Syst. Biol. 58(1):130–145, 2009

DOI:10.1093/sysbio/syp017
Advance Access publication on May 21, 2009

The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference

ALAN R. LEMMON[1,2,3,*], JEREMY M. BROWN[1], KATHRIN STANGER-HALL[4], AND EMILY MORIARTY LEMMON[1,3]

[1]*Section of Integrative Biology, University of Texas at Austin, 1 University Station C0930, Austin, TX 78712, USA;*
[2]*Present address: Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4120, USA;*
[3]*Present address: Department of Biological Science, Florida State University, Tallahassee, FL 32306, USA;*
[4]*Plant Biology Department, University of Georgia, 403 Biosciences Building, Athens, GA 30602, USA;*
**Correspondence to be sent to: Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4120, USA; E-mail: alemmon@evotutor.org.*

Abstract.—Although an increasing number of phylogenetic data sets are incomplete, the effect of ambiguous data on phylogenetic accuracy is not well understood. We use 4-taxon simulations to study the effects of ambiguous data (i.e., missing characters or gaps) in maximum likelihood (ML) and Bayesian frameworks. By introducing ambiguous data in a way that removes confounding factors, we provide the first clear understanding of 1 mechanism by which ambiguous data can mislead phylogenetic analyses. We find that in both ML and Bayesian frameworks, among-site rate variation can interact with ambiguous data to produce misleading estimates of topology and branch lengths. Furthermore, within a Bayesian framework, priors on branch lengths and rate heterogeneity parameters can exacerbate the effects of ambiguous data, resulting in strongly misleading bipartition posterior probabilities. The magnitude and direction of the ambiguous data bias are a function of the number and taxonomic distribution of ambiguous characters, the strength of topological support, and whether or not the model is correctly specified. The results of this study have major implications for all analyses that rely on accurate estimates of topology or branch lengths, including divergence time estimation, ancestral state reconstruction, tree-dependent comparative methods, rate variation analysis, phylogenetic hypothesis testing, and phylogeographic analysis. [Ambiguous characters; ambiguous data; Bayesian; bias; maximum likelihood; missing data; model misspecification; phylogenetics; posterior probabilities; prior.]

Phylogenetic analysis has become well established as an important research tool in the biological sciences (Harvey et al. 1996; Avise 2006), with applications spanning broad fields of research, including evolution (Murphy et al. 2001; Bowers et al. 2003; McKenna and Farrell 2005), ecology (Armbruster 1992; Webb 2000), and medicine (Bush et al. 1999; Hillis 2000; Eickmann et al. 2003). Numerous studies have demonstrated that model misspecification can affect the accuracy of phylogenetic estimates (Kuhner and Felsenstein 1994; Yang et al. 1994; Sullivan et al. 1995; Lockhart et al. 1996; Lemmon and Moriarty 2004). An important, but unresolved, question is whether ambiguous data affect the accuracy of phylogenetic estimates (Kearney 2002; de Queiroz and Gatesy 2007; Wiens 2003a, and references therein). The answer to this question is becoming increasingly relevant as more studies combine data sets. With partial sequences of more than 165 000 taxa now available in large sequence databases, such as GenBank, EMBL, or DDBJ, an increasing number of studies will use large-scale combinations of sequences from these databases in meta-analyses. Incomplete sequences and sampling biases in these databases have led researchers to build phylogenetic data sets that have large numbers of ambiguous characters and gaps (Driskell et al. 2004).

The effects of ambiguous data are unclear, at least in part, because the terminology used to describe the problem has neither been defined carefully nor used consistently across studies. Consequently, we begin by clarifying our terminology. We define the data as a matrix of cells with rows and columns corresponding to sequences and homologous sites, respectively. The value in each cell represents the character state for the corresponding sequence and site. The state of each character is *unambiguous* (taking the state "A," "C," "G," or "T"), *partially ambiguous* (taking the state "B," "D," "H," "V," "S," "W," "R," "Y," "K," or "M"), or *ambiguous* (taking the state "?" or "N"). *Ambiguous character* is used to refer to a character with an ambiguous state. Note that unless explicitly modeled, a gap (represented by the state "–" and resulting from an insertion or a deletion) will have the same effect as an ambiguous character. Also note that partially ambiguous character states are not considered here for simplicity. We use the term *ambiguous site* to refer to a site containing 1 or more ambiguous characters and the term *ambiguous sequence* to refer to a sequence containing 1 or more ambiguous characters. Last, we use *invariable site* to refer to a site in which all unambiguous characters have the same state. To ensure clarity, we henceforth avoid using the term "missing data," although the reader may think of ambiguous or gap characters as missing data.

Because of the complexity of the problem and the fact that conclusions from simulation studies are conflicting, the potential impact of ambiguous characters is still unclear. Early evidence suggested that ambiguous characters can reduce phylogenetic accuracy, especially when taxa have large numbers of ambiguous characters (Platnick et al. 1991). Subsequent studies of the effects of ambiguous characters disagree in their conclusions. For example, Wiens (1998, 2003a, 2003b) argued that adding ambiguous sequences or sites to a phylogenetic analysis has no detrimental effect and that the addition of more

of these sequences increases accuracy, even if it means adding ambiguous characters. This argument seems counterintuitive because adding in-group taxa will decrease the average length of internal branches and thus make the estimation of phylogenetic relationships more difficult (Jermiin et al. 2004). His analyses also suggest that even highly ambiguous sequences have little impact on the phylogenetic relationships of the unambiguous sequences (Wiens 2003a, 2003b). In contrast, other studies (Huelsenbeck 1991; Hillis et al. 1992; Bull et al. 1993; Wiens and Reeder 1995; Dragoo and Honeycutt 1997) found that ambiguous characters resulted in reduced phylogenetic accuracy, although the severity of the effect was variable. This variation was attributed to the number of ambiguous characters (Huelsenbeck 1991; Hillis et al. 1992; Bull et al. 1993; Wiens and Reeder 1995), the type of data (e.g., DNA vs. restriction sites; Wiens and Reeder 1995), the taxonomic distribution of the ambiguous characters (Dragoo and Honeycutt 1997), or the topological information in the data set (Dragoo and Honeycutt 1997; Wiens 2003b; Philippe et al. 2004).

Most of the past research into the effects of ambiguous characters focused on maximum parsimony analysis (Huelsenbeck 1991; Platnick et al. 1991; Wiens 1998; Kearney and Clark 2003; Wiens 2003b). More recently, studies also have considered the effects of ambiguous characters on neighbor joining (Wiens 2003a), maximum likelihood (ML) (Dunn et al. 2003; Wiens 2003a, 2006; Gouveia-Oliveira et al. 2007), or Bayesian (Wiens 2006; Wiens and Moen 2008) analyses. Although Wiens (2006) concluded that adding ambiguous sequences or sites to a data set increased phylogenetic accuracy for maximum parsimony, neighbor joining, ML, as well as Bayesian analyses, Dunn et al. (2003) found a 50% reduced accuracy for maximum parsimony and no reduction in accuracy for ML when ambiguous characters were added.

There are 2 conflicting views on how ambiguous characters affect accuracy. Some authors argue that reduced accuracy is due to a lack of information (i.e., not enough unambiguous characters) rather than due to the ambiguous characters per se (Kearney and Clark 2003; Wiens 2003a). They view ambiguous character states simply as representing the unknown, with little impact on the outcome of a phylogenetic analysis (Kearney and Clark 2003). The practical implication of this view is that all available data, including ambiguous sequences and sites, should be used in a phylogenetic analysis to maximize the available information. In contrast, other studies suggest that ambiguous characters bias the resulting phylogeny to an extent that goes beyond the lack of information (Huelsenbeck 1991). The recommendation in this case would be to reduce the data set in an effort to eliminate as many ambiguous characters as possible. Even though rarely stated explicitly, this latter strategy is widely used, either by setting arbitrary limits for including ambiguous sequences into an analysis (Kearney and Clark 2003) or by excluding the leading and trailing ends of sequences in a data set. Though some researchers are careful to remove ambiguous characters, others disregard their potential effects and use largely ambiguous matrices in an effort to maximize the number of taxa and genes sampled. Because simulation studies have drawn conflicting conclusions regarding the effects of ambiguous characters, and the exact mechanism by which ambiguous characters may affect phylogenetic accuracy remains unknown (Wiens 2006), further investigation is needed.

Previous studies conflict because the approaches taken by authors have confounded the effects of ambiguous characters with the effects of phylogenetic information (due to nucleotide substitutions). More specifically, authors have manipulated data sets by either adding sites containing more than 2 unambiguous characters or by changing the state of characters from unambiguous to ambiguous. In either case, authors have inadvertently manipulated the amount of phylogenetic information along with the number of ambiguous characters. Furthermore, different simulation studies varied widely in their assumptions. For example, Wiens (1998, 2003a, 2003b) assumed that all characters evolved at the same rate and that branch lengths were equal. In contrast, Dunn et al. (2003) assumed that characters and lineages evolved at different rates and allowed for different branch lengths. To determine the consequences of including ambiguous characters in phylogenetic analyses, it is necessary to separate these confounding variables and explore the effects of ambiguous characters across a wide range of parameter space.

The goals of this study are 1) to determine whether ambiguous characters bias estimates of phylogeny, and if they do, 2) to understand the mechanism by which this bias is introduced, and 3) to identify the factors that contribute to the direction and magnitude of the bias. Our approach differs from previous studies in that we only introduce ambiguous sites that should be topologically uninformative if ambiguous characters have no effect. In this way, we are able to remove the confounding factors described above and arrive at a clear understanding of the effects of ambiguous characters. We show that at least 5 factors determine the direction and magnitude of bias resulting from ambiguous characters: the number and taxonomic distribution of ambiguous characters, the strength of topological support from unambiguous characters, the degree of among-site rate variation, and the method and assumptions of the analysis (including the priors assumed in a Bayesian analysis). Although we focus on ambiguous characters, we expect gaps due to insertions and deletions to have the same effect, unless they are explicitly modeled. We conclude by discussing the implications of this work and introduce several possible solutions to the problem.

Methodological Overview

In the following section, we outline the simulation conditions and the general conditions under which the analyses were conducted. In the Results, we present the specific conditions under which the analyses were

conducted along with the result to which they pertain. The factors found to contribute to the effects of ambiguous characters are presented with increasing complexity, beginning with individual factors and ending with combinations of factors. In this way, the reader can more easily understand the effect of each factor as well as the interactions among the factors.

METHODS

In order to gain a clear understanding of the effects of ambiguous characters on estimates of phylogeny, our analyses incorporated the following 4 properties: First, we primarily used simulated data (instead of empirical data) in order to gain control over the factors affecting our phylogenetic estimates and to vary those factors independently. Second, we simplified the simulations as much as possible, focusing on variables that were of immediate interest. Consequently, our analyses were based on 4-taxon simulations under simple models of evolution. Third, our simulated data sets contained 2 regions (sets of nucleotide sites). The first region, which was of fixed length, contained only unambiguous characters and provided a baseline amount of support for the true topology. The second region was of variable length and may have contained ambiguous characters. Fourth, we were careful to include ambiguous characters in such a way that we could eliminate other confounding factors. More specifically, ambiguous sites provided no topological information because only 2 of the 4 characters had unambiguous states. In this way, we were able to remove the effects of substitutions in the ambiguous sites that could affect support for the true topology. In order to determine the effects of Bayesian priors, we also compared results from ML and Bayesian analyses (Hillis et al. 1996; Felsenstein 2004), when possible.

Data Simulation

We first generated 6 alignments, each comprising 500 nucleotides, using a 4-taxon tree (in which each branch length was equal to 1.0 My) and the Jukes–Cantor (JC) model of evolution (Jukes and Cantor 1969). The 6 types of data differed in that they were simulated under 6 different rates of evolution: 0.000015, 0.0015, 0.015, 0.15, 1.3, and 15.0 substitutions per site per My (refer to Fig. 1; note that 1.3 is not a typographical error). These rates were chosen, based on preliminary simulations, to produce data sets containing a range of phylogenetic information, resulting in posterior probabilities (given 500 unambiguous sites) for the true topology of 1/3, 2/3, 1, 1, 2/3, and 1/3, respectively. The lowest rate of evolution produced data sets that were invariable at all sites and the highest rate produced data sets that were saturated. All data sets were simulated with Seq-Gen 1.2.5 (Rambaut and Grassly 1997).

To introduce among-site rate variation, we produced 36 types of combined data sets by concatenating pairwise combinations of the 6 rate types outlined above. Each of the 36 combined data sets thus contained 1000 sites. We refer to the first 500 sites as Gene A and the remaining sites as Gene B. To vary the taxonomic distribution of ambiguous characters across data sets, we replaced all 500 characters in Gene B with ambiguous characters (taking the state "?") for 2 of the taxa (either sister or nonsister on the 4-taxon tree) or none of the taxa (Gene B unambiguous). Last, we varied the length of Gene B by removing between 0 and 500 of the sites at the end of Gene B (in increments of 50). One hundred replicates of each of the 36 types were created for a total of 118 800 data sets (36 rate combinations × 100 replicates × 3 ambiguous character distributions × 11 lengths).

ML Analyses

Two types of ML analyses were performed. The first type was performed to identify the effect of ambiguous characters on estimates of topology. To identify the ML topology for a given 4-taxon data set, each of the 3 possible topologies was scored using PAUP* v.4.0b10 (Swofford 2003) under the JC model of evolution. Branch lengths were optimized using default settings, which include collapsing short branches ($<10^{-8}$ substitutions per site) to polytomies. The topology with the highest likelihood after optimization was chosen as the ML topology. We also computed the likelihood of the data, given topologies with fixed branch lengths (see description below). In order to accommodate rate heterogeneity in data sets with rate variation across genes, we conducted additional ML analyses using TreeFinder (Jobb 2008).

The second type of ML analysis was performed to identify the effect of ambiguous characters on estimates of branch lengths. Because we simulated the data using an ultrametric tree, we expect the tips of the estimated phylogeny to be equidistant from the root if ambiguous characters have no effect. Therefore, a molecular clock test can be used to determine whether relative branch lengths are significantly affected. We first computed the likelihood of the data given the true topology and branch lengths optimized under the JC model. We then computed the likelihood of the data given the true topology and branch lengths optimized with a molecular clock assumption enforced (also under the JC model; the root is assumed to be between the 2 internal nodes). The molecular clock assumption forces the tips to be equidistant from the root. The ratio of these likelihoods was then computed to assess whether any departure from a clock-like evolutionary process was significant (χ^2 test with 2 df; Felsenstein 1981, 1988). The Type I error rate was computed as the proportion of replicates in which the clock model was rejected.

Bayesian Analyses

Bayesian analyses were performed to assess the effect of ambiguous characters on estimates of topological support in the form of bipartition posterior probabilities. Posterior distributions were estimated

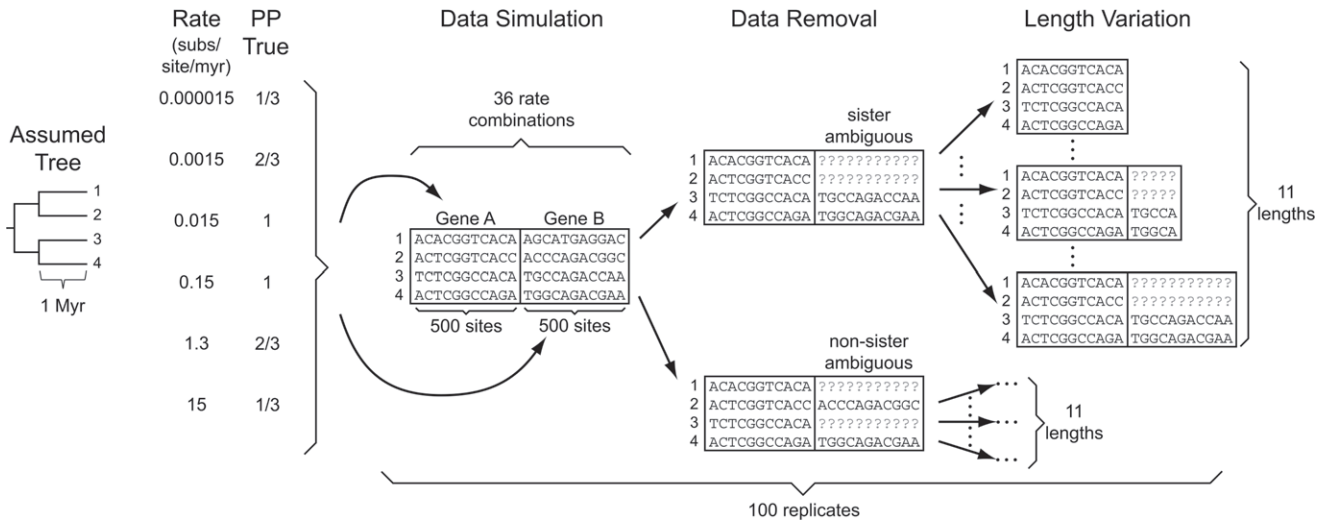

FIGURE 1. Simulation design. Among-site rate variation was simulated using 6 rates of evolution (chosen to produce the desired PP for the true tree with 500 sites) combined across 2 genes to form 36 rate combinations. Gene A contained unambiguous sites, whereas Gene B contained ambiguous sites. Ambiguous characters were present for either sister or nonsister taxa. Although Gene A always contained 500 sites, the length of Gene B varied from 0 to 500 sites. Note that Gene B contained no topological information, regardless of the rate of evolution. PP = posterior probabilities.

using MrBayes v.3.1.1 (Ronquist and Huelsenbeck 2003) with 4 incrementally heated chains (temperature = 0.2). Unless specified otherwise, we assumed the following priors. For the topology, a uniform prior (across all possible resolved topologies) was assumed (the default in MrBayes v.3.1.1). Note that this prior places a zero prior probability on polytomies. For the branch lengths, the exponential prior with mean equal to 0.1 was assumed (the default in MrBayes v.3.1.1). Note that this prior penalizes long branch lengths and requires branches to take lengths greater than 0 (i.e., does not allow polytomies). Markov chains were sampled every 10 generations. MrConverge v1.2b (written by A.R.L.; http://www.evotutor.org/MrConverge) was used to assess burn-in and convergence of 4 independent runs (see Brown and Lemmon [2007] for details).

Each data set was analyzed under the JC model of evolution. In addition, 3 different models of among-site rate variation were considered: gamma-distributed rates with 4 discrete categories (Γ_4) (Steel et al. 1993; Yang 1993, 1994), invariable sites (I) (Gu et al. 1995; Waddell and Penny 1996), and unlinked rates across partitions (P) (Ronquist and Huelsenbeck 2003). The priors assumed for these 3 models of rate variation were uniform(0,50), uniform(0,1), and dirichlet(1,1), respectively. The latter model was used by partitioning the sites according to the gene to which they belong (note that in this case, the partition boundaries are known). In addition, we also enforced a strong prior at the true values for the JC + I and JC + P models to assess the effect of the flat priors on the posterior distribution.

Manipulated Empirical Data

To confirm that the biases caused by ambiguous characters in our simulated data sets could also affect estimates of topology derived from empirical data sets, we manipulated an empirical data set that originally contained very few ambiguous characters. An 8-taxon, single-gene (16S) subset of data was taken from Mueller et al. (2004). To the original data set, we appended up to 1000 additional sites in 2 different schemes. In the first scheme (referred to as sister variable), we randomly chose sites in which the character states of 2 sister species (*Hydromantes italicus* and *Hydromantes brunus*) differed, appended copies of them to the original matrix, and changed the character states of the other 6 species at all appended sites to ambiguous ("?"). In the second scheme (referred to as distant invariable), we randomly chose sites in which the character states of 2 nonsister species (*Desmognathus fucus* and *Ensatina eschscholtzii*) did not differ, appended copies of them to the original matrix, and changed the character states of the other 6 species at all appended sites to ambiguous ("?"). Phylogenetic trees were then inferred from each of these new data sets using ML and Bayesian methods (settings were the same as described above, except that the unpartitioned GTR + I + Γ model was assumed). Because the appended sites in both types of manipulated data sets contain unambiguous characters for only 2 taxa, they should carry no topological information (i.e., their addition should not affect topological support values).

RESULTS

Ambiguous Characters and Branch-Length Priors

Effectively invariable data.—We begin by describing the results from analyses of effectively invariable data sets (i.e., rate = 0.000015 substitutions per site per My). We use the term "effectively invariable" because the

rate of evolution was so low that all simulated data sets were completely invariant at all sites. Here, the JC model is assumed (no rate heterogeneity). In this simple case, we expect the support for each of the 3 possible topologies to be equal, regardless of the length of Gene B or the distribution of ambiguous characters. This expectation is met in the ML framework (Fig. 2*a*). In contrast, the expectation of equal support is not met in the Bayesian framework (Fig. 2*b*). In this case, support for the true topology increases if Gene B is ambiguous for sister taxa, but decreases if Gene B is ambiguous for nonsister taxa. The magnitude of the bias (deviation from the expectation) increases as the length of Gene B increases. When Gene B is ambiguous for none of the taxa (the control), support for the true tree remains at the expected value. One interesting pattern is that the bias caused by the ambiguous data is asymmetric, with the positive bias (Gene B ambiguous for sister taxa) being approximately twice that of the negative bias (Gene B ambiguous for nonsister taxa). This asymmetry is due to the fact that the posterior probability estimate is positively biased for 1 tree and negatively biased for each of the other 2 trees. Because the posterior probabilities of all 3 trees must sum to 1, the magnitude of the bias observed for *each* of the 2 negatively biased trees is less than the magnitude of the bias observed for the single positively biased tree (Supplementary Table S1, http://www.sysbio.oxfordjournals.org).

One possible factor that could lead to the difference between the ML and the Bayesian results (Fig. 2*a*,*b*) is the fact that branch lengths may be collapsed to polytomies in the ML analyses but not in the Bayesian analyses. Branch lengths are not collapsed in the Bayesian analyses because the prior on topologies gives a 1/3 probability to each of the 3 possible (resolved) topologies. This prior places a zero probability on a branch length of 0 (i.e., a polytomy), which can lead to the star tree problem (Suzuki et al. 2002; Cummings et al. 2003; Lewis et al. 2005; Yang and Rannala 2005; Kolaczkowski and Thornton 2006; Steel and Matsen 2007; Yang 2007). Thus, branch lengths are forced to take small but nonzero values. To investigate whether the ability to collapse polytomies could be responsible for the difference between the ML and the Bayesian results, we conducted ML analyses with branch lengths constrained to small but nonzero values. Our results show that this factor indeed drives the misleading posterior probabilities in the Bayesian framework (Fig. 2*c*). As in the Bayesian case, support for the true topology changes with the length of Gene B, and the direction of the change depends on the distribution of ambiguous characters. Note that neither the ambiguous characters nor the prior alone produces substantial bias in topological support. Instead, it is the interaction between ambiguous characters and the prior that produces the bias. In the Supplemental Material (http://www.oxfordjournals.org/our_journals/sysbio/), we present a mathematical argument suggesting that when polytomies are given a zero prior probability (only nonzero branch lengths are allowed), topological

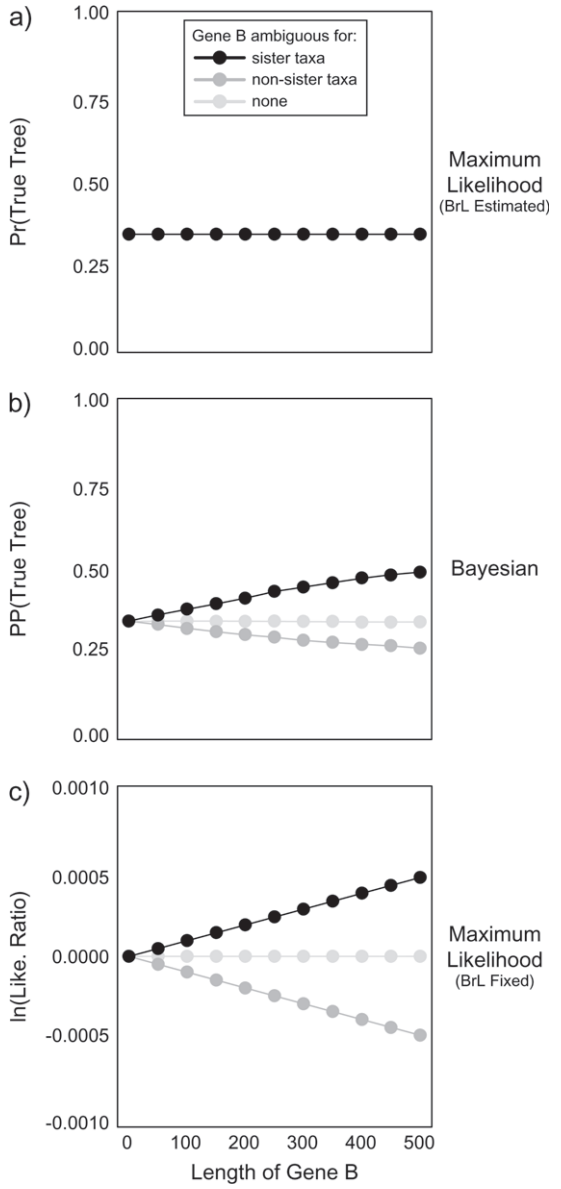

FIGURE 2. The effect of ambiguous characters on topological support when both genes are effectively invariable (rate = 0.000015 substitutions per site per My). On each graph, we plot the support for the true tree as a function of the length of Gene B. Each point represents the mean across 100 replicate data sets. In an ML framework (*a*), ambiguous characters do not affect topological support (Pr: calculated as the proportion of 100 replicates in which the true tree was chosen, with a value of 1/3 given to replicates with equal support across topologies) when branch lengths can be collapsed to polytomies. In a Bayesian framework (*b*), topological support (PP) changes as a function of the length of Gene B and whether Gene B is ambiguous for sister (black) or nonsister (dark gray) taxa. When Gene B is unambiguous (light gray), topological support is unaffected by the length of Gene B. When branch lengths are forced to take a small but nonzero value (10^{-6} substitutions/site) in an ML framework (*c*), ambiguous characters bias topological support (measured as the ratio of likelihood scores for the true to one of the false trees) in the manner seen in the Bayesian framework. Note that in a Bayesian framework, the flat prior on bifurcating topologies requires branch lengths to take a nonzero value. PP = posterior probability; Pr = probability.

support may be biased. Additional studies are needed to determine how efficiently a nonzero prior on polytomies would reduce or eliminate the bias.

Variable data.—Here, we describe the results of analyses of simulated data in which the rate of evolution was sufficient to produce variable sites but remained the same for ambiguous and unambiguous sites. Again, the JC model is assumed (no rate heterogeneity). Recall that the rate of Gene A determines the baseline level of support for the true topology. Gene B should not provide topological information when present in just 2 taxa. If ambiguous characters have no effect on topological support, then we expect support for the true topology to vary systematically with the rate of evolution but not with the length of Gene B. This expectation is met in the ML framework (Fig. 3*a*). Although some stochastic error is present when the sequences have evolved under very high rates, this error would disappear if the number of replicates was increased. In contrast, this expectation is not met in the Bayesian framework (Fig. 3*b*), where topological support changes as a function of the length of Gene B, the distribution of ambiguous characters, and the rate of evolution. The observed bias is highest when the rate of evolution produces either effectively invariable or saturated data and lowest when the rate of evolution is intermediate. No bias is observed when Gene A provides strong support for the true topology (Fig. 3, center 2 columns). Interestingly, the direction of bias is opposite for low and high rates of evolution. When the rate is low, support for the true topology is *positively* biased when sister taxa have ambiguous characters. Conversely, when the rate is high, support for the true topology is *negatively* biased when sister taxa have ambiguous characters.

One factor that could lead to bias in the Bayesian framework when the rate of evolution is high is the branch-length prior. Recall that an exponential branch-length prior (with mean equal to 0.1) was assumed in the Bayesian analyses presented in Figure 3*b*. Under this prior, the density decreases as the length of the branch increases, thus favoring short branches over long branches. If the exponential branch-length prior does not contribute to the bias in topological support, we expect the posterior probability of the true tree to be the same when a different prior is assumed. To determine whether this expectation is met, we estimated the posterior distributions assuming a uniform (flat) prior (0,100) on branch lengths. As expected, the bias disappears under a flat branch-length prior (Fig. 3*c*), suggesting that the combination of the exponential branch-length prior and the ambiguous characters produces the bias. Note that changing the prior has no effect on the bias observed for data sets that evolved under low rates of evolution because zero length branches are currently not allowed under any branch-length prior (as described above).

We also computed the likelihood of the saturated data assuming each of the 3 possible topologies with branch lengths fixed at an arbitrary value that is large (1.0 substitutions/site) but still smaller than the true value (15.0 substitutions/site). This has the effect of mimicking the Bayesian exponential prior, which negatively biases the branch-length estimates. As in the Bayesian analyses, support for the true topology decreases with the length of Gene B when sister taxa have ambiguous characters but increases with the length of Gene B when nonsister taxa have ambiguous characters (Fig. 3*d*). As expected, no bias is present when Gene B is ambiguous for none of the taxa. This result demonstrates that constraining branch lengths to values lower than their optimum in an ML setting has the same effect as assuming a prior that favors short branches in a Bayesian setting.

Ambiguous Characters, Rate Priors, and Model Misspecification

Here, we describe results from the analyses in which the rate of evolution is different for ambiguous and unambiguous sites. In this case, the correct model of evolution is the JC model with separate rates for the 2 genes. We used this model by partitioning the data set by gene (i.e., with known boundary) in both ML and Bayesian analyses. In the Bayesian analysis, the rate prior was set to dirichlet(1,1).

The magnitude and direction of bias in topological support are a function of the relative rates of the ambiguous and unambiguous sites in Bayesian (Fig. 4) but not in ML (Supplementary Fig. S1) analyses. For the Bayesian analyses, substantial bias is observed when the rate of evolution of Gene A is low (Fig. 4, left columns) or when the rates of evolution at both genes are high (Fig. 4, lower right corner). This suggests that weakly supported bipartitions are more sensitive to the effects of ambiguous characters. The rate of evolution of Gene B can affect support for the true topology when the baseline support (from Gene A) is weak. Support for the true topology is strongly biased when Gene B is evolving faster than Gene A. When Gene A is evolving faster, a much smaller bias is typically observed.

The magnitude of the bias caused by ambiguous characters also differs depending on the assumed model of among-site rate variation. This is shown in Figure 5, which presents results from Bayesian analyses of data sets in which Gene A is variable and Gene B is effectively invariable. If rate priors do not interact with ambiguous characters to produce biased topological estimates, then we expect the posterior probability estimates not to vary with the assumed model of rate variation, as long as that model matches the simulation conditions. To test this expectation, we compared results for 3 models of rate variation (Fig. 5): discrete gamma (Γ), invariable sites (I), and partitioned with variable rate prior (P). In principle, the latter 2 models should match the simulation conditions. The direction and magnitude of bias are similar for the discrete gamma and invariable sites

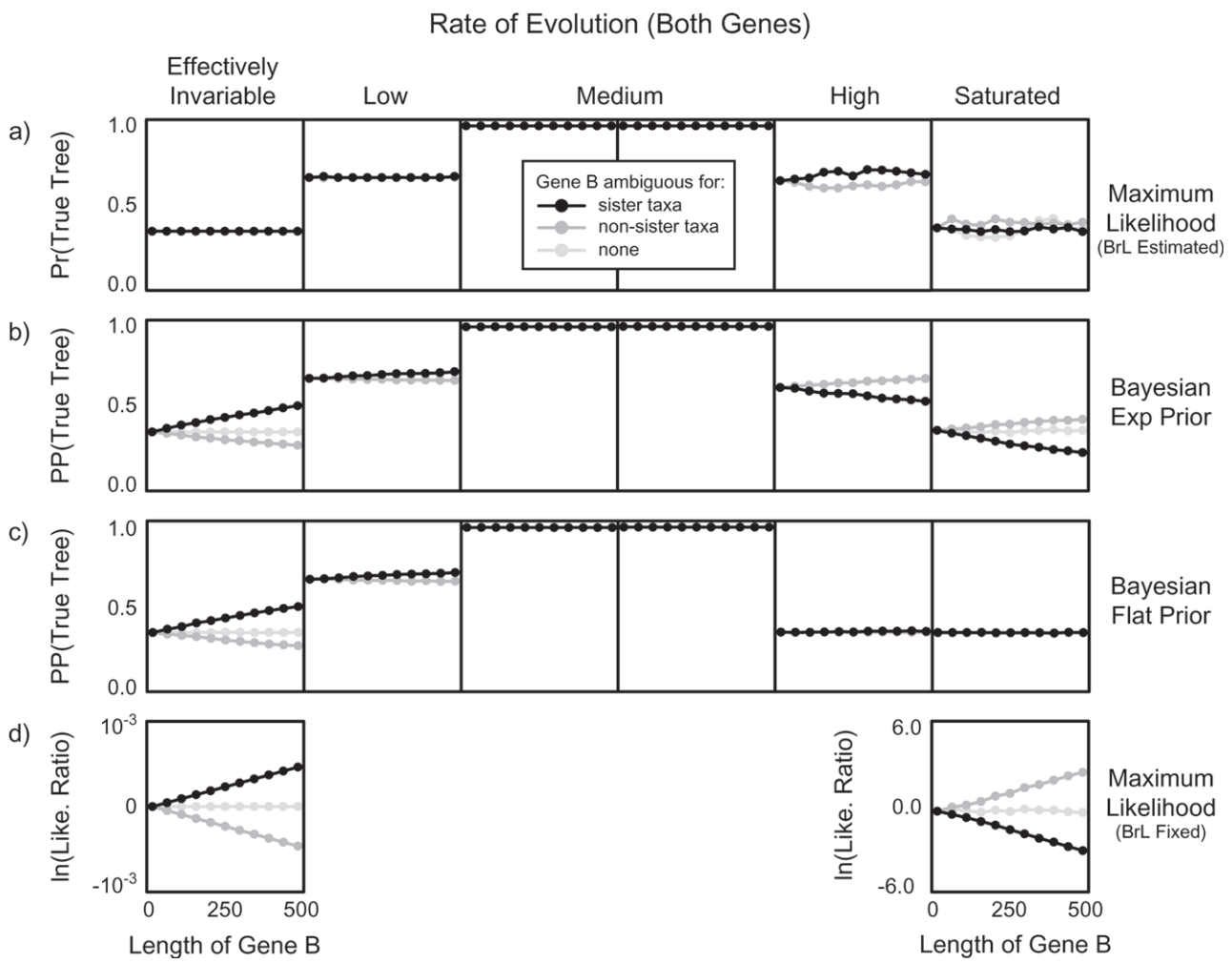

FIGURE 3. The effect of ambiguous characters on topological support when both genes are evolving at the same rate. Axes and shades of gray are the same as in Figure 2. Note that the graphs in the left column of (*a*), (*b*), and (*d*) are identical to those presented in Figure 2. In an ML framework (*a*), ambiguous characters do not lead to a systematic bias in topological support, regardless of the rate of evolution (increasing from left to right columns). In a Bayesian framework (*b*), however, the magnitude and direction of the bias are a function of the rate of evolution. This bias is strongest when the rate of evolution is low or high and weakest when the rate of evolution is intermediate (e.g., when Gene A provides strong support for the true tree). When the rate of evolution is high, the bias exists when an exponential branch-length prior is assumed (*b*) but is absent when a uniform branch-length prior is assumed (*c*). The type of bias seen in the Bayesian framework can be demonstrated in the ML framework (*d*) if branch lengths are fixed at an arbitrarily low value (results for 10^{-6} substitutions per site per My are shown in the lower left graph) or a very high value (results for 1.0 substitutions per site per My shown in the lower right graph) data set. Note that in the Bayesian framework, the flat topological prior prohibits zero-length branches and the exponential branch-length prior penalizes long branches.

models. Surprisingly, the bias is much more substantial for the partitioned model when the rate of Gene A is low but much less substantial when the rate of Gene A is high. Results from the ML analyses largely support these conclusions (Supplementary Fig. S2), although systematic bias was only observed when Gene A was evolving at a high rate.

In order to study the effect of the priors on the rate variation parameters, we performed additional analyses in which we used strong priors to effectively fix parameters at their true values for the invariable sites and partitioned models. The prior on the proportion of invariable sites (uniform from 0 to 1) has a small effect on the bias (compare second and third rows of Fig. 5), despite the fact that the proportion of invariable sites is only accurately estimated when Gene B is ambiguous for none of the taxa (Supplementary Table S2). In contrast, the prior on the relative rates in the partitioned model appears to have a substantial effect (compare the fourth and fifth rows of the left column in Fig. 5). When the prior is set such that strong weight is placed on the true values (i.e., dirichlet(10 000, 10 000)), the bias for effectively invariable data sets (left column) approximates the bias seen when the JC model was assumed (Fig. 2). Because the ratio of rates of evolution is infinity when Gene A is variable and Gene B is effectively invariable, the rate prior could not be fixed at the true values for some of the rate conditions.

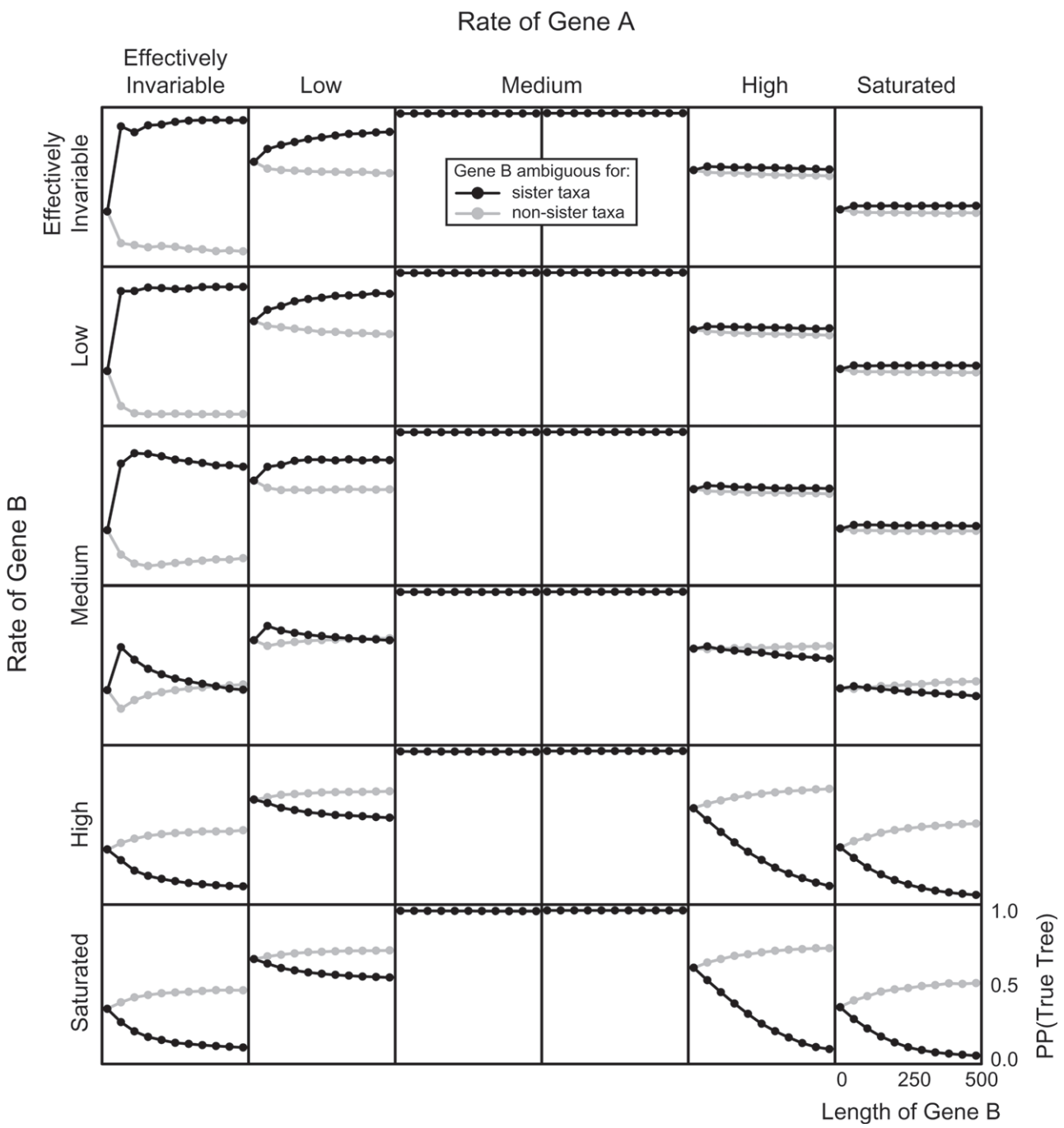

FIGURE 4. The effect of ambiguous characters on Bayesian posterior probabilities when rates differ between unambiguous (A) and ambiguous (B) genes. In each graph, the average posterior probability of the true tree (*y*-axis) is plotted as a function of the length of Gene B (*x*-axis), pattern of ambiguous characters (shade of gray), rate of Gene A (column), and rate of Gene B (row). Graphs show results from analyses in which rate variation was modeled in a partitioned analysis (partitioned by gene) with a dirichlet(1,1) rate prior. Therefore, the model of evolution is overparameterized along the diagonal (equal rates; analogous to Figure 3*b*) and correctly parameterized off the diagonal. Note that the magnitude and direction of bias are a function of the relative rates of the ambiguous and unambiguous genes. Also note that in some cases, the bias is strongest when the number of ambiguous sites is low.

Ambiguous Characters and the Molecular Clock

If ambiguous characters have no effect on branch-length estimates, then we would expect estimated trees to be ultrametric and the Type I error rate for a molecular clock test to be independent of the number of ambiguous sites. Our analyses demonstrate that this is not the case: Ambiguous characters can substantially inflate the Type I error rate for the molecular clock test (Fig. 6). In some cases, the Type I error rate can increase rapidly (from 5% to 100%) with the addition of very

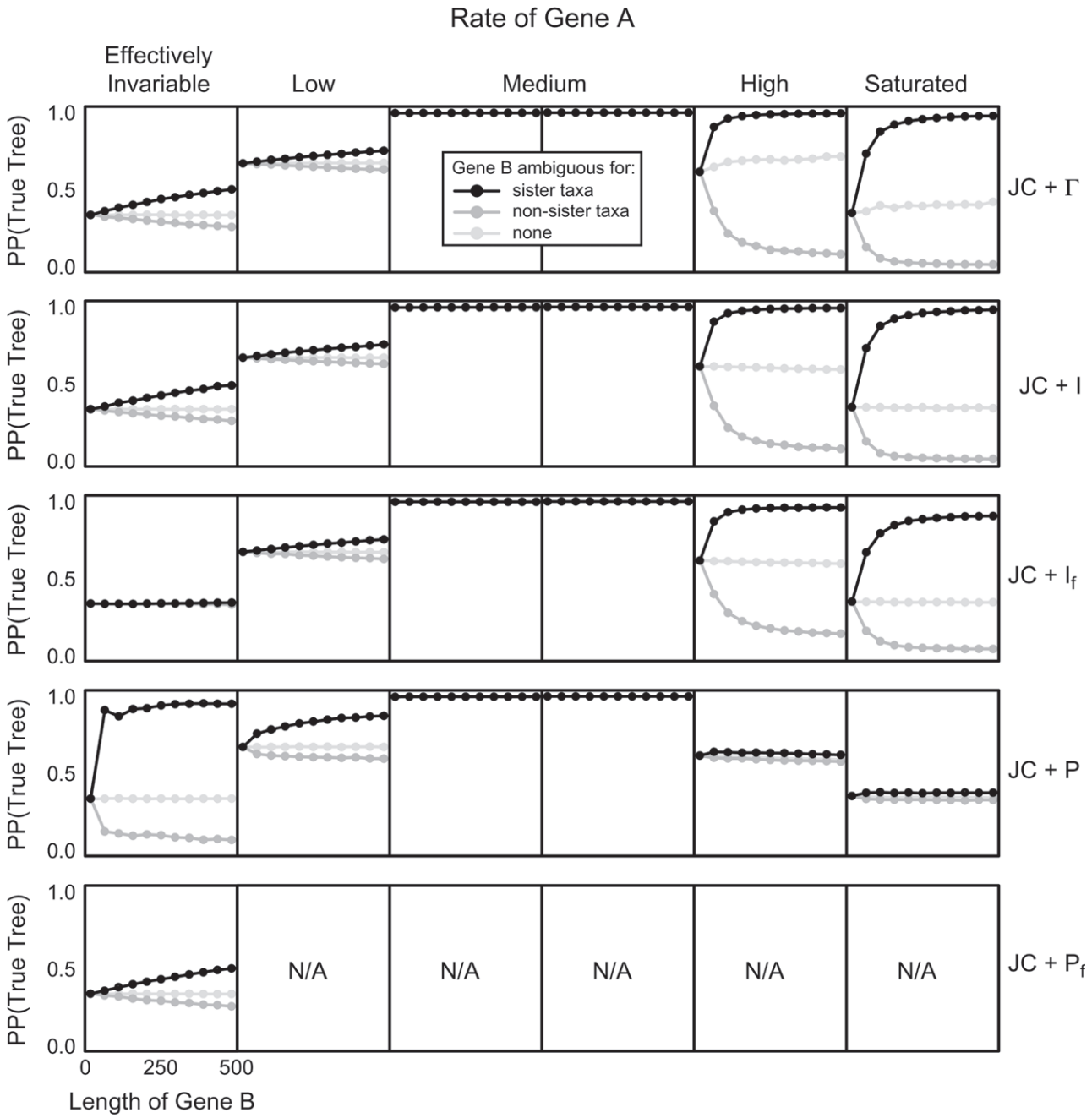

FIGURE 5. The effect of ambiguous characters on Bayesian posterior probabilities under different models of among-site rate variation. Axes and shades of gray have the same meaning as in Figure 2*b*. Each row corresponds to the top row of Figure 4 (Gene B is effectively invariable). Results for models of rate variation are shown: discrete gamma with 4 rate categories (Γ), invariable sites (I), and partitioned with variable rate prior (P). The subscript f indicates that the rate variation parameter was fixed at the true value, removing the effect of the prior on that parameter. Note that in each case, the light gray lines show the results from analyses of data sets in which both Genes A and B were completely unambiguous (i.e., the control). Under the (incorrect) gamma model of rate heterogeneity, the posterior probabilities were slightly biased even for the unambiguous data sets.

few ambiguous sites (e.g., 50). Inflation of the Type I error rate is greatest when Genes A and B are evolving at very different rates. Note, however, that when the 2 genes are evolving under the same rate (graphs along the diagonal in Fig. 6) or when Gene B is unambiguous for all taxa (light gray lines in Fig. 6), the Type I error rate is independent of the length of Gene B. These results suggest that the interaction between ambiguous characters and rate variation among sites can lead to estimates of trees that are significantly nonultrametric. Also note that we are assuming the JC model of evolution (no rate heterogeneity), so the results displayed in the off-diagonal cells of Figure 6 are, in fact, underparameterized. Thus, we cannot say whether the bias

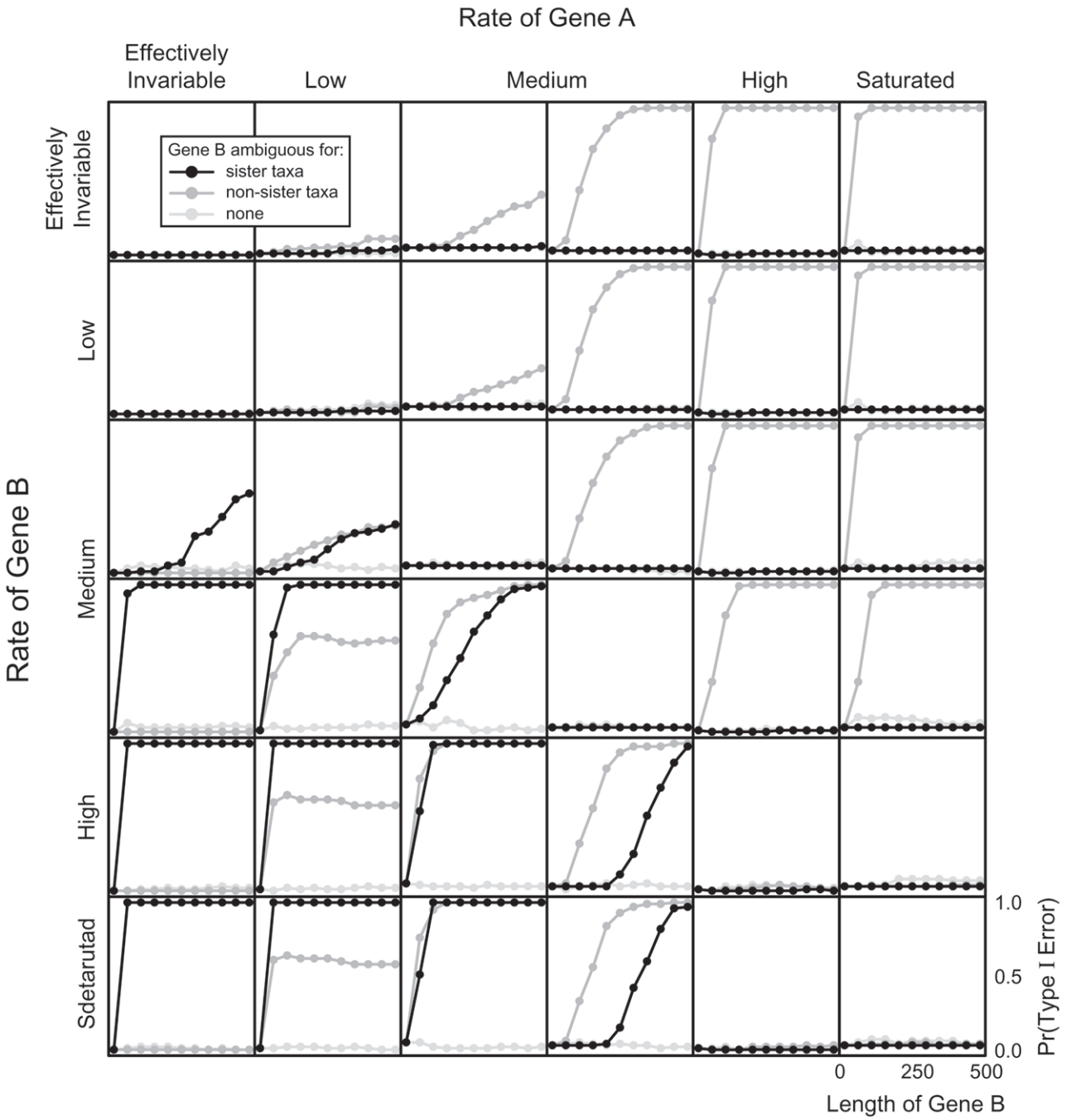

FIGURE 6. The effect of ambiguous characters on the probability of incorrectly rejecting a molecular clock model of evolution in an ML framework. The proportion of 100 replicates in which the clock model was rejected in a χ^2 test (df = 2; *y*-axis) is plotted against the length of Gene B (*x*-axis), distribution of ambiguous characters (shade of gray), the rate of Gene A (columns), and the rate of Gene B (rows). Because rate heterogeneity was not accommodated in these analyses (see text), the model of evolution was underparameterized in analyses presented off the diagonal. Note that substantial inflation of Type I error requires both rate variation (off diagonal) and ambiguous characters (black or dark gray points).

would disappear if rate heterogeneity were properly modeled (though we expect the bias would disappear).

Manipulated Empirical Data

Recall that in both schemes (sister variable and distant invariable), the appended sites contained unambiguous characters for only 2 of the 8 taxa, so topological support should remain the same as sites are appended if ambiguous characters have no effect. Our results clearly demonstrate, however, that ambiguous characters in empirical data sets can strongly bias estimates of topological support and branch lengths (Fig. 7). In particular, when variable sites are added (Fig. 7*a*), sister taxa are

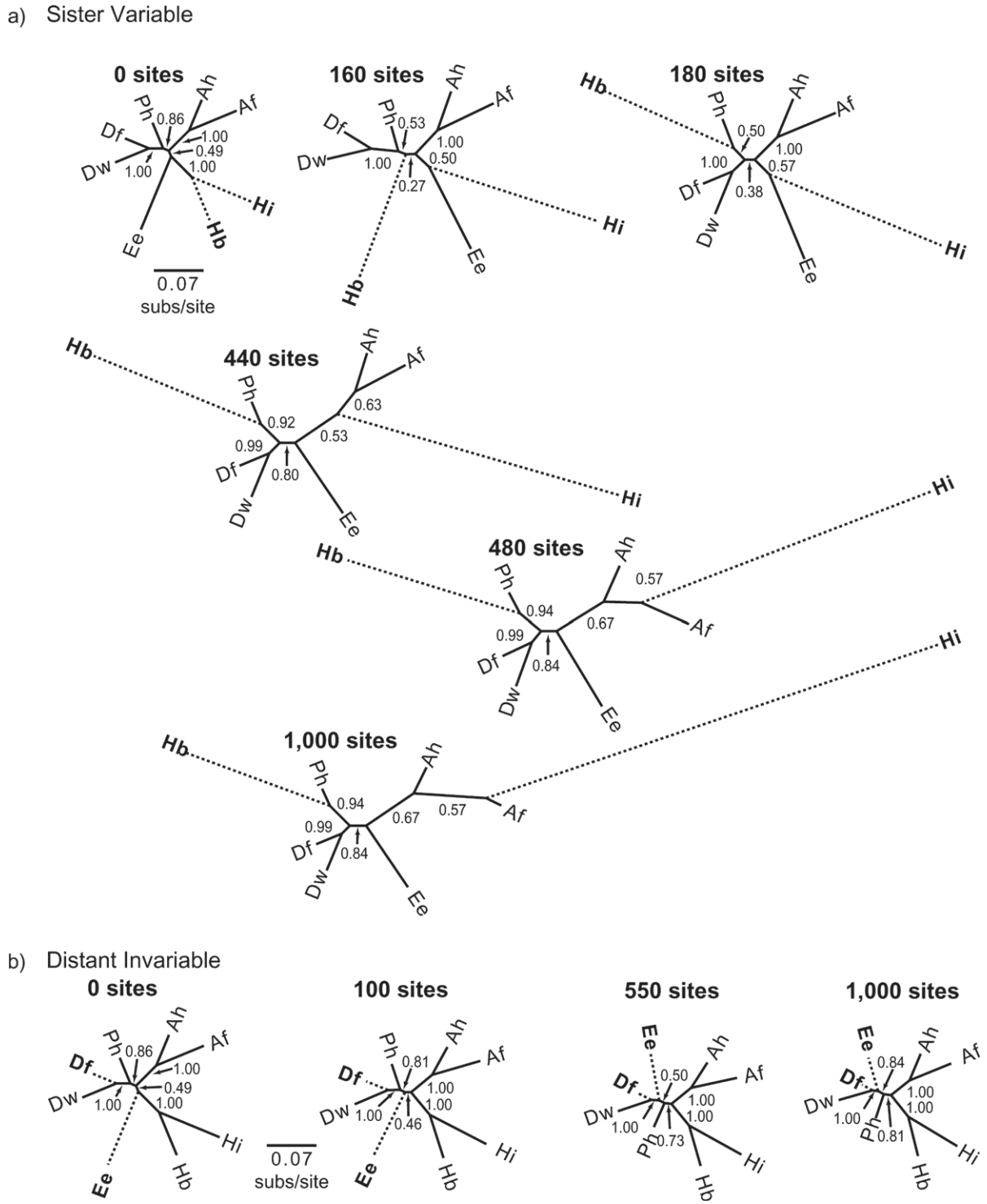

FIGURE 7. The effect of ambiguous characters on estimates of an empirical phylogeny estimated in a Bayesian framework. In (*a*), we present results based on an empirical data set with up to 1000 *variable* sites appended. The character state at each appended site was unambiguous but different for the sister taxa *Hydromantes brunus* (Hb) and *Hydromantes italicus* (Hi) and was ambiguous ("?") for the other 6 taxa: *Aneides flavipunctatus* (Af), *Aneides hardii* (Ah), *Desmognathus fucus* (Df), *Desmognathus wrighti* (Dw), *Ensatina eschscholtzii* (Ee), and *Phaeognathus hubrichti* (Ph). The number of appended sites is given above each phylogeny, and the bipartition posterior probability estimate is given at each internal branch. In (*b*), we present results based on the same empirical data set but with up to 1000 *invariable* sites appended. Here, the character state at each appended site was identical for the distant taxa Df and Ee and was ambiguous for the other 6 taxa: Af, Ah, Dw, Hb, Hi, and Ph. Note that when variable sites are added, taxa with unambiguous characters are pushed apart on the phylogeny, whereas when invariable sites are added, taxa with unambiguous characters are pulled together. Topologies estimated in an ML framework matched those estimated in a Bayesian framework.

pushed apart on the phylogeny. For example, when the data set contains no ambiguous sites, *H. brunus* and *H. italicus* are sister taxa supported by a posterior probability of 1.0. When the data set contains 1000 ambiguous sites, however, these 2 taxa are on opposite sides of the phylogeny (the branches separating them are supported by posterior probabilities of 0.67, 0.84, and 0.94). Conversely, when invariable sites are added (Fig. 7*b*), distant taxa are pulled together. For example, when the data set contains no ambiguous sites, *D. fucus* and *E. eschscholtzii* are separated by 3 branches with posterior probabilities equal to 1.0, 0.86, and 0.49. When the data set contains 1000 ambiguous sites, however, these 2 taxa are only separated by 1 internal branch. In both cases, the result is strong support for bipartitions that do not appear in the topology estimated without the ambiguous sites. Trees inferred using the ML criterion produced the same pattern of bias.

Discussion

Ambiguous characters can strongly bias estimates of topology and branch lengths in ML and Bayesian phylogenetic inference. Gaps due to insertions or deletions will have the same effect unless explicitly modeled (note that most software, including MrBayes, treat gaps as ambiguous characters because explicit models of indels are rarely implemented). We have shown that the magnitude and direction of the bias vary as a function of the number of ambiguous characters, the topological position of ambiguous sequences, the relative rates of evolution for ambiguous and unambiguous sites, and the model of sequence evolution assumed. Furthermore, topological bias is likely to be most pronounced in a Bayesian framework due to the additional interaction between the ambiguous characters and the priors. Even so, estimates of branch length and topology can be biased in an ML framework when rate variation across sites is not properly modeled. These results are in sharp contrast to recent opinions that the effects of ambiguous characters are overstated in the literature (e.g., de Queiroz and Gatesy 2007).

Bipartitions that are strongly supported by unambiguous sites are likely to remain strongly supported with the inclusion of ambiguous sites (e.g., Fig. 4, columns 3 and 4). False bipartitions that would otherwise be weakly supported, however, may become strongly supported with the inclusion of even a few ambiguous sites. In practice, therefore, it may be difficult to distinguish between true bipartitions that are strongly supported by real signal and false bipartitions that are strongly supported because of the effects of ambiguous characters. Note that although we focused our analyses on the 4-taxon case, we expect our conclusions to hold for weakly supported bipartitions in larger phylogenies, although the effects are expected to be more complex due to the interactions with additional bipartitions.

In contrast to several previous simulation studies that attributed a reduced phylogenetic accuracy to a lack of information in the ambiguous sites (leading to low resolution; Wiens 1998, 2003a), our study clearly shows that ambiguous characters actively produce misleading estimates of phylogeny through interaction with 2 other factors: Bayesian priors and model misspecification. Interaction with Bayesian priors can be understood by considering a Bayesian analysis of an invariable 4-taxon data set. Priors on topology (uniform over strictly bifurcating topologies) and branch lengths (typically uniform or exponential) result in sampled branch lengths that are small but nonzero. For a particular site in the data set, the conditional likelihood score is equal to 1.0 for any subtree containing only taxa with ambiguous character states (i.e., "?"). In effect, these portions of the tree are pruned out (ignored). Thus, the site likelihood is calculated only along branches connecting the sequences that are unambiguous for that site. For sites in which 2 of the sequences have unambiguous character states, this score is *not* identical across the 3 topologies. One of the topologies groups these 2 unambiguous taxa as sister, whereas the other 2 topologies position them in a nonsister arrangement. Two branches separate sister taxa, whereas 3 branches separate nonsister taxa. Given that branch lengths are nonzero and the site is invariable (i.e., both taxa with unambiguous characters have the same state), the likelihood under the topology placing the 2 unambiguous taxa as sister is greater than that under the topologies placing them as nonsister (the likelihood is greater when fewer branches separate taxa with the same character state). The priors ensure that only nonzero branch lengths are sampled and thus that the posterior probability of placing the 2 unambiguous taxa as sister is greater than 1/3. This posterior probability increases with an increasing number of such sites. A similar line of reasoning will lead to the opposite conclusion for saturated data sets. This explanation predicts the pattern of topological error seen in our analyses (Fig. 3*b*–*d*) and is confirmed by the mathematical argument shown in the Supplemental Material.

Ambiguous characters can also interact with model misspecification to produce misleading estimates of phylogeny. In order to understand this interaction, consider a 4-taxon data set in which Gene A is evolving at a slower rate than Gene B (refer to Fig. 8). Suppose that a pair of sister sequences contain ambiguous characters for all sites in Gene B. Under this scenario, the lengths of the branches connecting this sister pair will be estimated based only on the sites in Gene A, whereas the lengths of the branches connecting the other sister pair will be estimated based on all the sites (both genes). If among-site rate variation is not properly modeled, the branches connecting the sister pair with ambiguous characters in Gene B will be shorter than those connecting the other sister pair because Gene A is evolving at a slower rate than the average rate across all sites (both genes). As a result of this interaction, rate variation across sites is manifested as rate variation across branches, resulting in biased branch-length estimates (in fact, variation across sites in any model parameter could be manifested as

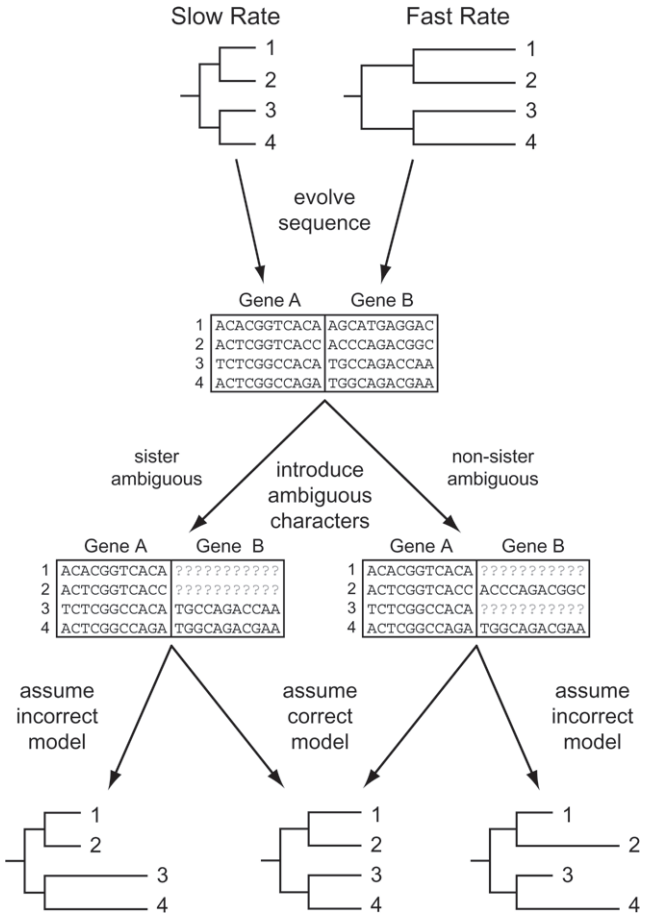

FIGURE 8. Ambiguous characters interact with model misspecification to produce misleading branch-length estimates. A data set is simulated using an ultrametric tree. Some sites evolve under a slow rate, whereas others evolve under a fast rate. Ambiguous characters are introduced nonrandomly with respect to rate and taxon. If rate variation is correctly modeled, the estimated tree is ultrametric. If rate variation is not correctly modeled, the estimated tree is non-ultrametric. The interaction between ambiguous characters and model misspecification causes among-site rate variation to be manifested as among-branch rate variation. Note that the pattern of branch lengths inferred depend on the taxonomic distribution of the ambiguous characters, even though the ambiguous sites contain no topological information.

parameter variation across branches through the same process). Bayesian priors introduce additional factors that may then interact to bias topological support.

The accuracy of many analyses may be jeopardized by the effects of ambiguous characters on branch-length estimates. For example, we have shown that ambiguous characters may increase the propensity to incorrectly reject a molecular clock (Fig. 6). Other branch length–dependent analyses that may be affected include divergence time estimation, ancestral state reconstruction, tree-dependent comparative methods, rate variation analysis, phylogenetic hypothesis testing, and phylogeographic analysis. Future studies are needed to determine the indirect effects of ambiguous characters on the accuracy of each type of analysis.

Many interesting problems stem from the potential for topological bias due to ambiguous characters; we outline several of them here. The rogue taxon problem, wherein 1 taxon is particularly labile in what is otherwise a well-supported tree, could result from the inclusion of ambiguous characters. Related to the rogue taxon problem is conflict among gene trees. Although unambiguous sites might support the same, true topology, genes with a large proportion of ambiguous characters may support alternative topologies due solely to misinterpretation of the ambiguous characters. A researcher could be deceived into believing that multiple phylogenetic signals exist across genes (interpreted as hybridization, horizontal gene transfer, or incomplete lineage sorting), when in fact all support for alternative topologies is due to the presence of ambiguous characters. One of us (K.S.-H.) has come across such an example of discordance among gene trees in empirical data from North American fireflies. Once ambiguous sites were excluded from the analysis, gene tree congruence increased substantially (Stanger-Hall et al. 2007). Last, statistical approaches to phylogenetic hypothesis testing (e.g., Bayesian posterior probabilities and ML bootstrap proportions) may also be rendered inaccurate by this bias. Hypothesis testing is of particular concern because changes in posterior probabilities or bootstrap proportions of only a few percent can alter conclusions of significance, even when the bias is not strong enough to alter the preferred topology.

The results of this study carry serious implications for the practice of combining data when inferring phylogenies, particularly when rates of evolution vary across data sets. For instance, consider the situation in which data are gathered from a large number of species for 2 genes: 1 slower-evolving nuclear gene is included to resolve deep relationships and 1 faster-evolving mitochondrial gene is included to resolve shallow relationships (note that this approach is increasingly common). Due to monetary or time constraints, not all species are sequenced for both genes. Our 4-taxon simulations suggest that the ambiguous characters will cause the taxa sequenced for only the fast gene to be pushed apart on the phylogeny, whereas the taxa sequenced for only the slow gene will be pulled together. Analyses of simulated 8-taxon data sets (Supplementary Fig. S3), as well as a manipulated empirical data set (Fig. 7), confirm these predictions. Supermatrix-style approaches that do not have nearly complete overlap in taxon sampling across data sets will be particularly prone to this type of error.

Although we expect no systematic error if the effects of priors are weak and rate variation across sites is correctly modeled, ensuring these 2 properties may be difficult in practice. For example, the branch-length prior is expected to have strong effects on any branch for which no substitutions have been observed, regardless of the dimensions of the data set. Correct modeling of rate variation across sites may be even more difficult. Ambiguous characters may appear in an alignment for a variety of reasons, such as monetary constraints,

desire to publish quickly, poor alignments, or technical difficulties with sequencing. Given these various causes for the inclusion of ambiguous characters, rates of evolution are unlikely to be discretely different between ambiguous and unambiguous sites. Determining a proper method for modeling rate variation is likely to be extremely difficult, especially as the proportion of ambiguous characters at each site increases. Heterotachy (changes in rates of evolution within a site across the tree), which has already proven problematic with complete data sets (Kolaczkawski and Thornton 2004; Philippe et al. 2005; Spencer et al. 2005; Steel 2005; Lockhart et al. 2006; Matsen and Steel 2007), may also interact with ambiguous characters to produce effects that may be difficult to avoid. One possible effect, for example, is for heterotachy to be manifested as among-site rate variation, thereby biasing estimates of among-site rate heterogeneity parameters.

We have not investigated the effectiveness of particular methods for correcting for the ambiguous character bias, although we suggest several here. The first (and most obvious) solution is to use only completely unambiguous data matrices when inferring phylogenies. To do so, either ambiguous characters should be filled in (through additional sequencing) or ambiguous sites should be removed from the alignment. A second potential solution is to use a technique known as statistical imputation (Kalton and Kish 1981; Ford 1983; David et al. 1986; Little and Rubin 2002; Marker et al. 2002). To impute data, each ambiguous site is filled in using characters from a randomly selected unambiguous site that has the same site pattern as the ambiguous site when cells containing ambiguous characters are ignored. Drawbacks of this approach include the need to account for the uncertainty associated with the filled data and the fact that imputing some sites may be impossible (due to lack of a matching unambiguous site), especially when the matrix contains a large number of sequences or a small number of sites. The third potential solution is to evaluate the effects of ambiguous characters on a data set–specific basis to see if a correction is needed. One approach is to analyze the data set with and without ambiguous sites and look for variation in the results. Note that in many cases, this approach may yield an unclear conclusion because the ambiguous sites could also contain true phylogenetic signal; this is the reason the ambiguous character problem is difficult to study using empirical data. The fourth and final solution we offer is to estimate the ambiguous character states as free parameters. In an ML framework, this would entail identifying the state for each ambiguous character that maximizes the likelihood of observing the unambiguous characters. In a Bayesian framework, a prior would be placed on the distribution of character states and the posterior distribution of character states for each ambiguous character would be estimated. The difficulty with this approach is that the number of parameters that would need to be estimated would be quite large for data sets containing a large number of ambiguous characters. This list of solutions is certainly not exhaustive; we look to future studies to identify the relative merits of various solutions.

We have demonstrated the potential for ambiguous characters to positively mislead ML and Bayesian phylogenetic inference. However, we have not investigated all possible variables that affect the magnitude of this bias (e.g., tree shape), and we leave such analyses for future studies. Much additional work is also needed to identify powerful and robust diagnostics for assessing when ambiguous characters may affect a particular data set as well to determine priors and models that minimize their effect. Until the costs of including ambiguous characters in empirical data sets can be more fully elucidated and methods for eliminating their effects can be developed, extreme caution should be taken when including ambiguous characters or indels in ML or Bayesian phylogenetic analyses.

Supplementary Material

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

Funding

A.R.L. and E.M.L. were supported by National Science Foundation (NSF) IGERT Fellowships in Computational Phylogenetics and Applications to Biology at the University of Texas at Austin (DGE-0114387). J.M.B was supported by an NSF Graduate Research Fellowship. K.S.-H. was supported by the NSF (DEB-0074953).

Acknowledgments

The authors thank David Bryant, Gavin Naylor, and the Computational Phylogenetics Group at the University of Texas at Austin for useful discussions. We are also grateful to Matt Morgan, Tracy Heath, Lars Jermiin, and 2 anonymous reviewers for comments on a previous version of this manuscript.

References

Armbruster W.S. 1992. Phylogeny and the evolution of plant-animal interactions. BioScience. 42:12–20.

Avise J. 2006. Evolutionary pathways in nature: a phylogenetic approach. New York: Cambridge University Press. p. 1–298.

Bowers J.E., Chapman B.A., Paterson A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 422:433–438.

Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.

Bull J.J., Cunningham C.W., Molineux I.J., Badgett M.R., Hillis D.M. 1993. Experimental molecular evolution of bacteriophage T7. Evolution. 47:993–1007.

Bush R.M., Bender C.A., Subbaro K., Cox N.J., Fitch W.M. 1999. Predicting the evolution of human influenza A. Science. 286:1921–1925.

Cummings M.P., Handley S.H., Myers D.S., Reed D.L., Rokas A., Winka K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

David M.H., Little R.J.A., Samuhel M.E., Triest R.K. 1986. Alternative methods for CPS income imputation. J. Am. Stat. Assoc. 81:29–41.
de Queiroz A., Gatesy J. 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22:34–41.
Dragoo J.W., Honeycutt R.L. 1997. Systematics of mustelid-like carnivores. J. Mammal. 78:426–443.
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.
Dunn K.A., McEachran J.D., Honeycutt R.L. 2003. Molecular phylogenetics of myliobatiform fishes (Chondrichthyes: Myliobatiformes), with comments on the effects of missing data on parsimony and likelihood. Mol. Phylogenet. Evol. 27:259–270.
Eickmann M., Becker S., Klenk H.D., Doerr H.W., Stadler K., Censini S., Guidotti S., Masignani V., Scarselli M., Mora M., Donati C., Han J.H., Song H.C., Abrignani S., Covacci A., Rappuoli R. 2003. Phylogeny of the SARS coronavirus. Science. 302:1504–1505.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22:521–565.
Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
Ford B.L. 1983. An overview of hot deck procedures. In: Madow W.G., Olkin I., Rubin D.B., editors. Incomplete data in sample surveys, vol II: theory and annotated bibliographies. New York: Academic Press. p. 185–207.
Gouveia-Oliveira R., Sackett P.W., Pedersen A.G. 2007. MaxAlign: maximizing usable data in an alignment. BMC. Bioinformatics. 8:312.
Gu X., Fu Y.-X., Li W.-H. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.
Harvey P.H., Leigh Brown A.J., Maynard Smith J., Nee S. 1996. New uses for new phylogenies. Oxford: Oxford University Press.
Hillis D.M. 2000. Origins of HIV. Science. 288:1757–1759.
Hillis D.M., Bull J.J., White M.E., Badgett M.R., Molineaux I.J. 1992. Experimental phylogenetics: generation of a known phylogeny. Science. 255:589–592.
Hillis D.M., Moritz C., Mable B.K. 1996. Molecular systematics. Sunderland (MA): Sinauer Associates.
Huelsenbeck J.P. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst. Zool. 40:458–469.
Jermiin L.S., Ho S.Y.W., Ababneh F., Robinson J., Larkum A.W.D. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638–643.
Jobb G. 2008. TreeFinder, version of April 2008. Munich (Germany). Available from: URL www.treefinder.de.
Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Kalton G., Kish L. 1981. Two efficient random imputation procedures. Proc. Survey Res. Methods Sec., Am. Stat. Assoc. 1981: 146–151.
Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst. Biol. 51: 369–381.
Kearney M., Clark J.M. 2003. Problems due to missing data in phylogenetic analyses including fossils: a critical review. J. Vertebr. Paleontol. 23:263–274.
Kolaczkowski B., Thornton J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431:980–984.
Kolaczkowski B., Thornton J.W. 2006. Is there a star tree paradox? Mol. Biol. Evol. 23:1819–1823.
Kuhner M.K., Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.
Lemmon A.R., Moriarty E.C. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53: 265–277.
Lewis P.O., Holder M.T., Holsinger K.E. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.
Little R.J.A., Rubin D.B. 2002. Statistical analysis with missing data, 2nd ed. Hoboken (NJ): Wiley-Interscience.
Lockhart P.J., Larkum A.W.D., Steel M.A., Waddell P.J., Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA. 93:1930–1934.
Lockhart P., Novis P., Milligan B.G., Riden J., Rambaut A., Larkum T. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol. Biol. Evol. 23:40–45.
Marker D.A., Judkins D.R., Winglee M. 2002. Large-scale imputation for complex surveys. In: Groves R.M., Dillman D.A., Eltinge J.L., Little R.J.A., editors. Survey nonresponse. New York: Wiley. p. 329–341.
Matsen F.A., Steel M. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst. Biol. 56:767–775.
McKenna D.D., Farrell B.D. 2005. Molecular phylogenetics and evolution of host plant use in the tropical rolled leaf "hispine" beetle genus *Cephaloleia* (Chevrolat) (Chrysomelidae: Cassidinae). Mol. Phylogenet. Evol. 37:117–131.
Mueller R.L., Macey J.R., Jaekel J., Wake D.B., Boore J.L. 2004. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proc. Natl. Acad. Sci. USA. 101: 13820–13825.
Murphy W.J., Eizirik E., Johnson W.E., Zhang Y.P., Ryder O.A., O'Brien S.J. 2001. Molecular phylogenetics and the origins of placental mammals. Nature. 409:614–618.
Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.
Philippe H., Zhou Y., Brinkmann H., Rodrigue N., Delsuc R. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5:50.
Platnick N.I., Griswold C.E., Coddington J.A. 1991. On missing entries in cladistic analysis. Cladistics. 7:337–343.
Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.
Ronquist R., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Spencer M., Susko E., Roger A.J. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22:1161–1164.
Stanger-Hall K.F., Lloyd J.E., Hillis D.M. 2007. Phylogeny of North American fireflies (Coleoptera: Lampyridae): implications for the evolution of light signals. Mol. Phylogenet. Evol. 45:33–49.
Steel M. 2005. Should phylogenetic models be trying to "fit an elephant"? Trends. Genet. 21:307–309.
Steel M., Matsen F.A. 2007. The Bayesian "star paradox" persists for long finite sequences. Mol. Biol. Evol. 24:1075–1079.
Steel M., Székely P.J.L., Erdös P.L., Waddell P.J. 1993. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. N. Z. J. Bot. 31:289–296.
Sullivan J., Holsinger K.E., Simon C. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol. Biol. Evol. 12:988–1001.
Suzuki Y., Glazko G.V., Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA. 99:16138–16143.
Swofford D.L. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sunderland (MA): Sinauer Associates.
Waddell P., Penny D. 1996. Evolutionary trees of apes and humans from DNA sequences. In: Lock A.J., Peters C.R., editors. Handbook of symbolic evolution. Oxford: Clarendon Press. p. 53–73.
Webb C.O. 2000. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am. Nat. 156: 145–155.
Wiens J.J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625–640.
Wiens J.J. 2003a. Incomplete taxa, incomplete characters and phylogenetic accuracy: is there a missing data problem? J. Vertebr. Paleontol. 23:297–310.
Wiens J.J. 2003b. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.

Wiens J.J. 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39:34–42.
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.
Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–214.
Yang Z. 2007. Fair-balance paradox, star-tree paradox and Bayesian phylogenetics. Mol. Biol. Evol. 24:1639–1655.
Yang Z., Goldman N., Friday A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–324.
Yang Z., Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.

First submitted 8 October 2007; reviews returned 10 January 2008; final acceptance 30 December 2008
Associate Editor: Lars Jermiin