Mistreating birth-death models as priors in phylogenetic analysis compromises our ability to compare models

Michael R. May,[1]* Carl J. Rothfels[1]

[1]University Herbarium and Department of Integrative Biology, University of California, Berkeley,

*To whom correspondence should be addressed; E-mail: mrmay@berkeley.edu

Time-calibrated phylogenetic trees are fundamental to a wide range of evolutionary studies. Typically, these trees are inferred in a Bayesian framework, with the phylogeny itself treated as a parameter with a prior distribution (a "tree prior"). This prior distribution is often a variant of the stochastic birth-death process, which models speciation events, extinction events, and sampling events (of extinct and/or extant lineages). However, the samples produced by this process are observations, so their probability should be viewed as a likelihood rather than a prior probability. We show that treating the samples as part of the prior results in incorrect marginal likelihood estimates and can result in model-comparison approaches disfavoring the best model within a set of candidate models. The ability to correctly compare the fit of competing tree models is critical to accurate phylogenetic estimates, especially of divergence times, and also to studying the processes that govern lineage diversification. We outline potential remedies, and provide guidance for researchers interested in comparing the fit of competing tree models.

Evolutionary inferences that have an explicit temporal component—when did a particular group arise, how do lineages disperse over space and diversify over time, etc.—depend on phylogenetic trees with branch lengths measured in units of time. Rates of evolution and durations of lineages are inherently difficult to distinguish: a short-lived lineage that evolves quickly will experience the same net amount of evolution as a long-lived lineage that evolves slowly. Methods for estimating time-calibrated trees therefore require additional information about rates of evolution or ages of clades; consequently, these methods are predominantly Bayesian (*1–3*), and use priors on rates of evolution (a "clock prior") and the durations of lineages (a "tree prior") to disentangle rate and time. The tree prior often takes the form of a stochastic birth-death process (*4, 5*), which is a model with parameters that govern the frequency of birth (speciation or lineage splitting) and death (extinction) events. Recently developed extensions to this class of model—*e.g.*, the "serially sampled" or "fossilized" birth-death processes (*6–9*)—also

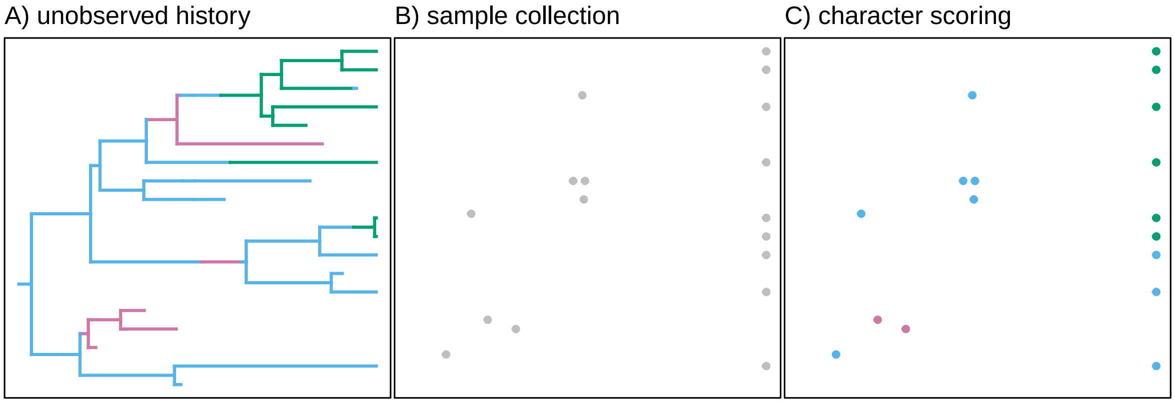

Figure 1: **The phylogenetic process and data.** A) We imagine that there is a true—but unobserved—history for our study group. This unobserved history is the outcome of random processes of lineage diversification (which gives rise to the tree topology and branch lengths) and character evolution (which assigns colors to the branches). Inferring this unobserved history—and the processes that generated it—is the focus of a phylogenetic analysis. B) We collect samples of extant and/or extinct members of our study group (grey dots). C) We score each sample for character data (color of each dot), and use that data to estimate the phylogeny of our study group. Traditionally, we treat the probability of the character data (C) evolving under a particular model of character evolution as the likelihood function, but the probability of the samples and their occurrence times (B) as part of the prior probability of the tree.

model sampling events over time, and allow researchers to estimate time-calibrated trees for mixed samples of extant and extinct (or otherwise non-contemporaneous) lineages. Variants of these processes allow birth, death, and sampling rates to vary over time (*10*), across the branches of the tree (*11, 12*), or as a function of an evolving trait (*13*) or geographic region (*14*). These models have the potential to reconcile major discrepancies between phylogenetic and paleontological estimates of the origin times of lineages (*15, 16*), and allow researchers to study epidemiological dynamics as they unfold over time and space (*17, 18*).

However, treating the tree model as a prior is conceptually problematic. Phylogenetic datasets are the outcome of a hierarchy of processes. First, the true (but unobserved) history of the group unfolds according to a process of lineage diversification and character evolution (Fig. 1A); this history depends on birth and death parameters of the birth-death process, but not the sampling parameters. Next, a researcher collects samples of extant and/or extinct lineages according to a random process governed by the sampling parameters (Fig. 1B). Finally, in order to estimate the phylogenetic relationships among the samples, a researcher may record the character data for each sample (Fig. 1C). The traditional phylogenetic treatment implicitly assumes that the character data generated in the third step (the morphological data, or molecular sequence data) are observations, and their probability is assigned to the likelihood of the model. In contrast, the samples generated in the second step are not considered observations;

rather, they are considered part of the (unobserved) tree, and their probability is assigned to the prior probability of the model. However, the samples themselves are observations that carry information about the underlying diversification process; indeed, paleontologists regularly treat these samples as observations, and use them to infer diversification rates (*19, 20*). It is more conceptually coherent to view the birth-death process as a *model* that produces observations (*i.e.*, the samples) and an unobserved tree, just as the model of character evolution produces observations (the character states of the samples) as well as an unobserved character history (colored branches, Fig. 1A).

In maximum-likelihood inference, there is a strong distinction between samples (which are random variables that are the outcome of a model, and thus have a probability) and parameters (which are fixed but unobserved variables, and thus have no probability). In phylogenetic inference, however, there is occasional disagreement as to whether, for example, ancestral character states are parameters (*21*) or random variables (*22*), or whether the phylogeny itself is a parameter (*23, 24*) (Supplementary Text, section "Samples versus parameters in phylogenetics"). In Bayesian inference, all variables have probability distributions, so the labeling of samples (and thus likelihoods) and parameters (and thus priors) is often innocuous. In particular, this labeling does not impact estimates of posterior distributions inferred by Markov chain Monte Carlo (MCMC). However, the marginal likelihood of a Bayesian model (the denominator in Bayes' theorem), which is the likelihood of the model averaged over all possible parameter values in proportion to their prior probability, does depend on which parts of the model are assigned to the likelihood and which are assigned to the prior. Marginal likelihoods are fundamental to one of the primary methods for comparing the fit of competing models, Bayes factors (*25, 26*). Model comparison, in turn, is critical for inferring accurate topologies and divergence times (*27–30*), and for making macroevolutionary inferences about how and why rates of lineage diversification vary across the tree or over time (*31*). In particular, because of the inherent interaction of rate and time (*30, 32*), divergence-time estimates are extremely sensitive to the tree model (*27–30*), and thus the ability to compare tree models is especially critical in this context.

Bayes factors can be calculated in two theoretically equivalent ways: the first, and most commonly used, is to compute the marginal likelihoods of each model (estimated using a special algorithm, *e.g*, the stepping-stone [SS] sampler (*35*)) and take the ratio of the two marginal likelihoods, and the second is to divide the ratio of each model's posterior probability (typically estimated using reversible-jump [RJ] MCMC (*36*)) by the corresponding prior probability ratio. For models of molecular evolution, the two approaches for computing Bayes factors produce the same value, as expected (Fig. 2A; Supplementary Text, section "Simulation Study"). However, when we compare the fit of competing birth-death models, the two approaches disagree (Fig. 2B,C). This error is consistent with the marginal likelihoods being calculated incorrectly for tree models and is not conservative—Bayes factors computed using marginal likelihoods are not simply lower in magnitude than those computed using posterior model probabilities—and can result in Bayes factors favoring the wrong model (the model that would be disfavored if the Bayes factors were calculated correctly; Fig. 2, shaded regions).

Intuitively, Bayes factors based on the marginal likelihood measure the evidence in the char-

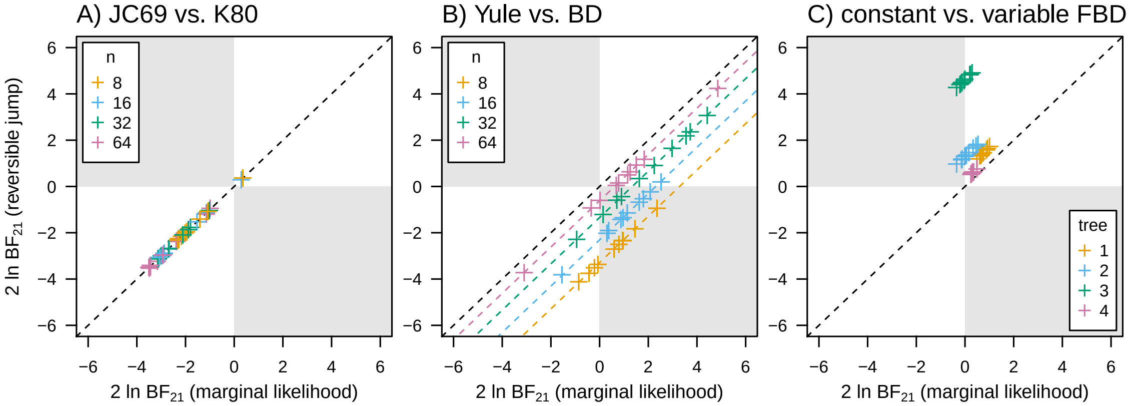

Figure 2: **Bayes factor discrepancies when comparing tree models.** A) The fit of two models of molecular evolution—JC69 (*33*) (the true model) and K80 (*34*)—to simulated datasets of n contemporaneous samples calculated with Bayes factors (BFs) computed using a stepping-stone sampler (SS, based on marginal likelihoods; (*35*)) and reversible-jump MCMC (RJ, based on posterior model probabilities; (*36*)). Positive values indicate support for the second model, and negative values indicate support for the first model. The BFs computed by these two methods are the same. B) The fit of two birth-death processes—the Yule model (with no extinction rate parameter) and the standard birth-death (BD) model (the true model)—to the datasets from panel A. There is a constant discrepancy between the two BFs that depends on the number of lineages (colored points), which is equal to the marginal likelihood ratio of the samples under the two models (dashed lines). This discrepancy can result in the two methods preferring different models (shaded regions). C) The fit of two fossilized birth-death models—one with a constant fossilization rate and another with a fossilization rate that varies over time (the true model)—to simulated datasets of non-contemporaneous samples. The resulting BFs demonstrate the same constant discrepancy as in the contemporaneous case, though in this case the ratio of the marginal likelihood of the samples is effectively impossible to calculate.

acter data for the competing birth-death models, but ignore the evidence from the samples. By contrast, Bayes factors based on posterior model probabilities extract all the relevant evidence from the samples about the relative fit of the birth-death models. Theory indicates that the discrepancy between Bayes factors calculated with these two approaches is equal to the ratio of the marginal likelihoods of the samples under the two models (Supplementary Text, section “Sequential Bayesian Analysis”). While this quantity can be computed for birth-death processes for contemporaneous samples (Fig. 2B, dashed lines), analytical solutions are currently unavailable for processes that generate non-contemporaneous samples (Fig. 2C).

To provide an example of the impact of this error on empirical inferences, we reanalyzed data from our study of marattialean ferns (*30*). Our original study supported the surprising conclusion that there was no evidence for fossilization rates varying over time, despite the fact that the majority of our fossil specimens were drawn from a narrow time interval (the Pennsylvanian). In our re-analysis (Supplementary Text, section "Empirical Analysis"), marginal likelihoods favor a model with constant fossilization rates over one that allows fossilization rates to vary ($2 \ln \mathrm{BF} \approx 3$), while RJ MCMC very strongly favors a model that allows fossilization rates to peak in the Pennsylvanian ($2 \ln \mathrm{BF} \approx 18$ in favor of the variable-rate model). The choice of model results in large (up to 15 My) differences in divergence-time estimates, as well as significant differences in underlying macroevolutionary inferences. Under the constant model, for example, we infer that there were about 400 (95% credible interval [36,4789]) marattialean lineages during the Pennsylvanian, which drops by nearly half—to 225 (95% CI [25,2757]) lineages—under the variable model: the constant model struggles to produce the observed number of fossils in this epoch without implying a very large number of unsampled lineages.

The most straightforward existing solution to this problem is to compare tree models using RJ MCMC. Although this approach has the additional advantage of averaging estimates of the tree (and all other parameters) over the uncertainty in the particular tree model, these algorithms need to be tailored to specific sets of models, which constrains the types of models we can compare. A more satisfying, long-term solution for comparing models would be to develop the theoretical and computational machinery to separate the probability of the samples from the probability of the tree (Supplementary Text, section "Factorizing Bayes' Theorem"). While this factorization is relatively straightforward for simple birth-death models that generate contemporaneous samples, it is currently analytically intractable for models that generate non-contemporaneous samples, especially when the exact ages of the samples are uncertain (as is often the case with fossils).

Acknowledging samples as data in a phylogenetic analysis will have significant benefits for theoreticians and empiricists alike. For theoreticians, it opens the door to a suite of statistical tools for assessing absolute model fit using posterior-predictive simulation (PPS) (*37–39*), which evaluates how well a model predicts observed data. As an absolute measure of model fit (as opposed to the relative measure of fit provided by Bayes factors), PPS helps not only to establish confidence (or skepticism) in particular estimates, but also to identify specific weaknesses in existing models, which enables further model development. In this case, posterior-predictive simulation would allow us to assess the model's ability to predict the number and temporal distribution of the samples themselves (Supplemental Text, section "Posterior-Predictive Simulation"). For empiricists, treating samples as data will enable them to robustly assess the relative and absolute performance of tree models using Bayes factors and posterior-predictive simulation. In turn, this will improve estimates of phylogeny and divergence times, allow for the accurate comparison of macroevolutionary processes, and consequently facilitate countless evolutionary and epidemiological inferences.

References

1. Z. Yang, *Molecular Evolution: A Statistical Approach* (Oxford University Press, 2014).

2. T. A. Heath, B. R. Moore, *Bayesian Phylogenetics: Methods, Algorithms, and Applications* pp. 277–318 (2014).
3. M. dos Reis, P. C. Donoghue, Z. Yang, *Nature Reviews Genetics* **17**, 71 (2016).
4. D. G. Kendall, *The Annals of Mathematical Statistics* **19**, 1 (1948).
5. Z. Yang, B. Rannala, *Molecular Biology and Evolution* **14**, 717 (1997).
6. T. Stadler, *Journal of Theoretical Biology* **267**, 396 (2010).
7. A. Gavryushkina, D. Welch, T. Stadler, A. J. Drummond, *PLoS Computational Biology* **10**, e1003919 (2014).
8. T. A. Heath, J. P. Huelsenbeck, T. Stadler, *Proceedings of the National Academy of Sciences, USA* **111**, E2957 (2014).
9. C. Zhang, T. Stadler, S. Klopfstein, T. A. Heath, F. Ronquist, *Systematic Biology* **65**, 228 (2016).
10. T. Stadler, *Proceedings of the National Academy of Sciences* **108**, 6187 (2011).
11. O. Maliet, F. Hartig, H. Morlon, *Nature Ecology & Evolution* **3**, 1086 (2019).
12. S. Höhna, *et al.*, *bioRxiv* p. 555805 (2019).
13. W. P. Maddison, P. E. Midford, S. P. Otto, *Systematic Biology* **56**, 701 (2007).
14. E. E. Goldberg, L. T. Lancaster, R. H. Ree, *Systematic Biology* **60**, 451 (2011).
15. F. Ronquist, N. Lartillot, M. J. Phillips, *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150136 (2016).
16. C. R. Marshall, *Frontiers in Genetics* **10** (2019).
17. N. F. Müller, *et al.*, *Science Translational Medicine* (2021).
18. S. A. Nadeau, T. G. Vaughan, J. Scire, J. S. Huisman, T. Stadler, *Proceedings of the National Academy of Sciences* **118** (2021).
19. M. Foote, *Paleobiology* **27**, 602 (2001).
20. D. Silvestro, J. Schnitzler, L. H. Liow, A. Antonelli, N. Salamin, *Systematic Biology* **63**, 349 (2014).
21. M. Pagel, *Systematic Biology* **48**, 612 (1999).
22. Z. Yang, S. Kumar, M. Nei, *Genetics* **141**, 1641 (1995).

23. A. W. Edwards, *Journal of the Royal Statistical Society: Series B (Methodological)* **32**, 155 (1970).

24. B. Rannala, Z. Yang, *Journal of Molecular Evolution* **43**, 304 (1996).

25. H. Jeffreys, *Mathematical Proceedings of the Cambridge Philosophical Society* (Cambridge University Press, 1935), vol. 31, pp. 203–222.

26. R. E. Kass, A. E. Raftery, *Journal of the American Statistical Association* **90**, 773 (1995).

27. F. L. Condamine, N. S. Nagalingum, C. R. Marshall, H. Morlon, *BMC Evolutionary Biology* **15**, 5134 (2015).

28. M. S. Y. Lee, A. M. Yates, *Proceedings of the Royal Society B: Biological Sciences* **285**, 20181071 (2018).

29. A. M. Wright, P. Wagner, D. Wright, *EcoEvoRxiv* (2020).

30. M. R. May, *et al.*, *Systematic Biology* (2021).

31. H. Morlon, *Ecology Letters* **17**, 508 (2014).

32. M. dos Reis, Z. Yang, *Journal of Systematics and Evolution* **51**, 30 (2012).

33. T. H. Jukes, C. R. Cantor, *Mammalian Protein Metabolism* **3**, 21 (1969).

34. M. Kimura, *Journal of Molecular Evolution* **16**, 111 (1980).

35. W. Xie, P. O. Lewis, Y. Fan, L. Kuo, M.-H. Chen, *Systematic Biology* **60**, 150 (2011).

36. P. J. Green, *Biometrika* **82**, 711 (1995).

37. A. Gelman, X.-L. Meng, H. Stern, *Statistica Sinica* pp. 733–760 (1996).

38. J. P. Bollback, *Molecular Biology and Evolution* **19**, 1171 (2002).

39. J. M. Brown, *Systematic Biology* **63**, 289 (2014).

Acknowledgments

We thank members of the Rothfels lab, Brian Moore, Carrie M. Tribble, Ixchel González Ramírez, Jenna T. B. Ekwealor, Ammon Thompson, Jiansi Gao, and the editor and X anonymous reviewers for providing valuable feedback on this manuscript. **Funding:** This research was supported by the National Science Foundation (NSF) grant DEB-1754705 to CJR and Cindy V. Looy. This research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer). **Author contributions:** MRM and CJR conceived the research; MRM performed the simulations and analyses; MRM and CJR wrote the manuscript. Both authors read and approved the final manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Supplementary scripts and data can be found in the Zenodo repository `http://doi.org/10.5281/zenodo.5072533` and the GitHub repository `https://github.com/mikeryanmay/bd_bayes_factors/releases/tag/initial_submission`.

Supplementary materials

Materials and Methods (Sections S1–S2)
Supplementary Text (Sections S3–S7)
Figs. S1–S8
Tables S1–S2
References *(40–55)*