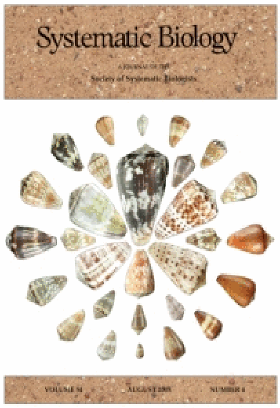

Systematic Biology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713658732

Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence

To link to this article: DOI: 10.1080/10635150601146041
URL: http://dx.doi.org/10.1080/10635150601146041

Syst. Biol. 56(1):17–24, 2007
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150601146041

Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence

Laura Salter Kubatko[1] and James H. Degnan[2]

[1]*Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio 43210, USA;*
E-mail: lkubatko@stat.ohio-state.edu
[2]*Department of Biostatistics, Harvard School of Public Health, Building 2, 4th Floor, 655 Huntington Avenue, Boston, Massachusetts 02115, USA*

Although multiple gene sequences are becoming increasingly available for molecular phylogenetic inference, the analysis of such data has largely relied on inference methods designed for single genes. One of the common approaches to analyzing data from multiple genes is concatenation of the individual gene data to form a single supergene to which traditional phylogenetic inference procedures—e.g., maximum parsimony (MP) or maximum likelihood (ML)—are applied. Recent empirical studies have demonstrated that concatenation of sequences from multiple genes prior to phylogenetic analysis often results in inference of a single, well-supported phylogeny. Theoretical work, however, has shown that the coalescent can produce substantial variation in single-gene histories. Using simulation, we combine these ideas to examine the performance of the concatenation approach under conditions in which the coalescent produces a high level of discord among individual gene trees and show that it leads to statistically inconsistent estimation in this setting. Furthermore, use of the bootstrap to measure support for the inferred phylogeny can result in moderate to strong support for an incorrect tree under these conditions. These results highlight the importance of incorporating variation in gene histories into multilocus phylogenetics. [Coalescence; concatenation; gene tree; maximum likelihood; species tree; statistical inconsistency; supergene.]

As sequence data for multiple genes (loci) become increasingly available, the fields of phylogenetics and phylogeography are faced with the challenge of adapting traditional inference procedures designed for single genes to appropriately analyze multigene data. Recent studies using real data (Chen and Li, 2001; Rokas et al., 2003) have claimed that the procedure of applying standard methods to concatenated multigene data leads to a strongly supported phylogenetic estimate, assumed to be the species tree. This approach has been further supported by simulation-based work that has shown high levels of phylogenetic accuracy as more genes are added to an alignment (Rokas and Carroll, 2005; Gadagkar et al., 2005). However, some authors have noted that differences in individual gene histories can cause the concatenation procedure to fail (Kolaczkowski and Thornton, 2004; Mossel and Vigoda, 2005), though neither of these studies explicitly modeled how such variation in individual gene histories could arise and thus did not address the frequency with which such problems occur with real data.

Numerous processes (e.g., horizontal transfer, gene duplication, incomplete lineage sorting) can lead to discord in the evolutionary histories of genes, but among these, lineage sorting is perhaps the best-studied because it can be mathematically described by the coalescent (Kingman, 1982; Hudson, 1983; Tajima, 1983). Using typical phylogenetic assumptions (e.g., no recombination within genes, no migration or other horizontal gene transfer), and other assumptions due to the coalescent being a large sample approximation to the Wright-Fisher model (large population sizes, panmictic populations, constant population sizes within populations, and selective neutrality; Nordborg, 2001), the coalescent allows computation of the probabilities of individual gene tree topologies for a given species phylogeny when that phylogeny represents the historical relationships among these species (or populations) (Tajima, 1983; Takahata and Nei, 1985; Pamilo and Nei, 1988; Rosenberg, 2002). Recent work (Degnan and Salter, 2005) has expanded the set of trees for which such probabilities can be calculated to include any number of taxa, thus allowing detailed exploration of the effects of tree shape and speciation times on the probability distribution of gene trees.

An important and surprising consequence of this exploration is the recognition that under the coalescent model, the gene tree with the topology that matches that of the species tree need not be the most probable topology (Degnan and Salter, 2005; Degnan and Rosenberg, 2006). Degnan and Rosenberg (2006) describe exact conditions under which a most-probable gene tree has a different topology from an underlying four-taxon species tree and discuss implications of the existence of such gene trees. In this paper, we consider the consequences of applying traditional inference methods to concatenated data under these conditions. In particular, we show using simulation that concatenating multigene data in this setting can lead to phylogenetic estimates that are statistically inconsistent as the number of genes increases, even when an estimation method that is consistent in the number of sites is used with the true mutation model. We further examine the effect on the bootstrap, a standard measure of phylogenetic support, and show that it can provide strong support for an incorrect phylogeny under these conditions.

Methods

Simulations

We assumed that one individual was sampled per species and we measured branch lengths in the species tree (which are the intervals of time between speciation events) in *coalescent units* of $t/(2N)$ where t is the number of generations and N is the effective population size for a diploid population. For example, for a population size of 10^5, a branch length of 0.1 coalescent units corresponds to 20,000 generations. The most probable gene tree can have a different topology from that of the species tree when branch lengths are sufficiently small in coalescent units, either due to a small number of generations or a large effective population size. We refer to gene trees more probable than the gene tree that matches the species tree topology as *anomalous* (Degnan and Rosenberg, 2006). When an anomalous gene tree (AGT) exists, a sample of gene trees generated from the underlying species tree is expected to have more trees with the topology of the AGT than with the topology of the species tree. Thus we expect that the high level of conflict between gene and species trees can make the true species-level relationships poorly supported for concatenated data, even for a very large number of genes.

To test this, we used maximum likelihood (ML) to estimate phylogenies from sequence data simulated on a varying number of gene histories generated under the coalescent process and subsequently concatenated. The model species tree was the asymmetric four-taxon tree (Fig. 1A). This model species tree produces anomalous gene trees whenever the internal branches, particularly the branch with length x, are sufficiently short. This occurs because short species tree branch lengths increase the likelihood that all four gene lineages coalesce prior to the root. When there are four lineages available to coalesce, the 15 rooted topologies are not equally probable, even though lineages are assumed to coalesce at random. This is because asymmetric topologies constrain coalescent events to occur in a particular order (for example C and D must coalesce "before" B, to produce the (A(B(CD))) topology), whereas symmetric topologies have fewer constraints in the order of coalescences, so that A and B can coalesce either before or after C and D coalesce for the ((AB)(CD)) topology. As a result, when all four lineages are available prior to the root, the 12 asymmetric topologies each have probability 1/18, and the three symmetric topologies each have probability 2/18 (Brown, 1994). The increased probability for symmetric topologies results in one or more symmetric topologies being anomalous when species tree branch lengths are sufficiently short.

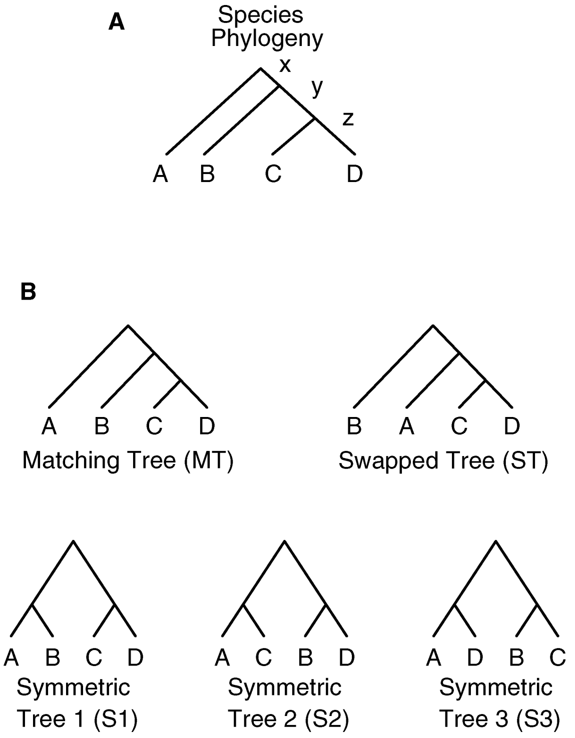

FIGURE 1. (A) Model species tree used in the simulations, with branch lengths x, y, and z. (B) For small x and moderate y, MT, ST, and S1 are the three most probable topologies. When x and y are both sufficiently small, the topologies S1, S2, and S3 are the three most probable. Note that S1, S2, and S3 are all anomalous for certain choices of (x, y) but that ST can never be anomalous (Degnan and Rosenberg, 2006).

In the figures and text that follow, we call the asymmetric gene tree with the topology that is identical to the model species trees the Matching Tree (MT) and the asymmetric tree with the labels on the two most basal taxa switched the Swapped Tree (ST). The three possible symmetric trees with four tips are called Symmetric Trees 1, 2, and 3 (S1, S2, and S3; see Fig. 1B). The remaining labeled topologies with four tips occur with very low frequency regardless of branch lengths under the coalescent model (unless both internal branches of the species tree are very close to zero, in which case the other 10 gene trees are approximately equally frequent) and are not shown here. The two internal branch lengths are represented by the pair (x, y) and determine gene tree probabilities under coalescence. Degnan and Rosenberg (2006) have shown that the collection of (x, y) pairs can be partitioned into two regions: an *anomaly zone* in which P(S1) > P(MT) and a region where there are no AGTs (Fig. 2). We additionally define the boundary of the anomaly zone to be the set of pairs for which P(S1) = P(MT), shown as the upper curve in Figure 2. Furthermore, note that when both x and y are small, the topologies S1, S2, and S3 can all be more probable than MT. This occurs for all points below the lower curve in Figure 2.

Several values of (x, y) were chosen to examine the effect of branch lengths in various regions of this space on the behavior of ML estimation as the number of genes becomes large (Fig. 2). Among the parameter settings considered in the anomaly zone were $(0.01, 1.0)$, $(0.05, 0.05)$, $(0.1\ 0.05)$, $(0.1568, 0.1568)$, and $(0.25, 0.01)$. Outside the anomaly zone, we considered $(0.01, 2.0)$, $(0.05, 1.0)$, and $(0.1, 1.0)$. The points $(0.1568, 0.1568)$ and $(0.25, 0.01)$ lie within the anomaly zone but very close to the boundary, whereas the points $(0.01, 2.0)$ and $(0.05, 1.0)$ also lie near but on the opposite side of the boundary. These points were chosen so that the results for all possible relationships between x and y ($x < y$, $x > y$, and $x = y$) for points close to the boundary could be examined. For each of these pairs, the coalescent was used to simulate samples of independent gene trees using the program COAL (www.coaltree.net) (Degnan and Salter, 2005) for $n = 10$, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, and 6000 genes. To convert gene tree branch lengths from coalescent units to mutation units, gene tree branch lengths were multiplied by $\theta/2$, where $\theta = 4N\mu$, so that all populations were assumed to have equal values for θ. Note that the resulting trees satisfied the molecular clock assumption.

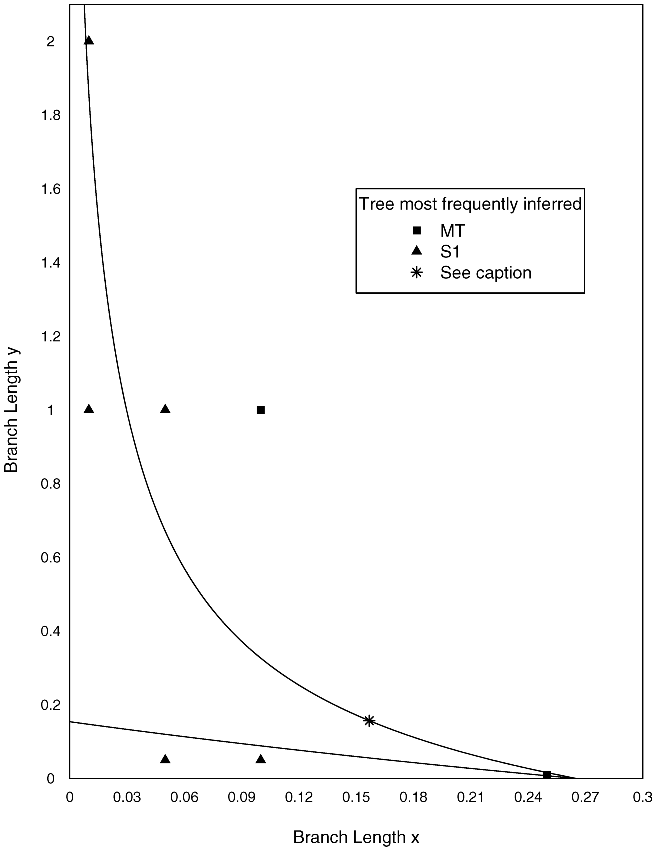

FIGURE 2. Plot of the anomaly zone as a function of the two internal branch lengths. The upper curve is the boundary of the anomaly zone for the species phylogeny in Figure 1A. The boundary extends infinitely in the y direction. For points below this curve, there is at least one AGT. For points below the lower curve creating the approximately triangular region, the three symmetric trees, S1, S2, and S3, are all anomalous. Both curves are based on Degnan and Rosenberg (2006), equations (4) and (5). For the point (0.1568, 0.1568), which is slightly within the boundary, MT is inferred most frequently when $\theta = 0.01$, but MT and S1 are inferred approximately equally often when $\theta = 0.001$ (see Figs. 3D and 4D).

Additionally, inference is affected by z, the time from the present to the most recent common ancestor of C and D, and $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per generation per base pair (bp). We considered $\theta = 0.001$ and $\theta = 0.01$ and examined several values for z including 0.001, 0.2, 1.0, 2.0, and 3.0, though we report only values for $z = 1.0$ here because values of z in this range had a negligible effect.

For each sample of n gene trees, DNA sequences of length 500 were generated for each gene tree under the Jukes-Cantor model using the program Seq-Gen (Rambaut and Grassly, 1997), and the resulting alignments were concatenated. This entire process was repeated 300 times for each n and for each (x, y). To examine the effect of increasing branch length x for fixed y and for $\theta = 0.001$, samples of gene trees were obtained as described above for x varying from 0.02 to 1.0 in increments of 0.02 for $y = 1.0$. This process was repeated 300 times for each x for $n = 10$, 20, 50, and 100.

Phylogenetic Analysis

Maximum likelihood phylogenetic estimates were obtained from the concatenated data using PAUP* v4.0b10 (Swofford, 2003) assuming both the Jukes-Cantor model and a molecular clock (to remove incorrect model specification as a source of error in phylogenetic estimates). To ensure that the exact ML tree was found, all 15 possible trees were evaluated exhaustively for each data set. For each simulation condition, the percentage of times that each topology was selected as the ML estimate was recorded. All replicates for which PAUP* reported a tie for the ML tree were deleted prior to tabulation of topology percentages.

Data sets for the bootstrap analysis were generated in the same manner as described above for (0.01, 1.0), $\theta = 0.001$, and $n = 100$. ML estimates of the phylogeny were obtained as above. PAUP* was then used to perform bootstrapping for 200 bootstrap samples, and the bootstrap support for all possible clades was recorded. This process was repeated 500 times, and the collection of bootstrap proportions for the clades of interest was used to construct Figure 6.

RESULTS

When S1 was anomalous or nearly anomalous, the frequency with which ML correctly inferred the species tree was generally low and was similar to the frequency with which S1 was inferred when the number of genes was small (Figs. 3 and 4). In addition, many choices of (x, y) showed an increase in the frequency of incorrectly inferring S1 with the number of genes, though the rate depended on the choice of (x, y), suggesting that ML estimation using concatenated data is statistically inconsistent for some values of (x, y). As anticipated, the behavior of ML was tied closely to the probabilities associated with each of the gene tree topologies, and convergence was typically faster when there was a greater disparity in the gene tree probabilities for S1 and MT.

The phenomenon of AGTs is particularly sensitive to short branches deep in the tree, and for most of the anomaly zone for this four-taxon tree, $x < y$ (Fig. 2). However, $x < y$ is not a necessary condition for S1 to be the most frequently inferred tree, as can be seen for the point (0.1, 0.05) (Figs. 3G and 4G). When x and y are both small, S1, S2, and S3 can all be anomalous, and MT was then the *fourth* most likely tree to be estimated (Figs. 3 and 4F and G).

To examine the relationship between the anomaly zone and the performance of ML on concatenated sequences, we considered four cases in which the pair (x, y) was near the boundary (see Fig. 2). The point (0.25, 0.01) lies in the anomaly zone, and although the frequency of inferring MT was fairly low for this point (approximately 53% with 6000 genes or three million bases when $\theta = 0.001$), MT was inferred much more often than S1, and this frequency grew with the number of genes (Fig. 3H). Similarly, the point $x = y = 0.1568$ is just inside the anomaly zone, and S1 and MT were roughly equally likely to be inferred as the number of genes increased (Fig. 3D). When

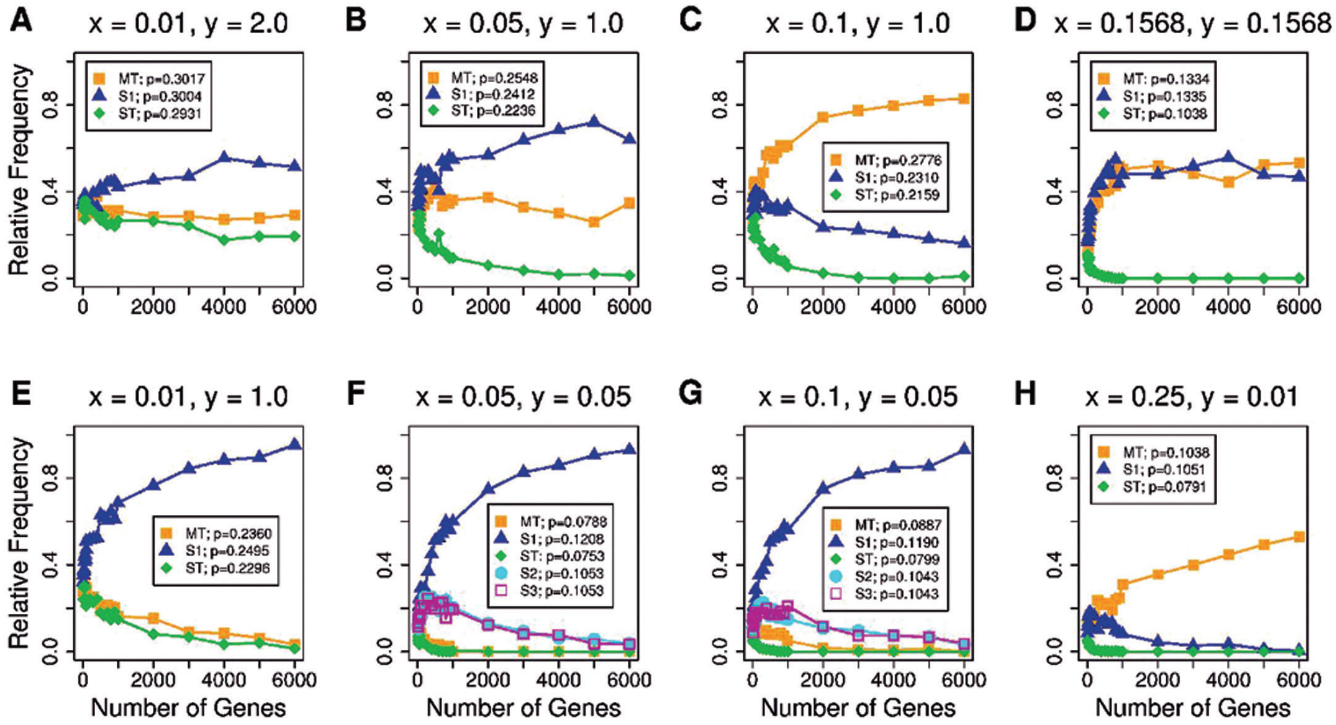

FIGURE 3. The proportion of times trees were estimated as the number of genes increased for $\theta = 0.001$. The legend gives the probability of each gene tree topology for the model species tree under coalescence. The frequencies of the trees displayed here do not always sum to 1.0 because trees other than these were sometimes inferred. Note that S1 is anomalous in all cases except (A), (B), and (C), and that S1, S2, and S3 are all anomalous in (F) and (G).

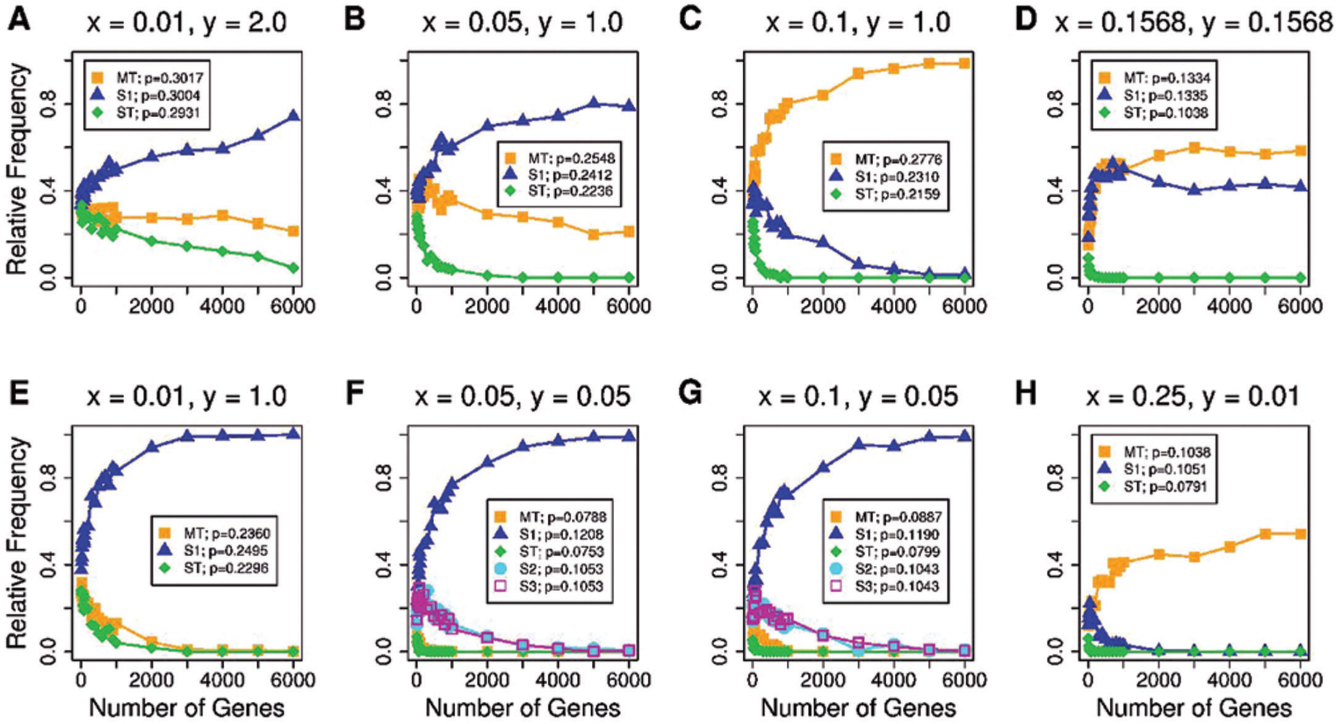

FIGURE 4. The proportion of times trees were estimated as the number of genes increased for $\theta = 0.01$. The legend gives the probability of each gene tree topology for the model species tree under coalescence. The frequencies of the trees displayed here do not always sum to 1.0 because trees other than these were sometimes inferred. Note that S1 is anomalous in all cases except (A), (B) and (C), and that S1, S2, and S3 are all anomalous in (F) and (G).

θ = 0.01 (Fig. 4D), MT was inferred slightly more often than S1 for this case, but both trees had a high probability of being inferred, even with 6,000 genes. Conversely, for the points (0.01, 2.0) and (0.05, 1.0), S1 is not anomalous but was inferred more frequently than MT (Figs. 3 and 4A and B), a surprising result given that S1 is neither the most frequently occurring topology nor the topology of the underlying species tree. Although these results indicate that the existence of an AGT is neither necessary nor sufficient for statistical inconsistency, they demonstrate that ML estimation from concatenated sequences can perform poorly for points in or even near the anomaly zone.

Inference from sequence data also depends on the parameters z and θ, although these values do not affect gene tree probabilities. Both of these parameters influence the amount of sequence evolution, and therefore the amount of variability in the DNA sequences. For extremely small values of θ, there is likely to be high sequence similarity among taxa, and inference of a single tree would be difficult without a large number of genes. For very large values of both z and θ, the sequence data are likely to be noisy due to the long branches over which mutations could accumulate, which would again make inference difficult. Intermediate and realistic values for these parameters showed very little effect on ML estimates. As an example, we considered changing θ from 0.001 to 0.01 (Figs. 3 and 4), which covers the range of typical values in the literature (Rannala and Yang, 2003; Jennings and Edwards, 2005; Kopp and Barmina, 2005). This had a small effect on the rate at which increasing the number of genes raised the frequency of inferring S1, with convergence occurring slightly more quickly in this case. Effects of changing z within a reasonable range were also small (results not shown).

Although branch lengths in this study were chosen to illustrate potential dangers in analyzing concatenated data, a practical question for researchers wishing to estimate species-level relationships is how long branches must be in order to use concatenation to estimate species trees reliably. For a fixed value of y and number of genes n, the probability of (correctly) inferring MT increased as a function of x (Fig. 5). In this setting, we found the frequency with which MT was inferred to be strongly correlated with the probability of MT given the species tree, P(MT), as well as the difference between the gene tree probabilities for MT and S1. We note, however, that in spite of this correlation, an unmanageably large number of genes may be required for the frequency of recovering MT to be close to 100%. For example, when $(x, y) = (0.4, 1.0)$, MT was recovered approximately 86% of the time for 100 genes, even though for these branch lengths, MT is more than twice as probable as S1 (P(MT) = 0.396 and P(S1) = 0.176). Similarly, when $x = 0.1$, $y = 1.0$, and $\theta = 0.001$, S1 is not anomalous, but the frequency that MT was inferred as a function of the number of genes grew very slowly, reaching only about 83% with as many as 6000 genes (Fig. 3C).

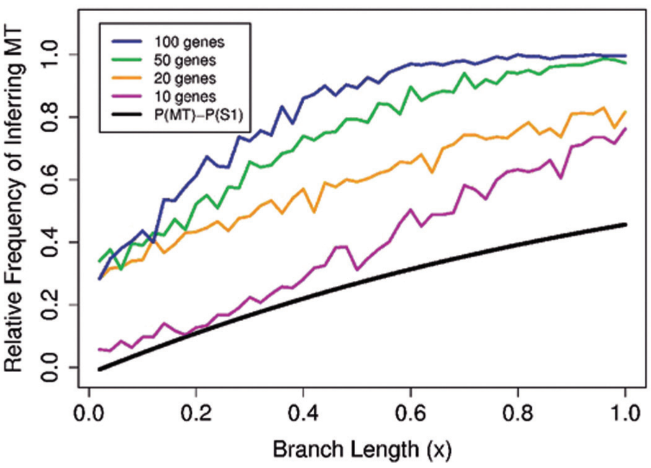

FIGURE 5. Frequency of inferring MT as a function of internal branch length x when $y = 1.0$ for n = 10, 20, 50, and 100 genes. The solid black line is the difference between the probability of MT and the probability of S1 as a function of x.

Confidence in phylogenetic estimates obtained using ML is often assessed through use of the bootstrap (Felsenstein, 1985; Efron, 1996). When data from multiple genes are concatenated, bootstrap resamples can be drawn from the resulting sequences. In the presence of one or more AGTs, more of the data are generated from the AGT(s) rather than from the gene tree matching the species tree, and the bootstrap should show support for the AGT(s). We examined this by considering the case (0.01, 1.0) and $\theta = 0.001$ for 100 genes. For 500 simulated data sets, the ML tree and bootstrap support for various clades were recorded. When S1 was the ML estimate, the bootstrap showed moderate to strong support for the (A,B) clade (Fig. 6A), which appears in S1 but not in the other two trees, whereas the support for (A,B) was lower when MT or ST were the ML estimates. When MT was the ML estimate, moderate to strong support for the clade containing B, C, and D, which is present in the species tree but not in the other trees, was observed, although this support was low when either of the other trees was estimated (Fig. 6B). The results (not shown) were analogous for the (A,C,D) clade (moderate to high bootstrap support for this clade when ST was the ML tree, low support when either of the other trees was the ML estimate). The bootstrap therefore failed to recognize the conflicting signal in the data, and generally gave moderate to strong support for whichever tree was inferred.

DISCUSSION

In this study, we have identified conditions under which concatenation of data from multiple loci can lead to poor performance of standard phylogenetic estimates. These conditions include (1) evolution according to standard phylogenetic and coalescent assumptions; (2) widespread incomplete lineage sorting, due to species tree branch lengths that are short relative to effective population size; and (3) sampling one individual per species. In this section, we examine these assumptions

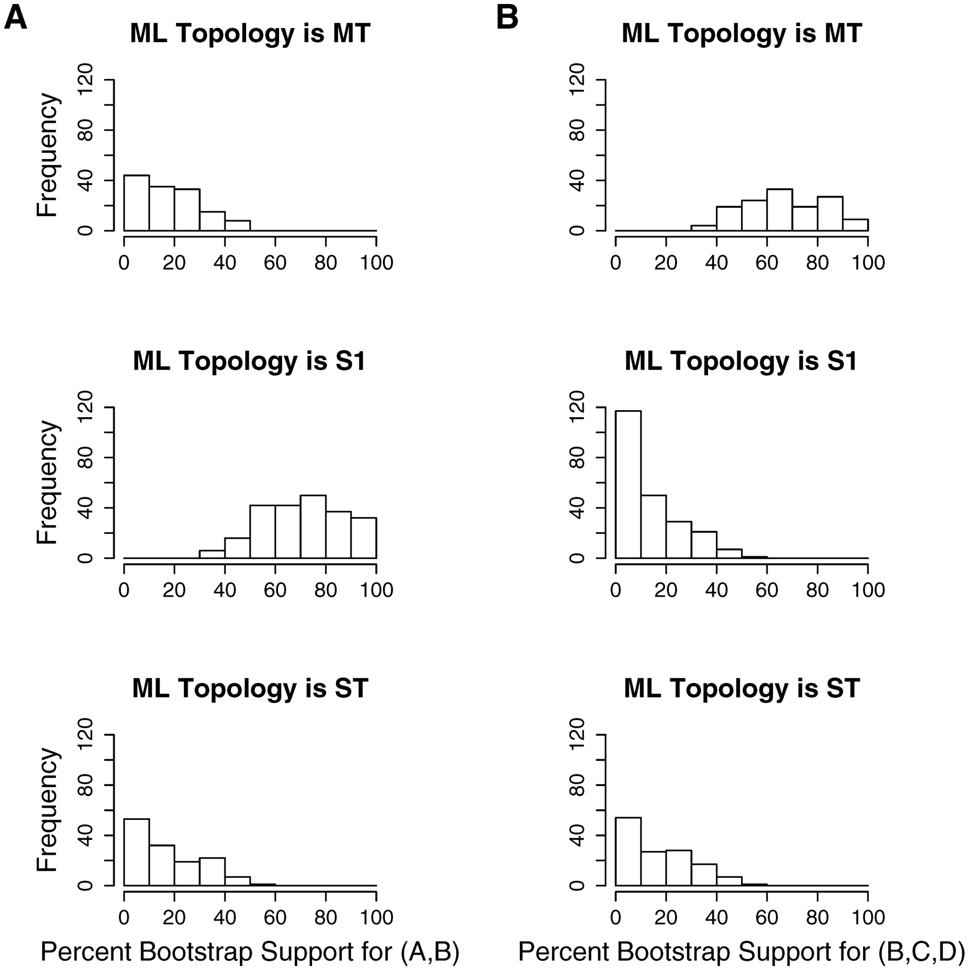

FIGURE 6. Bootstrap support for various clades in the simulated data. (A) Support for the (A,B) clade. (B) Support for a clade containing taxa B, C, and D.

and discuss both how frequently they can be expected to be met for real data and how changes in these conditions might affect phylogenetic performance.

We first address the issue of widespread incomplete lineage sorting, the most severe instance of which occurs when an AGT exists. Degnan and Rosenberg (2006) thoroughly characterize the conditions required for existence of four-taxon AGTs and show that for three taxa, no gene tree can be anomalous for any species tree. For four taxa, the asymmetric tree used here produces AGTs when branch lengths are in the regions delineated in Figure 2, which correspond to either one or both internal branches in the species phylogeny being "short." We return to a discussion of what constitutes a "short" branch below. Degnan and Rosenberg further demonstrate that there are no AGTs for the symmetric four-taxon species phylogeny, but that any species tree topology with five or more taxa has at least one AGT, regardless of the level of symmetry (Degnan and Rosenberg, 2006, proposition 2). In addition, we note that, even when species tree branches are not short enough to produce AGTs, moderately short branches in the species tree can still lead to poor performance of standard phylogenetic inference procedures applied to concatenated data. For example, the species phylogeny is estimated correctly less than 80% of the time for 10 genes when both internal branches on the species tree are 1.0 coalescent units (Fig. 5).

Although it is possible for any species tree topology with more than four taxa to have an AGT, the existence of an AGT requires that one or more branch lengths within the species phylogeny are short. Short species tree branches indicate a small number of generations relative to the effective population size. For example, a branch of length 0.01 coalescent units (used several times in our simulations) might correspond to 1000 generations in a population with effective size 50,000 or to 10,000 generations in a population with effective size 500,000. The largest branch length used in our simulation study was 1.0 coalescent units, which, for example, could correspond to 1,000,000 generations for a population with effective size 500,000. It is therefore helpful to consider scenarios under which such short branches are likely to arise.

First, short branches in species-level phylogenies can result from adaptive radiation or rapid diversification, as has been hypothesized for several species of birds (Poe and Chubb, 2004; Edwards et al., 2005; McCracken and Sorensen, 2005), for fruit flies (Kopp and Barmina, 2005), and for some fish (Verheyen et al., 2003). Second, as taxon sampling increases within a fixed group, we would necessarily expect branches to become shorter and problems with incomplete lineage sorting to become more pronounced (Edwards et al., 2005; Maddison and Knowles, 2006). Finally, we note that short branches are expected to primarily reflect more recent divergences, due to the fact that reciprocal monophyly is generally achieved within five coalescent time units (Rosenberg, 2003). In addition, deeper clades in large trees might be expected to be separated by long branches as a result of extinction. A consequence is that in larger trees with short branches confined to edges near the tips of the tree, the extent of incongruence will be relatively minor, and deeper clades might be relatively well-resolved using concatenated data. However, Edwards et al. (2005) argue that there is the potential for incomplete lineage sorting to be important deeper in the tree even in large phylogenies, because the phenomenon depends mainly on the length of the internal edge, rather than on the depth of that edge within the tree. Maddison and Knowles (2006) note that when short branches occur deep within the tree, it will be difficult to overcome the effects of incomplete lineage sorting by increased sampling of either genes or individuals within species. We also note that for larger trees, a single short branch deep in the tree can lead to an AGT that is topologically close to the species tree. Thus, although this single branching event might be difficult to resolve even with large amounts of data, mistaking the AGT for the species tree topology in this case would still lead to an estimated tree that was topologically close to the species tree.

In empirical studies, we will not know either the length of species tree branches or whether an AGT exists. However, the potential for this problem can be recognized when substantial discord is observed among the gene trees estimated independently for each locus. An empirical example is the study of Australian grass finches conducted by Jennings and Edwards (2005), in which they consider the coalescent as a possible explanation. They obtained estimates of the gene trees from thirty loci for three taxa and found that 16, 7, and 5 of the genes supported each of the three possible phylogenies for three taxa (two gene trees were unresolved). A rough estimate of the internal branch length for their tree can be obtained by using their ML estimates of ancestral population size and number of generations (their table 3), resulting in an estimated branch length of 0.3. Although this is larger than most of the branch lengths considered in our simulation study (Figs. 3 and 4), Figure 5 demonstrates that the frequency of inferring the true tree for a branch of this magnitude might be expected to be fairly low, even though there is not an AGT in this case. We also note that under the coalescent model with an internal branch of length 0.3, the three gene tree topologies have probabilities 50.6%, 24.7%, and 24.7% (Pamilo and Nei, 1988), which are an excellent fit with the frequencies of the three gene topologies observed by Jennings and Edwards. Thus, it is reasonable to conclude (as Jennings and Edwards do) that the coalescent is likely a factor in the high levels of incongruence observed in their study. In addition to looking at empirical studies, simulation studies that model speciation (for example, Huelsenbeck and Lander, 2003, used a linear birth-death process) could be useful in examining expected branch lengths and the effects of extinction and taxon sampling on these branch lengths.

Maddison and Knowles (2006) have recently used simulation to address the possibility of substantial gene tree incongruence arising as a consequence of coalescence, and the effect that this may have on the ability to infer the species trees. They examined the effect of incomplete lineage sorting on two procedures for estimating species phylogenies that incorporate the coalescent in the estimation procedure to some extent. They also examined the effect of sampling more lineages within taxa. Their findings are similar to ours, in that shorter branches in their species phylogenies led to more difficulty in estimation of the species tree using either of the approaches they considered. In situations in which substantial levels of gene tree incongruence can be expected, they observed that it was more beneficial (in terms of phylogenetic accuracy) to sample more individuals per species than to sample more genes, because each individual from a species provides an independent opportunity to observe coalescence with an individual from the sister species. For species phylogenies with longer branches, however, there appeared to be a slight advantage to sampling more genes with only a single individual per species. Also, if external branches are long (e.g., more than five coalescent time units; see Rosenberg, 2003), but there are short internal branches, the benefits of sampling individuals within species may be lost because monophyly is likely to be achieved on the external branches, and therefore the number of lineages in the critical short internal branches would not be increased (Degnan and Rosenberg, 2006).

Although we have focused here on the coalescent as the sole process that generates discord among gene trees, many other evolutionary processes might lead to substantial incongruence, including gene flow, selection, hybridization, and gene duplication. Many of these processes undoubtedly make estimation of the species tree even more challenging, and they, too, should be modeled and examined. In addition, violation of coalescent assumptions might either diminish or augment problems in species tree inference. For example, if population sizes fluctuate, the long-term effective population sizes are reduced, which would have an effect similar to lengthening a branch and would therefore reduce incomplete lineage sorting and gene tree incongruence. Here, we demonstrate that even in the simple setting of constant population size, no selection, and no population stratification, species tree estimation using concatenated data can perform poorly. Importantly, we find that bootstrap support for the species tree estimated from concatenated data can be high, even when that estimate is incorrect. The potential for this problem with the bootstrap in similar settings where substantial differences in the evolutionary processes of different genes exist has been recognized by others (Gadagkar et al., 2005; Delsuc et al., 2005).

We have demonstrated that estimation of species trees using ML on concatenated data can be statistically inconsistent when substantial incomplete lineage sorting results from short internal branches in the species trees. However, in studies in which speciation events are separated by large time intervals relative to the effective population size, leading to high levels of congruence between individual gene trees, standard phylogenetic methods applied to concatenated multilocus data may still be expected to perform well (Gadagkar et al., 2005). When substantial discord in individual gene trees is observed and is believed to be due to incomplete lineage sorting, sampling more individuals per species may be beneficial (Maddison and Knowles, 2006). Although increased taxon sampling can be helpful in some settings, such as breaking up long branches (Hillis, 1998), if short branches occur deep in the tree (Maddison and Knowles, 2006) or incongruence is due to some other factor (e.g., Kolaczkowski and Thornton, 2004), then increased levels of sampling are not expected to improve phylogenetic accuracy. Recent advances in understanding the probability distribution on gene trees under the coalescent (Rannala and Yang, 2003; Degnan and Salter, 2005) make possible the development of probabilistic methods such as ML and Bayesian techniques (Liu and Pearl, 2006) that explicitly incorporate the coalescent, as suggested by several authors (Felsenstein, 1988, 2004; Maddison, 1997). We agree with others (Maddison, 1997; Edwards et al., 2005; Maddison and Knowles, 2006) who have suggested that these methods will be useful tools for species tree estimation in cases such as those examined in this study.

Acknowledgments

We thank N. Rosenberg, W. B. Jennings, S. Edwards, M. Hickerson, T. Lowrey, C. Moritz, T. Turner, N. Laird, J. Galovich, and A. J. Powell for helpful discussions. We thank Roderick Page, Tim Collins, and the anonymous reviewers for valuable comments during the review process. Both authors were supported by grants from NSF (DMS0104290 and 0505265) to L. K. J. D. was also supported by NIH grant MH59532 to N. Laird. This work was completed while L. K. was supported by the Mathematical Biosciences Institute at the Ohio State University.

References

Brown, J. K. M. 1994. Probabilities of evolutionary trees. Syst. Biol. 43:78–91.

Chen, F.-C., and W.-H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68:444–456.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics 2:762–768.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

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

Edwards, S. V., W. B. Jennings, and A. M. Shedlock. 2005. Phylogenetics of modern birds in the era of genomics. Proc. R. Soc. B 272:979–992.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:7085–7090.

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

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22:521–565.

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

Gadagkar, S. R., M. S. Rosenberg, and S. Kumar. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. J. Exp. Zool. 304B:64–74.

Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.

Jennings, W. B., and S. V. Edwards. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution 59:2033–2047.

Kingman, J. F. C. 1982. The coalescent. Stoch. Proc. Appl. 13:235–248.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and maximum likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Kopp, A., and O. Barmina. 2005. Evolutionary history of the *Drosophila bipectinata* species complex. Genet. Res. 85:23–46.

Liu, L., and D. K. Pearl. 2006. Reconstructing posterior distributions of a species phylogeny using estimated gene tree distributions. Mathematical Biosciences Institute Technical Report #53. The Ohio State University, Columbus, Ohio.

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

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

McCracken, K. G., and M. D. Sorensen. 2005. Is homoplasy or lineage sorting the source of incongruent mtDNA and nuclear genes in the stiff-tailed ducks? Syst. Biol. 54:35–55.

Mossel, E., and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309:2207–2209.

Nordborg, M. 2001. Coalescent theory. Pages 179–212 *in* Handbook of statistical genetics (D. Balding, M. Bishop, and C. Cannings, eds.). Wiley, Chichester.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Poe, S., and A. L. Chubb. 2004. Birds in a bush: Five genes indicate explosive evolution in avian orders. Evolution 58:404–415.

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

Rannala, B., and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

Rokas, A., and S. B. Carroll. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337–1344.

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

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.

Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species: Probabilities of monophyly paraphyly, and polyphyly in a coalescent model. Evolution 57:1465–1477.

Swofford, D. L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.

Verheyen, E., W. Salzburger, J. Snoeks, and A. Meyer. 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science 300:325–329.

First submitted 7 February 2006; reviews returned 21 April 2006; final acceptance 7 July 2006

Associate Editor: Tim Collins