenetic analysis. Quart. Rev. Biol. 73:141–162.

Friedman M. 2008. The evolutionary origin of flatfish asymmetry. Nature. 454:209–212.

Gatesy J., Amato G., Norell M., DeSalle R., Hayashi C. 2003. Combined support for wholesale taxic atavism in gavialine crocodylians. Syst. Biol. 52:403–422.

Gauthier J., Kluge A.G., Rowe T. 1988. Amniote phylogeny and the importance of fossils. Cladistics. 4:105–209.

Grande L., Bemis W.E. 1998. A comprehensive phylogenetic study of amiid fishes (Amiidae) based on comparative skeletal anatomy, an empirical search for interconnected patterns of natural history. Soc. Vertebr. Paleontol. Mem. 4:1–690.

Hallstrom B.M., Kullberg M., Nilsson M.A., Janke A. 2007. Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups. Mol. Biol. Evol. 24:2059–2068.

Hasegawa M., Kishino H., Yano T. 1985. Dating the human–ape split by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22: 160–174.

Hermsen E.J., Nixon K.C., Crepet W.L. 2006. The impact of extinct taxa on understanding the early evolution of Angiosperm clades: an example incorporating fossil reproductive structures of Saxifragales. Plant Syst. Evol. 260:141–169.

Hillis D.M., Wiens J.J. 2000. Molecular versus morphological systematics: conflicts, artifacts, and misconceptions. In: Wiens J.J., editor. Phylogenetic analysis of morphological data. Washington, D.C.: Smithsonian Institution Press. p. 1–19.

Huelsenbeck J.P. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst. Zool. 40:458–469.

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

Hugall A.F., Foster R., Lee M.S.Y. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. Syst. Biol. 56:543–563.

Jordan G.J., Hill R.S. 1999. The phylogenetic affinities of *Nothofagus* (Nothofagaceae) leaf fossils based on combined molecular and morphological data. Int. J. Plant Sci. 160:1177–1188.

Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst. Biol. 51:369–381.

Lewis P.O. 2001. A likelihood approach to inferring phylogeny from discrete morphological characters. Syst. Biol. 50:913–925.

Maddison W.P., Maddison D.R. 2004. Mesquite: a modular system for evolutionary analysis. Version 1.05. Available from: http://mesquiteproject.org.

Magallón S. 2007. From fossils to molecules: phylogeny and the core Eudicot floral groundplan in Hamamelidoideae (Hamamelidaceae, Saxifragales). Syst. Bot. 32: 317–347.

Manos P.S., Soltis P.S., Soltis D.E., Manchester S.R., Oh S.-H., Bell C.D., Dilcher D.L., Stone D.E. 2007. Phylogeny of extant and extinct Juglandaceae inferred from the integration of molecular and morphological data sets. Syst. Biol. 56:412–430.

Novacek M.J. 1992. Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals. Syst. Biol. 41:58–73.

O'Keefe F.R. Wagner P.J. 2001. Inferring and testing hypotheses of cladistic character dependence using character compatibility. Syst. Biol. 50:657–675.

O'Leary M.A. 1999. Parsimony analysis of total evidence from extinct and extant taxa, and the cetacean-artiodactyl question. Cladistics. 15:315–330.

O'Leary M.A., Gatesy J. 2008. Impact of increased character sampling on the phylogeny of Cetartiodactyla (Mammalia): combined analysis including fossils. Cladistics. 24:397–442.

Organ C.L., Schweitzer M.H., Zheng W., Freimark L.M., Cantley L.C., Asara J.M. 2008. Molecular phylogenetics of mastodon and *Tyrannosaurus rex*. Science. 320:499.

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.

Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H., Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Roelants K., Gower D.J., Wilkinson M., Loader S.P., Biju S.D., Guillaume K., Bossuyt F. 2007. Patterns of diversification in the history of modern amphibians. Proc. Natl. Acad. Sci. USA. 104: 887–892.

Rokas A., Carroll S.B. 2006. Bushes in the Tree of Life. PLoS. Biology. 4:e352.

Rokas A., Krueger D., Carroll S.B. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science. 310:1933–1938.

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

Rothwell G.W., Nixon K.C. 2006. How does the inclusion of fossil data change our conclusions about the phylogenetic history of euphyllophytes? Int. J. Plant Sci. 167:737–749.

Rowe T. 1988. Definition, diagnosis, and origin of Mammalia. J. Vertebr. Paleontol. 8:241–264.

Shaffer H.B., Meylan P., McKnight, M.L. 1997. Tests of turtle phylogeny: molecular, morphological and paleontological approaches. Syst. Biol. 46:235–268.

Smith S.A., Arif S., Nieto Montes de Oca A., Wiens J.J. 2007. A phylogenetic hotspot for evolutionary novelty in Middle American treefrogs. Evolution. 61:2075–2085.

Sun G., Ji Q., Dilcher D.L., Zheng S., Nixon K.C., Wang X. 2002. Archaefructaceae, a new basal angiosperm family. Science. 296: 899–904.

Swofford D.L. 2002. PAUP*: phylogenetic analysis using parsimony*. Version 4.0b10. Sunderland (MA): Sinauer.

Takezaki N., Figueroa F., Zaleska-Rutczynska Z., Takahata N., Klein J. 2004. The phylogenetic relationship of tetrapod, coelacanth, and lungfish revealed by the sequences of 44 nuclear genes. Mol. Biol. Evol. 21:1512–1524.

Wible J.R., Rougier G.W., Novacek M.J., Asher R.J. 2007. Cretaceous eutherians and Laurasian origin for placental mammals near the K-T boundary. Nature. 477:1003–1006.

Wiens J.J. 1998a. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625–640.

Wiens J.J. 1998b. Testing phylogenetic methods with tree-congruence: phylogenetic analysis of polymorphic morphological characters in phrynosomatid lizards. Syst. Biol. 47:411–428.

Wiens J.J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.

Wiens J.J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54:731–742.

Wiens J.J., Bonett R.M., Chippindale P.T. 2005. Ontogeny discombobulates phylogeny: paedomorphosis and higher-level salamander phylogeny. Syst. Biol. 54:91–110.

Wiens J.J., Fetzner J.W., Parkinson C.L., Reeder T.W. 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Syst. Biol. 54:719–748.

Wiens J.J., Kuczynski C.A., Smith S.A., Mulcahy D., Sites J.W. Jr., Townsend T.M., Reeder T.W. 2008. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. Syst. Biol. 57:420–431.

Wiens J.J., Moen D.S. 2008. Missing data and the accuracy of Bayesian phylogenetics. J. Syst. Evol. 46:307–314.

Wiens J.J., Reeder T.W. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Syst. Biol. 44:548–558.

Wilkinson M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44:501–514.

Won H., Renner S.S. 2006. Dating dispersal and radiation in the gymnosperm *Gnetum* (Gnetales)—clock calibration when outgroup relationships are uncertain. Syst. Biol. 55:610–622.

Xiang Q., Manchester S.R., Thomas D.T., Zhang W., Fan C. 2005. Phylogeny, biogeography, and molecular dating of cornelian cherries (*Cornus*, Cornaceae): tracking Tertiary plant migration. Evolution. 59:1685–1700.

Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

First submitted 22 July 2008; reviews returned 11 September 2008; final acceptance 30 December 2008
Associate Editor: Susanne Renner