Missing the point (estimate): Bayesian and likelihood phylogenetic reconstructions of morphological characters produce generally concordant inferences. A comment on Puttick *et al.* (2017)

Joseph W. Brown[1*], Caroline Parins-Fukuchi[1*], Gregory W. Stull[1], Oscar M. Vargas[1], and Stephen A. Smith[1]

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA

[*]Equal authorship. Emails: josephwb@umich.edu, cfukuchi@umich.edu

Abstract

Puttick *et al.* (2017) performed a simulation study to compare accuracy between methods inferring phylogeny from discrete morphological characters. They report that a Bayesian implementation of the Mk model (Lewis, 2001) was most accurate (but with low resolution), while a maximum likelihood (ML) implementation of the same model was least accurate. They conclude by strongly advocating that Bayesian implementations of the Mk model should be the default method of analysis for such data. While we applaud investigations into accuracy and alternative methods of analysis, this conclusion is based on an inappropriate comparison of the ML point estimate with the Bayesian consensus. We revisit these issues through simulation by considering uncertainty in ML reconstructions, and demonstrate that Bayesian and ML estimates are generally concordant when conventional edge support thresholds are considered. We therefore disagree with the conclusions of Puttick *et al.* (2017), and consider their prescription of *any* default method to be unfounded. Instead, we recommend caution and thoughtful consideration of the model or method being applied to a morphological dataset.

Key words: phylogeny, morphology, paleontology, Bayesian, likelihood

Comparing point estimates to consensus summaries

Puttick *et al.* (2017) report that ML tree inference under the Mk model results in higher topological error than Bayesian implementations. However, this result is driven precisely by the comparison of maximum likelihood point estimates (MLE) to Bayesian majority-rule (BMR) consensus trees. MLE topologies are fully resolved, but this stems from the standard binary tree searching algorithms employed and not from an explicit statistical rejection of unresolved nodes. Therefore, individual MLE estimates may contain edges with negligible statistical support. On the other hand, consensus summaries, independent of phylogenetic method, may have reduced resolution as a product of uncertainty arising by summarization across conflicting sampled topologies. Thus, a direct comparison between a consensus tree (i.e., BMR) and a point estimate (i.e., MLE) is inappropriate. BMR topologies of Puttick *et al.* (2017) are more accurate because poorly supported conflicted edges were collapsed, while MLE topologies were fully resolved, even if poorly supported. While contrasting MLE and Bayesian maximum *a posteriori* (MAP) trees would be a more appropriate comparison of optimal point estimates, the incorporation of uncertainty is an integral part of all phylogenetic analysis. Therefore, comparison of consensus trees from Bayesian and ML analyses hold more practical utility for systematists. For these reasons, we argue that the results of Puttick *et al.* (2017) are an artifact of their comparison between fundamentally incomparable sets of trees.

Support metrics are available for morphological characters

To avoid drawing untenable conclusions, it is *de rigueur* of any phylogenetic analysis to explicitly assess edge support. Systematists often accomplish this via non-parametric bootstrap sampling (Felsenstein, 1985), though other measures exist (see below). Puttick *et al.* (2017) did not assess edge support in their ML estimates, stating that morphological (but not genetic) data do not meet an underlying assumption of the bootstrap statistical procedure that phylogenetic signal is distributed randomly among characters. The authors do not explain the meaning of this statement, and no references are provided to support the assertion. Non-parametric bootstrapping has been a staple of phylogenetic reconstruction for decades, including for the analysis of discrete morphological characters. Bootstrapping works via the assumption that the observed characters are a representative sample from a population of possible characters evolving under the same process, and thus can be resampled to assess confidence in parameters (Felsenstein, 1985). While morphological matrices typically include only variable characters (i.e., an ascertainment bias), this is an informative subset of the possible characters, and should not be thought of as misleading calculations. Were this otherwise, the original sample would be likewise suspect, as the use of model-based phylogenetic inference (such as Mk) explicitly assumes characters evolve according to the same process. Concerns about the interpretation and use of the bootstrap exist (Sanderson, 1995), the primary of which involves the assumption that individual characters are statistically independent. However, it is reasonable to assume that individual sites in a morphological matrix would be more independent than adjacent sites from the same gene, and genetic datasets are routinely bootstrapped. We therefore disagree with the claims of Puttick *et al.* (2017) that bootstrapping is inappropriate for morphological data, or at least any *more* inappropriate than for genetic data.

There are also other methods researchers can use to assess edge support in a likelihood framework. Jackknifing, unlike bootstrapping, samples without replacement, conditioning on strict subsets of the observed data. More recently, the SH-like test (Guindon *et al.*, 2010) computes support for each internal edge in the MLE tree by considering all nearest neighbour interchanges (NNIs). This test is implemented in several software packages including RAxML (Stamatakis, 2014), one of the programs used by Puttick *et al.* (2017). Alternatively, ML programs frequently offer an option to collapse edges on a MLE tree that fall below some minimum threshold length. Use of any of these options would enable a fairer comparison of likelihood and Bayesian reconstructions.

ML and Bayesian comparisons incorporating uncertainty

To measure the effect of comparing BMR and MLE trees, we used the simulation code from Puttick *et al.* (2017) to generate 1000 character matrices, each of 100 characters on a fully pectinate tree of 32 taxa, as these settings generated the most discordant results in Puttick *et al.* (2017). Each matrix was analyzed in both Bayesian and ML frameworks using the Mk+G model (Lewis, 2001). Bayesian reconstructions were performed using MrBayes v3.2.6 (Ronquist *et al.*, 2012), using the same settings as Puttick *et al.* (2017): 2 runs, each with 5 x 10^5 generations, sampling every 50 generations, and discarding the first 25% samples as burnin. As in Puttick *et al.* (2017), we summarized each analysis with a BMR consensus tree (i.e. only edges with >= 0.5 posterior probability are represented). Likelihood analyses were performed in RAxML v8.2.9 (Stamatakis, 2014). For each simulated matrix we inferred both the MLE tree and 200 nonparametric bootstrap trees. Accuracy in topological reconstruction was assessed using the Robinson-Foulds (RF) distance (Robinson and Foulds, 1981), which counts the number of unshared bipartitions between trees. We measured the following distances from the true simulated tree: d_{BMR}, the distance to the Bayesian majority-rule consensus; d_{MLE}, the distance to the MLE tree; d_{ML50}, the distance to the MLE tree which has had all edges with <50% bootstrap support collapsed. Finally, for each matrix we calculate $D_{MLE} = d_{MLE} - d_{BMR}$, and $D_{ML50} = d_{ML50} - d_{BMR}$. These paired distances measure the relative efficacy of ML and Bayesian reconstructions: values of D greater than 0 indicate that ML produces less accurate estimates (that is, with a greater RF distance from the true generating tree).

As demonstrated by Puttick *et al.* (2017), MLE trees are indeed less accurate than BMR trees (Figure 1; D_{MLE}), with MLE trees on average having an RF distance 17.6 units greater than the analogous Bayesian consensus distance. However, when collapsing MLE edges with less than 50% bootstrap support, Bayesian and ML differences are normally distributed around 0 (Figure 1; D_{ML50}), indicating that when standardizing the degree of uncertainty in tree summaries there is no difference in topology reconstruction accuracy. These results support the argument that the original comparisons made in Puttick *et al.* (2017) of MLE and BMR trees are inappropriate. Depending on the level of uncertainty involved, an optimal point estimate from a distribution (e.g., MLE or MAP) may be arbitrarily distant from a summary of the same distribution. And so, the differences in MLE vs. BMR are not expected to be consistent.

The expected concordance of Bayesian and ML results

Our results reveal much greater congruence between Bayesian and ML estimates than suggested by Puttick *et al.* (2017). This is to be expected. ML and Bayesian tree construction methods should yield similar results under the conditions in which they are often employed. While Bayesian tree reconstruction differs from ML by incorporating prior distributions, the methods share likelihood functions. In phylogenetics, researchers typically adopt non-informative priors, with a few exceptions (e.g., priors on divergence time parameters). Arguments can be made for pseudo-Bayesian approaches when care is taken to ensure that priors used are truly uninformative, which result in posterior probabilities that mirror the likelihood and are therefore congruent with ML (Alfaro and Holder 2006; Gelman *et al.* 2014). If prior distributions are formulated thoughtfully, as with Wright *et al.* (2016), in shaping the Mk model using hyperpriors to accommodate character change heterogeneity, Bayesian methods can outperform ML.

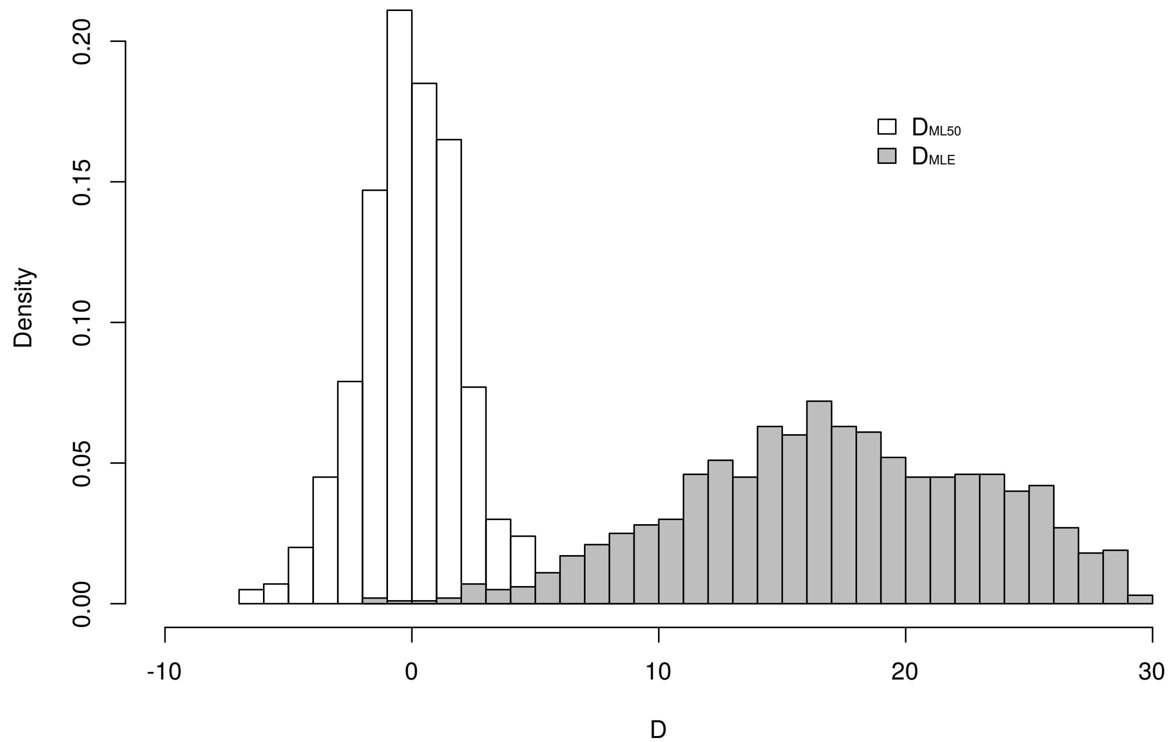

Figure 1: Topological accuracy of ML vs. Bayesian reconstructions. D measures how much larger ML distances are from the true tree (d_{ML}) than are Bayesian distances (d_{BMR}). MLE trees are indeed less accurate than BMRs (D_{MLE}), but when conventional bootstrap thresholds are employed (D_{ML50}) the difference in efficacy disappears.

Alternatively, inappropriate priors can positively mislead (Gelman *et al.*, 2014). Generally, when informative prior distributions are known or can be estimated using hierarchical approaches, Bayesian reconstruction methods may be strongly favored over ML. It is unclear whether Puttick *et al.* (2017) intend to draw the comparisons discussed above as they do not describe any reasons to prefer Bayesian over ML in principle.

Although our results demonstrate general concordance between ML and Bayesian approaches when uncertainty is represented, further simulation work is needed to determine the extent and conditions of this concordance. Issues surrounding the application of Bayesian methods are particularly important in paleontology, where researchers often conduct inference upon very limited data. In these cases, it may be desirable to construct informative prior distributions when conducting Bayesian analyses (Gelman *et al.*, 2014). The questions posed by Puttick *et al.* (2017) are critically important as statistical morphological phylogenetics moves forward. However, their inappropriate comparison between ML and Bayesian approaches leaves the relative performance of the two implementations of the Mk model unresolved.

We conclude by stating that we are not advocating one method over another for morphological phylogenetic reconstruction. Methods differ in model (Mk vs. parsimony), inferential paradigm (parsimony vs. ML/Bayesian), assumptions (prior distributions, model adequacy), interpretation, and means to incorporate uncertainty (ML/parsimony vs. Bayesian). We therefore recommend caution and thoughtful consideration of the biological question being addressed and then choosing the method that will best address that question. All inferential approaches possess strengths and weaknesses, and it is the task of researchers to determine the most appropriate given available data and the questions under investigation. The excitement of new morphological data sources and new means for analyzing these data should not overshadow the obligation to apply methods thoughtfully.

Authors' contributions

J.W.B. conceived the design of the study and performed the analyses; J.W.B. and C.F.-P. drafted the manuscript; all authors contributed to the interpretation of results and the writing of the manuscript.

Acknowledgements

We thank Mark Puttick for sharing datasets and thoughts on bootstrap resampling. We thank members of the Smith laboratory for thoughtful discussions on an earlier draft of this manuscript. J.W.B. and C.F.-P. thank Annika Hansen for being a stalwart leading example of objective criticism. This is paper #1 of the PRUSSIA working group at UM.

References

Alfaro, M. E. and Holder, M. T. 2006. The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 37: 19–42.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4): 783–791.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2014. *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Systematic Biology*, 59(3): 307–321.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6): 913–925.

Puttick, M. N., O'Reilly, J. E., Tanner, A. R., Fleming, J. F., Clark, J., Holloway, L., Lozano-Fernandez, J., Parry, L. A., Tarver, J. E., Pisani, D., and Donoghue, P. C. J. 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society of London B: Biological Sciences*, 284(1846): 20162290.

Robinson, D. F. and Foulds, L. R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2): 131–147.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3): 539–542.

Sanderson, M. J. 1995. Objections to bootstrapping phylogenies: a critique. *Systematic Biology*, 44(3): 299–320.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

Wright, A. M., Lloyd, G. T., and Hillis, D. M. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, 65(4): 602–611.