Propagating clade and model uncertainty to confidence intervals of divergence times and branch lengths

David R. Bickel

Propagating clade and model uncertainty to confidence intervals of divergence times and branch lengths

August 17, 2021

David R. Bickel

Informatics and Analytics

University of North Carolina at Greensboro

The Graduate School

241 Mossman Building, CAMPUS

Greensboro, NC 27402-6170

dbickel@uncg.edu

Abstract

Confidence intervals of divergence times and branch lengths do not reflect uncertainty about their clades or about the prior distributions and other model assumptions on which they are based. Uncertainty about the clade may be propagated to a confidence interval by multiplying its confidence level by the bootstrap proportion of its clade or by another probability that the clade is correct. (If the confidence level is 95% and the bootstrap proportion is 90%, then the uncertainty-adjusted confidence level is $(0.95)(0.90) = 86\%$.) Uncertainty about the model can be propagated to the confidence interval by reporting the union of the confidence intervals from all the plausible models. Unless there is no overlap between the confidence intervals, that results in an uncertainty-adjusted interval that has as its lower and upper limits the most extreme limits of the models. The proposed methods of uncertainty quantification may be used together.

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

Introduction

You have a sequence alignment and fossil data for generating a time tree. The divergence time of primary interest has a 95% confidence interval of [28 MY, 230 MY], at least according to the first substitution model you try. The proportion of bootstrap samples supporting the corresponding clade is 84%. You wonder how that 84%, not being 100%, should affect the 95% confidence level of the interval. You also try a very different but equally plausible model, obtaining a new confidence interval of [36 MY, 270 MY] with a bootstrap proportion of 93%. You already have three types of uncertainty:

1. Assuming the reconstructed clade is correct, the sources of uncertainty that each model accounts for are reflected in its 95% confidence interval.

2. Apart from those sources of uncertainty, there is uncertainty about which sequences belong together in a clade, uncertainty reflected in the bootstrap proportion, even if the assumptions of the model were known to hold.

3. There is also uncertainty about the assumptions underlying each model, leading to different bootstrap scores and different confidence intervals. Uncertainty in divergence times can differ between models by as much as hundreds of millions of years in cases of cyanobacteria and animals (Bromham, 2019).

How should 95% confidence level be adjusted to reflect the uncertainty about the model and about the clade given each model?

In theory, uncertainty about the model should propagate to the results of data analysis. In practice, the opposite has been observed: researchers have a bias toward models supporting the rejection of null hypotheses (Sun and Zhang, 2021), as if uncertainty in the model should lead to more certain conclusions. For that reason, the advice to try different models when their assumptions are uncertain (e.g., Bromham, 2016, p. 448) is not enough. We need a practical way to follow the recommendation of Bromham et al. (2018) to "report the range of results obtained across the candidate models ... and appropriately summarizing the results." Ideally, we would have an objective and simple approach to propagating uncertainty from the models to the conclusions.

Such an approach is proposed below along with a complementary approach to propagating the uncertainty about the clade. The new approaches work together to propagate both types of uncertainty to the results in the form of adjusted confidence intervals. In general, they not only have wider limits than before but also have their 95% levels reduced to numbers reflecting the

uncertainty involved.

While the fully Bayesian approach would also propagate both types uncertainty, it requires a joint prior distribution not only over the clades but also over the models (Li and Drummond, 2012). Apart from the potential burden of slow computation, the fully Bayesian approach by its nature does not account for the uncertainty in that prior distribution (Bromham et al., 2018). For example, making different assumptions about the prior has lead to divergence time estimates differing by a factor of two in a study of placental mammals (Bromham, 2019). That is why fully Bayesian methods require varying the prior distribution to assess the effect on the results (e.g., Battistuzzi et al., 2018). But then, considering each prior distribution as a different high-level model, we are back to the need for a new approach that specifies how to propagate uncertainty about the model to the conclusions.

New Approaches

The new approaches take the form widening the confidence intervals to account for uncertainty about the model and decreasing the 95% level of a confidence interval to a lower number that quantifies the uncertainty about the clade. The first subsection provides a simple derivation of new approaches applicable not only to confidence intervals but also to Bayesian credible intervals. The second subsection derives numerically equivalent methods from standard frequentist theory.

Approaches for both credible intervals and confidence intervals

The "95%" level of a 95% credible interval has a different technical meaning than the "95%" of a 95% confidence interval. While that difference will become more important in the next subsection, it is stated here to clarify how the same methods apply to both types of intervals.

Let ϑ denote a branch length, divergence time, or other parameter of interest. Assuming the prior distribution is known and that all other model assumptions hold, there is a 95% posterior probability that the true value of ϑ lies within the limits of the 95% credible interval computed from the sequence data. That requires ϑ to have a probability distribution called a *Bayesian posterior distribution* . Whereas ϑ is random, the credible interval computed from the actual data set is fixed.

By contrast, standard frequentist theory would take the branch length or divergence time to be a fixed but unknown number θ, written differently from the ϑ that has a posterior distribution. The sequence data are considered to be randomly generated from a substitution model. Envisioning a

large number of data sets to be generated from the model, each data set has its own 95% confidence interval. With enough of those hypothetical data sets, 95% of their 95% confidence intervals will include the true value of θ.

It is then natural to conclude that there is an estimated 95% probability that the 95% confidence interval computed from the actual data set includes the true value. That type of reasoning is common among scientists (e.g., Felsenstein and Kishino, 1993), and yet statisticians typically regard it as fallacious since it would require that the true value has a distribution, as in the case of credible intervals, for the confidence interval from the real data is fixed (e.g., Morey et al., 2016). Historically, that type of reasoning is associated with Fisher's "fiducial argument" (Zabell, 1992) even though it predated him.

However, since the turn of the century, the statistics community has experienced renewed interest in building mathematical theories and methods for various flavors of fiducial reasoning; see Nadarajah et al. (2015) for a review. While some flavors lean toward standard frequentism, others regard the parameter of interest, again denoted by ϑ, to have a posterior distribution called a *fiducial distribution* that, needing no prior distribution, differs from a Bayesian posterior distribution. A fiducial distribution is called a *confidence distribution* if it has the property that there is a 95% probability that ϑ lies within the limits of a 95% confidence interval computed from the fixed data set. A confidence distribution may be interpreted as an estimate of a Bayesian posterior distribution based on a prior distribution that assumes as little as possible (e.g., Bickel, 2021).

Returning to the example of the introduction, let ϑ_1 denote what the divergence time would be were the first substitution model true, and let ϑ_2 be what it would be under the second model. Then we have both $\Pr(28\,\mathrm{MY} \leq \vartheta_1 \leq 230\,\mathrm{MY}) = 95\%$ and $\Pr(36\,\mathrm{MY} \leq \vartheta_2 \leq 270\,\mathrm{MY}) = 95\%$. A conservative way to combine the two confidence intervals is to take their union:

$$[28\,\mathrm{MY}, 230\,\mathrm{MY}] \cup [36\,\mathrm{MY}, 270\,\mathrm{MY}] = [28\,\mathrm{MY}, 270\,\mathrm{MY}]\,.$$

If the true divergence time ϑ is either ϑ_1 or ϑ_2, then by the rules of probability we have that $\Pr(28\,\mathrm{MY} \leq \vartheta \leq 270\,\mathrm{MY})$ satisfies one of these inequalities:

$$\begin{aligned}
\Pr(28\,\mathrm{MY} \leq \vartheta \leq 270\,\mathrm{MY}) &= \Pr(28\,\mathrm{MY} \leq \vartheta_1 \leq 270\,\mathrm{MY}) \\
&= \Pr(28\,\mathrm{MY} \leq \vartheta_1 \leq 230\,\mathrm{MY}) + \Pr(230\,\mathrm{MY} < \vartheta_1 \leq 270\,\mathrm{MY}) \qquad (1) \\
&\geq \Pr(28\,\mathrm{MY} \leq \vartheta_1 \leq 230\,\mathrm{MY}) = 95\%;
\end{aligned}$$

$$\begin{aligned}\Pr(28\,\text{MY} \le \vartheta \le 270\,\text{MY}) &= \Pr(28\,\text{MY} \le \vartheta_2 \le 270\,\text{MY}) \\ &= \Pr(28\,\text{MY} \le \vartheta_2 < 36\,\text{MY}) + \Pr(36\,\text{MY} \le \vartheta_2 \le 270\,\text{MY}) \qquad (2) \\ &\ge \Pr(36\,\text{MY} \le \vartheta_2 \le 270\,\text{MY}) = 95\%.\end{aligned}$$

Regardless of whether $\vartheta = \vartheta_1$ or $\vartheta = \vartheta_2$, we see that $\Pr(28\,\text{MY} \le \vartheta \le 270\,\text{MY}) \ge 95\%$. It follows that $[28\,\text{MY}, 270\,\text{MY}]$ may be reported as a conservative 95% confidence interval. The method of combining confidence intervals by taking their union can be used to combine any number of confidence intervals from different models with the guarantee of having at least 95% probability of including ϑ. If there are gaps, as when combining two non-overlapping confidence intervals, then the resulting union, not being an interval in the strict since, is better called a *conservative 95% confidence region*.

That addresses the uncertainty about the model but not the uncertainty about the clade, which may be quantified as the proportion P_d of replicate trees from bootstrap samples that reproduce the clade observed in the original tree, where d represents the actual sequence data. Simulation studies (Zharkikh and Li, 1992; Hillis and Bull, 1993) have indicated that the bootstrap proportion tends to be conservative as an estimate of the probability that the method of phylogenetic tree reconstruction would obtain the correct clade given a data set D randomly generated from the substitution model. More precisely, the expected bootstrap proportion is no greater than that probability:

$$E(P_D) \le \Pr(\widehat{c}_D = c), \qquad (3)$$

where P_D is the bootstrap proportion based on D, c is the true clade, and $\widehat{c}_D$ is the clade in the tree reconstructed from D. By the essentially fiducial argument of Felsenstein and Kishino (1993), that probability that the model would yield the correct clade is equal to the probability that the true clade is the clade generated by the actual data:

$$\Pr(\widehat{c}_D = c) = \Pr(C = \widehat{c}_d),$$

where $\widehat{c}_d$ is the clade in the tree reconstructed from D and C is a clade that, rather than being fixed, has the confidence distribution approximated by the histogram of bootstrap replicates. Then we have

$$E(P_D) \le \Pr(C = \widehat{c}_d),$$

expressing of the conservatism of the bootstrap proportion P_d as an unbiased estimate of $E(P_D)$

and a conservative estimate of $\Pr(C = \widehat{c}_d)$.

Since each divergence time or branch length ϑ depends on the accuracy of the computed clade $\widehat{c}_d$ for its meaning, the 95% probability that ϑ lies in its 95% confidence interval is conditional on $C = \widehat{c}_d$. A bound on non-conditional or marginal probability can then be computed as follows. With $[28\,\text{MY}, 230\,\text{MY}]$ as the confidence interval,

$$\begin{aligned}
\Pr(28\,\text{MY} \leq \vartheta_1 \leq 230\,\text{MY}) &= \Pr(C = \widehat{c}_d)\Pr(28\,\text{MY} \leq \vartheta_1 \leq 230\,\text{MY} \mid C = \widehat{c}_d) \\
&\quad + \Pr(C \neq \widehat{c}_d)\Pr(28\,\text{MY} \leq \vartheta_1 \leq 230\,\text{MY} \mid C \neq \widehat{c}_d) \\
&\geq \Pr(C = \widehat{c}_d)\Pr(28\,\text{MY} \leq \vartheta_1 \leq 230\,\text{MY} \mid C = \widehat{c}_d) \\
&\geq E(P_D)\Pr(28\,\text{MY} \leq \vartheta_1 \leq 230\,\text{MY} \mid C = \widehat{c}_d) \\
&\geq E(P_D)(95\%).
\end{aligned}$$

Thus, $0.95 \times P_d$ is a conservative estimate of $\Pr(28\,\text{MY} \leq \vartheta_1 \leq 230\,\text{MY})$. That is $(0.95)(0.84) = 80\%$ using $P_d = 84\%$ under the first substitution model of the introduction. The confidence level of the confidence interval is in that way reduced from 95% to 80% to propagate the uncertainty about the clade. Similarly, since $P_d = 93\%$ according to the second model, $0.95 \times 0.93 = 88\%$ is a conservative estimate of $\Pr(36\,\text{MY} \leq \vartheta_2 \leq 270\,\text{MY})$.

How does uncertainty about the clade affect the confidence level of $[28\,\text{MY}, 270\,\text{MY}]$, the confidence interval combined across the two models? By following the steps in equations (1) and (2), we obtain 80% as a conservative estimate of $\Pr(28\,\text{MY} \leq \vartheta_1 \leq 270\,\text{MY})$ and 88% as a conservative estimate of $\Pr(28\,\text{MY} \leq \vartheta_2 \leq 270\,\text{MY})$. Uncertain about which model is true, we can report 80% as the estimate of $\Pr(28\,\text{MY} \leq \vartheta \leq 270\,\text{MY})$ since that estimate, being the lower of the model-specific estimates, is conservative regardless of whether $\vartheta = \vartheta_1$ or $\vartheta = \vartheta_2$. More generally, the lowest bootstrap proportion across all models would be a conservative estimate of the probability that the divergence time or branch length is in the union of the confidence intervals computed from the sequence data.

Since the rules of probability were applied above with a parameter ϑ having a probability distribution and with confidence intervals computed from actual data, the approaches apply to credible intervals without modification. The above inequalities can be seen to hold under Bayesian model averaging regardless of how much prior probability is assigned to each model. The interval-union inequalities also hold when there is uncertainty about whether to use a frequentist method or a Bayesian method. By contrast, the approaches of the next section do not apply to credible

intervals that do not approximate confidence intervals.

A derivation specific to confidence intervals

The new methods of this section, while coinciding numerically with the above methods, have the advantage of the standard frequentist interpretation. The method of propagating uncertainty about models by taking the union of their confidence intervals can be retained without modification since the resulting unions of confidence intervals generated from randomly drawn data sets include the fixed true branch length or divergence time with at least 95% probability.

To propagate uncertainty about clades, we need to make equation (3) more precise by setting a threshold $P_{\min}$ that the bootstrap proportion must exceed for conservatism; the results of Zharkikh and Li (1992) and Hillis and Bull (1993) suggest $P_{\min} = 75\%$ and $P_{\min} = 50\%$, respectively. Then

$$E\left(P_D \,|\, P_D > P_{\min}\right) \leq \Pr\left(\widehat{c}_D = c \,|\, P_D > P_{\min}\right).$$

With $\left[\theta_D^{\text{lower}}, \theta_D^{\text{upper}}\right]$ as the 95% confidence interval for a random set of sequences D that is valid whenever $\widehat{c}_D = c$ and that is approximately independent of the event that $P_D > P_{\min}$, the joint probability of that event and the event that $\left[\theta_D^{\text{lower}}, \theta_D^{\text{upper}}\right]$ covers θ, the true branch length or divergence time, is

$$\begin{aligned}\Pr\left(\widehat{c}_D = c, \theta_D^{\text{lower}} \leq \theta \leq \theta_D^{\text{upper}} |\, P_D > P_{\min}\right) &\approx \Pr\left(\widehat{c}_D = c \,|\, P_D > P_{\min}\right) \Pr\left(\theta_D^{\text{lower}} \leq \theta \leq \theta_D^{\text{upper}} | \widehat{c}_D = c\right) \\ &\gtrapprox E\left(P_D \,|\, P_D > P_{\min}\right)(95\%),\end{aligned}$$

which is conservatively estimated by $0.95 \times P_d$ if $P_d > P_{\min}$.

Application to nucleotide sequence data

The running example's confidence intervals and bootstrap proportions are those of Figure 1, clade A. The results for the other clades are given in Table 1.

The bootstrap proportions and confidence intervals were computed by maximum likelihood, Gamma distributed rates, complete deletion, Gram-positive species defining the outgroup, and calibrated by the divergence between *E. coli* and *S. enterica* at 100 MY −160 MY (Hall, 2018, chapter 15) using MEGA X (Kumar et al., 2018; Stecher et al., 2020) with the models of Table 1, otherwise with default settings.

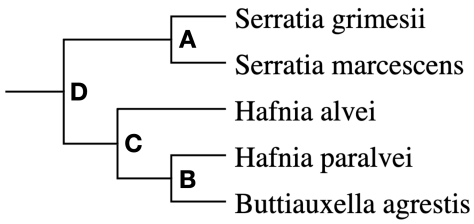

Figure 1: Part of the trees reconstructed

Clade	3-parameter (Tamura, 1992)	General time-reversible (Nei and Kumar, 2000)	Uncertain model
A	80%; [28 MY, 230 MY]	88%; [36 MY, 270 MY]	80%; [28 MY, 270 MY]
B	30%; [31 MY, 250 MY]	27%; [44 MY, 320 MY]	27%; [31 MY, 320 MY]
C	77%; [37 MY, 250 MY]	73%; [55 MY, 330 MY]	73%; [37 MY, 330 MY]
D	86%; [66 MY, 370 MY]	86%; [93 MY, 460 MY]	86%; [66 MY, 460 MY]

Table 1: $0.95 \times P_d$; divergence time confidence interval

Acknowledgments

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

Data Availability

The data underlying this article are in the MEGA (Kumar et al., 2018) sequence alignment file named "ebgC.meg" that is available in the Hall (2018) resources at `https://learninglink.oup.com/access/hall-5e-student-resources`.

References

Battistuzzi, F.U., Tao, Q., Jones, L., Tamura, K., Kumar, S., 2018. RelTime Relaxes the Strict Molecular Clock throughout the Phylogeny. Genome Biology and Evolution 10, 1631–1636.

Bickel, D.R., 2021. Null hypothesis significance testing interpreted and calibrated by estimating probabilities of sign errors: A Bayes-frequentist continuum. The American Statistician 75, 104–112.

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.

Bromham, L., Duchéne, S., Hua, X., Ritchie, A.M., Duch'ene, D.A., Ho, S.Y.W., 2018. Bayesian molecular dating: opening up the black box. Biological Reviews 93, 1165–1191.

Felsenstein, J., Kishino, H., 1993. Is there something wrong with the bootstrap on phylogenies? a reply to hillis and bull. Systematic Biology 42, 193–200.

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

Hillis, D.M., Bull, J.J., 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. Systematic Biology 42, 182–192.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution 35, 1547.

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.

Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E.J., 2016. The fallacy of placing confidence in confidence intervals. Psychonomic Bulletin & Review 23, 103–123.

Nadarajah, S., Bityukov, S., Krasnikov, N., 2015. Confidence distributions: A review. Statistical Methodology 22, 23–46.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Oxford.

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

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.

Tamura, K., 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+ C-content biases. Mol Biol Evol 9, 678–687.

Zabell, S.L., 1992. R. A. Fisher and the fiducial argument. Statistical Science 7, 369–387.

Zharkikh, A., Li, W.H., 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Molecular Biology and Evolution 9, 1119–1147.