Propagating uncertainty about molecular evolution models and prior distributions to phylogenetic trees

David R. Bickel

Informatics and Analytics

University of North Carolina at Greensboro

The Graduate School

241 Mossman Building, CAMPUS

Greensboro, NC 27402-6170

USA

dbickel@uncg.edu

Abstract

Phylogenetic trees constructed from molecular sequence data rely on largely arbitrary assumptions about the substitution model, the distribution of substitution rates across sites, the version of the molecular clock, and, in the case of Bayesian inference, the prior distribution. Those assumptions affect results reported in the form of clade probabilities and error bars on divergence times and substitution rates. Overlooking the uncertainty in the assumptions leads to overly confident conclusions in the form of inflated clade probabilities and short confidence intervals or credible intervals.

This paper demonstrates how to propagate that uncertainty by combining the models considered along with all of their assumptions, including their prior distributions. The combined models incorporate much more of the uncertainty than Bayesian model averages since the latter tend to settle on a single model due to the higher-level assumption that one of the models is true. Nucleotide sequence data illustrates the proposed model combination method.

Keywords: bootstrap; confidence interval; credible interval; molecular phylogenetics; uncertainty propagation; uncertainty quantification

1 Introduction

A state of crisis has been declared in many fields of science. To an alarming degree, results once considered established have failed to replicate when studies in psychology, neuroscience, and biomedicine were repeated. A major cause of that *replication crisis* is the practice of analyzing the same data using multiple statistical models and then settling on a model that achieves a p-value less than 0.05. The finding is then published as statistically significant without mentioning the process of model selection that in many cases ensured that result (Bausell, 2021).

An increasingly popular solution is to register the methods of analysis before data are collected, but that solution is problematic for observational studies as opposed to controlled experiments such as clinical trials (Simmons et al., 2021). In the case of inferring phylogenetic trees and divergence times, the fossil and molecular data are typically available before methods of estimation are applied. For that reason, estimated phylogenetic trees and divergence times are only hypotheses or possibility explanations that are as tentative as the results from other historical sciences (Bromham, 2019). Indeed, the fact that geological and evolutionary history cannot be repeated means that, unlike psychology (Hantula, 2019) and other experimental fields, evolutionary biology will not have the benefit of a replication crisis as a symptom that the certainty of many reported findings is exaggerated due to the flexibility in method choice.

The ability of researchers to select arbitrary data analysis methods in order to make results appear more certain is already a recognized problem in evolutionary biology (Nakagawa et al., 2021; Sun and Zhang, 2021). The problem occurs in molecular phylogenetics whenever biologists select Bayesian methods because they make phylogenies appear more certain; that practice is widespread according to Bromham (2016, p. 431). But even after deciding whether or not to use a Bayesian method or a frequentist method, the many largely arbitrary choices of substitution models, clock models, and distributions of rates across sites leave plenty of flexibility for reporting narrower confidence intervals or higher probabilities to make a story more interesting and a paper more publishable.

It would be more objective to make those choices using the Akaike information criterion (AIC), Bayesian information criterion (BIC), or another a mathematical method of model selection. However, reporting phylogenetic trees or confidence intervals of divergence times only on the basis of the best (e.g., lowest-AIC) model, as if that model were known to be true, also suppress uncertainty.

The traditional Bayesian alternative to selecting a single model is to average models with respect to their posterior probabilities (Li and Drummond, 2012; Wu et al., 2013; Bouckaert and Drummond,

2017). In this approach, the term *model* is used broadly enough to include all possible data-generating processes and a joint prior distribution over those processes. By decision theory, Bayesian model averaging would be ideal if the set of models to be averaged contained the true, data-generating model and if the prior probabilities of the models were known. But since neither is the case, Bayesian model averaging leads, given enough data, to misleadingly assigning practically 100% of the posterior probability to a single model as if it were known to be true (e.g., Cerquides and de Mántaras, 2005; Le and Clarke, 2017; Yao et al., 2018). That gives the same result as selecting the single model, again suppressing uncertainty about the model (Kittler et al., 1998).

That is exactly what is seen in molecular phylogenetics. For example, Barido-Sottani et al. (2018) explain that since the Bayes factor comparing two clock models is $e^{8.8}$ for hepatitis B virus data, there is "overwhelming support" favoring a relaxed clock model over a strict clock model. In fact, the posterior probability for the strict clock model would only be

$$\frac{1}{1+e^{8.8}} \approx \frac{1}{1+6700} \approx 1.5 \times 10^{-4}$$

by Bayes's theorem, conditional on the truth of one of those models and assuming they have equal prior probabilities. With a probability that close to zero, any credible interval and posterior probabilities computed from Bayesian model averaging would be the same as those from the relaxed clock alone. Since, under, general conditions (e.g., Claeskens and Hjort, 2008), a difference between the BIC scores of two models is approximately the difference in their marginal likelihoods (MLs) and since the Bayes factor between two models is approximately $e^{\text{difference in MLs}}$ (Barido-Sottani et al., 2018), the near-100% concentration of posterior probability in one model is seen whenever BIC scores or MLs differ between the best-performing model and other models by more than 5 or so (cf. Kass and Raftery, 1995). That is more the rule than the exception (e.g., Baele et al., 2012, Tables 1-2, Figures 1, 3; Duchêne et al., 2020, Fig. 2). Since Bayesian model averaging tends to give the same results as Bayesian model selection, it fails to combine the strengths of the phylogenetic models considered, as noted above for other domains.

Section 2 addresses the problem by providing methods that combine the results of prior distributions and other model assumptions made in molecular evolution studies. The methods are applied in Section 3 to bacterial sequence data in order to address uncertainty in whether to use Bayesian methods of estimating clade probabilities and in order to combine credible intervals across models.

2 Theory: Proposed methods

2.1 Combining probabilities about clades or divergence times

How can the uncertainty about model assumptions be propagated to a probability about an aspect of a phylogenetic tree? For example, if a maximum likelihood model says the probability of a clade is 87% and a Bayesian model says it is 100%, how can the uncertainty about the model by propagated to the probability of the clade?

Fortunately, there are a number of methods of doing that by model combination, called *ensemble learning* in the machine learning literature and *opinion pooling* in the statistics and decision making literature. Some of the simplest methods involve taking the arithmetic mean or the geometric mean (normalized to give 100% as the total probability) of the distributions to be combined (Genest and Zidek, 1986; Kittler et al., 1998). Given m models, the proposed methods are based on combining posterior probabilities $p_1, p_2, \ldots, p_m$ by their weighted arithmetic mean:

$$\overline{p} = w_1\, p_1 + w_2\, p_2 + \cdots + w_m\, p_m,$$

where $w_1, w_2, \ldots, w_m$ are positive weights that add up to 1. If the models are weighted equally, the combined probability is the usual average:

$$\begin{aligned}\overline{p} &= \frac{1}{m}\, p_1 + \frac{1}{m}\, p_2 + \cdots + \frac{1}{m}\, p_m \\ &= \frac{p_1 + p_2 + \cdots + p_m}{m}.\end{aligned}$$

Similarly, the normalized weighted geometric mean is

$$\overline{p} \propto p_1^{w_1} \times p_2^{w_2} \times \cdots \times p_m^{w_m},$$

reducing in the case of equal weights to

$$\overline{p} \propto \left(p_1 \times p_2 \times \cdots \times p_m\right)^{\frac{1}{m}},$$

with the constant of proportionality set to ensure that the normalized weighted geometric means add up to 1.

But why favor the arithmetic mean over the normalized geometric mean and the wealth of other ways to combine probability distributions? An important reason is that the arithmetic mean

	$t < 100$ MYA	$t = 150 \pm 50$ MYA	$t < 200$ MYA	Total probability
Model 1	1%	50%	49%	100%
Model 2	49%	50%	1%	100%
Arithmetic mean	25%	50%	25%	100%
Geometric mean	7%	50%	7%	**64%**
Normalized geometric mean	11%	**78%**	11%	100%

Table 1: Posterior probabilities of divergence times, highlighting an advantage of the arithmetic mean as the method of combining probabilities from different models. The probabilities in **boldface** are problematic since probabilities should add up to 100% and should agree with those of the models when the models agree on a probability assignment.

ensures that the combined probability is the same as that assigned by each of the models whenever all models agree on that probability (Stone, 1961; Cooke, 1991, p. 173).

For example, consider these three hypotheses about a divergence time: it is less than 100 MYA, it is between 100 MYA and 200 MYA, and it is greater than 100 MYA. Two models agreeing that the probability of the second hypothesis is 50% should yield a combined probability of 50%, but that is not necessarily the case unless the arithmetic mean is used. That is seen in Table 1: whereas the geometric mean would also give a combined probability of 50%, its total probability is only **64%** (bolded as in the table). That is remedied by normalization to 100%, but then the combined probability that the divergence time is 150 ± 50 MYA is **78%** even though the models agree that it is only 50%.

The same considerations apply to bootstrap proportions since they may be interpreted as conservative estimates of posterior probabilities in spite of their non-Bayesian origins (Bickel, 2022). Thus, bootstrap proportions from different models may be combined using the weighted arithmetic mean. More generally, bootstrap proportions from maximum likelihood models may be combined with posterior probabilities from Bayesian models, as will be illustrated in Section 3.1.

2.2 Combining models by mixtures of their posterior distributions

Another advantage of combining probabilities using the arithmetic mean is its compatibility with mixtures of probability distributions. Mathematically, any weighted arithmetic mean of probabilities is equal to the probability derived from this mixture probability density with the same weights:

$$\overline{f} = w_1 f_1 + w_2 f_2 + \cdots + w_m f_m,$$

where $f_1, f_2, \ldots, f_m$ are the probability densities from the m models. That property will be exploited when combining confidence intervals or their Bayesian counterparts, called *credible intervals*.

2.3 Combining confidence intervals and credible intervals

A 95% credible interval by definition has a 95% posterior probability of containing the divergence time or another quantity of interest according to some posterior probability distribution. That means 95% credible intervals from different models may be combined in two steps:

1. Combine the posterior probability distributions of the models into a mixture distribution (Section 2.2).

2. Derive the 95% credible interval from the that mixture distribution, possibly in the same way as the original 95% credible intervals were derived.

For example, if the models' credible intervals are highest-density intervals of divergence times, then the combined 95% credible interval could be the interval of 95% posterior probability that has the divergence times of highest posterior densities according to the mixture distribution.

Under broad conditions, posterior distributions approximate normal distributions given enough data (Bickel and Doksum, 2000, §5.5). In that case, this algorithm combines the credible intervals from m different models:

1. For each of the m credible intervals to be combined, the corresponding normal distribution is determined by these steps:

   (a) The mean of the distribution is the midpoint between the lower and upper limits of the credible interval $[L, U]$:

   $$\mu = \frac{L + U}{2}.$$

   (b) The standard deviation is the number that makes $U - \mu$, the difference between the upper limit and the mean, equal to the number of standard deviations needed to achieve the same probability as the level of the credible interval. For a 95% credible interval, that difference is 1.96 standard deviations, implying that the standard deviation is

   $$\sigma = \frac{U - \mu}{1.96}.$$

2. Define the combined posterior distribution as the mixture of those m normal distributions (Section 2.2).

3. Define the combined credible interval as the interval with the same amount of posterior probability, according to the combined distribution, as each of the credible intervals according

to their posterior distributions. The simplest way is to do so such that there is equal combined posterior probability below the lower limit and above the upper limit of the combined interval. For example, when combining 95% credible intervals, the combined interval will have 2.5% posterior probability below the combined lower limit and 2.5% posterior probability above the combined upper limit according to the mixture distribution.

That method of combining credible intervals may be easily applied to the output of phylogenetics software using the web app available at `https://davidbickel.shinyapps.io/MixtureUncertainty/` for the case of equal weights.

Just as bootstrap proportions estimate posterior probabilities, 95% confidence intervals estimate 95% credible intervals (Bickel, 2022). For that reason, the methods used to combine credible intervals also may be used to combine confidence intervals. They also apply to combining confidence intervals with credible intervals when there is uncertainty about whether to rely on Bayesian methods.

3 Results: Applications to nucleotide sequence data

3.1 Clade probabilities under uncertainty about Bayesian and frequentist models

Section 2.1's method of propagating model uncertainty to clade probabilities is illustrated in Table 2. Its clade probability conservatively estimated by the bootstrap proportion using maximum likelihood is lower than the probabilities estimated by the Bayesian methods, as is often the case (Bromham, 2016, p. 431). Averaging over the Bayesian models before averaging with the maximum likelihood model (Table 2) is equivalent to the weighted mean with $1/2$ of the weight on the bootstrap proportion and $1/6$ of the weight on each of the three posterior probabilities from the Bayesian methods.

3.2 Divergence times under uncertainty about clock models

Section 2.3's method of propagating model uncertainty to estimated divergence times is illustrated in Figure 1. The results displayed are from the software of Figure 2.

	Clade probability	Intermediate average	Combined probability to report
bootstrap	87%	$\frac{87\%}{1} = 87\%$	$\frac{87\%+100\%}{2} = 94\%$ (over all four models)
SC	100%	$\frac{100\%+100\%+100\%}{3} = 100\%$ (over the Bayesian models)	
RCE	100%		
RCLN	100%		

Table 2: Estimated probability that there is a clade consisting only of *Serratia grimesii* and *Serratia marcescens*, assuming the substitution model of Hasegawa et al. (1985). Whereas the "bootstrap" row gives the bootstrap proportion using maximum likelihood estimation, the other three rows give posterior probabilities from Bayesian methods assuming different clock models: strict clock (SC), relaxed clock exponential (RCE), and relaxed clock lognormal (RCLN).

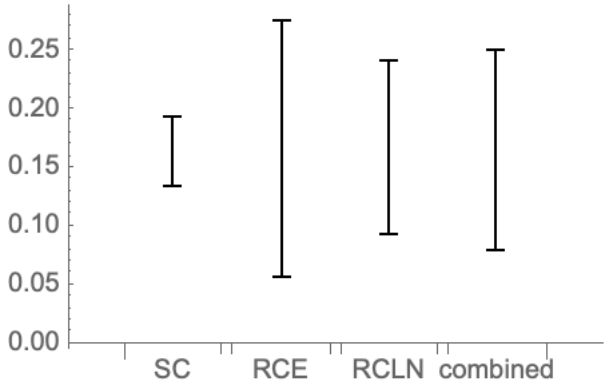

Figure 1: Bayesian 95% credible intervals from the strict clock (SC) model, the relaxed clock exponential (RCE) model, the relaxed clock lognormal (RCLN) model, and the combination of the three models using equal weights for the total tree height of the 40 bacterial DNA sequences of the alignment. The divergence times are not calibrated but are in units relative to a single substitution per site (see Bagley, 2011).

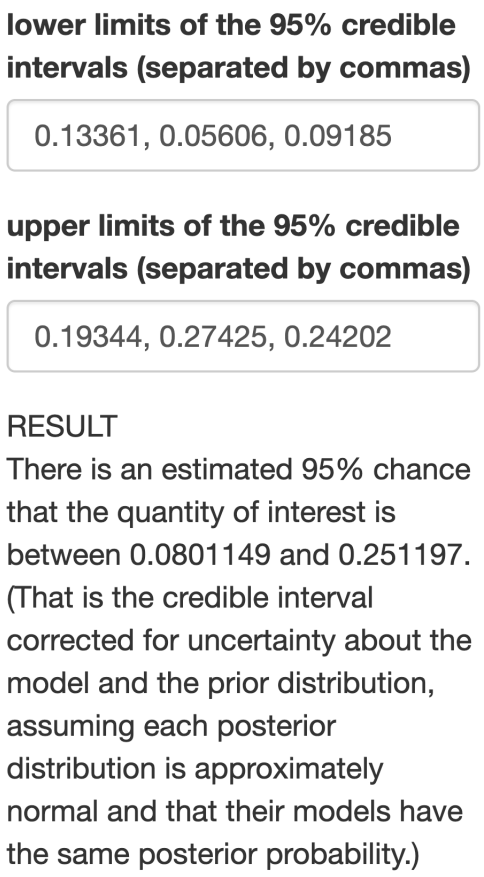

Figure 2: A screenshot from the web app at `https://davidbickel.shinyapps.io/MixtureUncertainty/` (accessed 22 December 2021). Each number in a box is a limit of a 95% credible interval from one of the three models of Figure 1. The "RESULT" gives the combined 95% credible interval of Figure 1.

3.3 Data analysis details

The data analyzed in this section are in the MEGA (Stecher et al., 2020) bacterial DNA sequence alignment file "ebgC.meg" from the Hall (2018) resources at `https://learninglink.oup.com/access/hall-5e-student-resources` (accessed 29 October 2021).

Except as specified above, the bootstrap results were found using the default settings of MEGA 10.2.6 (Stecher et al., 2020).

The Bayesian posterior probabilities and credible intervals were found using BEAST 2.6.6 (Bouckaert et al., 2014), with 10% burnin, 0 posterior probability limit, maximum clade credibility tree, and common ancestor heights in TreeAnnotator 2.6.6, and with other settings at their defaults, except that the clock models were varied as specified above.

Acknowledgments

This research was supported by the University of North Carolina at Greensboro.

References

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M.A., Alekseyenko, A.V., 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution 29, 2157–2167.

Bagley, J., 2011. BEAST and the BEAST basics: molecular clocks and how to input rates into BEAST. Web page, accessed 10 November 2021 URL: `https://bit.ly/3sByH4d`.

Barido-Sottani, J., Bošková, V., Plessis, L.D., Kühnert, D., Magnus, C., Mitov, V., Müller, N.F., PečErska, J., Rasmussen, D.A., Zhang, C., Drummond, A.J., Heath, T.A., Pybus, O.G., Vaughan, T.G., Stadler, T., 2018. Taming the BEAST - A community teaching material resource for BEAST 2. Systematic Biology 67, 170–174.

Bausell, R.B., 2021. The Problem with Science: The Reproducibility Crisis and what to Do about it. Oxford University Press, Oxford.

Bickel, D.R., 2022. Propagating clade and model uncertainty to confidence intervals of divergence times and branch lengths. Molecular Phylogenetics and Evolution 167, 107357.

Bickel, P.J., Doksum, K.A., 2000. Mathematical statistics: basic ideas and selected topics, volume I. CRC Press, New York.

Bouckaert, R., Drummond, A., 2017. bModelTest: Bayesian phylogenetic site model averaging and model comparison. BMC Evolutionary Biology 17, 1–11.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Computational Biology 10.

Bromham, L., 2016. An Introduction to Molecular Evolution and Phylogenetics. Oxford University Press, Oxford.

Bromham, L., 2019. Six impossible things before breakfast: Assumptions, models, and belief in molecular dating. Trends in Ecology & Evolution 34, 474–486.

Cerquides, J., de Mántaras, R.L., 2005. Robust bayesian linear classifier ensembles, in: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (Eds.), Machine Learning: ECML 2005, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 72–83.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge.

Cooke, R.M., 1991. Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press.

Duchêne, D.A., Tong, K.J., Foster, C.S., Duchêne, S., Lanfear, R., Ho, S.Y., 2020. Linking branch lengths across sets of loci provides the highest statistical support for phylogenetic inference. Molecular biology and evolution 37, 1202–1210.

Genest, C., Zidek, J.V., 1986. Combining Probability Distributions: A Critique and an Annotated Bibliography. Statistical Science 1, 114–135.

Hall, B., 2018. Phylogenetic Trees Made Easy: A How-To Manual. Sinauer Associates, New York.

Hantula, D.A., 2019. Replication and reliability in behavior science and behavior analysis: A call for a conversation. Perspectives on Behavior Science 42, 1–11.

Hasegawa, M., Kishino, H., Yano, T.a., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. Journal of Molecular Evolution 22, 160–174.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 226–239.

Le, T., Clarke, B., 2017. A Bayes Interpretation of Stacking for $\mathcal{M}$-Complete and $\mathcal{M}$-Open Settings. Bayesian Analysis 12, 807 – 829.

Li, W.L.S., Drummond, A.J., 2012. Model Averaging and Bayes Factor Calculation of Relaxed Molecular Clocks in Bayesian Phylogenetics. Molecular Biology and Evolution 29, 751–761.

Nakagawa, S., Lagisz, M., Jennions, M., Koricheva, J., Noble, D., Parker, T., Sánchez-Tójar, A., Yang, Y., O'Dea, R., 2021. Methods for testing publication bias in ecological and evolutionary meta-analyses. Methods in Ecology and Evolution doi:10.1111/2041-210X.13724.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2021. Pre-registration: Why and how. Journal of Consumer Psychology 31, 151–162.

Stecher, G., Tamura, K., Kumar, S., 2020. Molecular evolutionary genetics analysis (MEGA) for macOS. Molecular Biology and Evolution 37, 1237–1239.

Stone, M., 1961. The opinion pool. The Annals of Mathematical Statistics 32, pp. 1339–1342.

Sun, M., Zhang, J., 2021. Rampant false detection of adaptive phenotypic optimization by ParTI-based Pareto front inference. Molecular Biology and Evolution 38, 1653–1664.

Wu, C.H., Suchard, M.A., Drummond, A.J., 2013. Bayesian Selection of Nucleotide Substitution Models and Their Site Assignments. Molecular Biology and Evolution 30, 669–688.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A., 2018. Using Stacking to Average Bayesian Predictive Distributions (with Discussion). Bayesian Analysis 13, 917–1007.