

LETTER

Choice of Topology Estimators in Bayesian Phylogenetic Analysis

Jeet Sukumaran and Charles W. Linkem

Department of Ecology and Evolutionary Biology, Biodiversity Institute and Natural History Museum, University of Kansas

Wheeler WC and Pickett KM (2008. Topology-Bayes versus clade-Bayes in phylogenetic analysis. *Mol Biol Evol.* 25:447–453.) discuss two ways of summarizing the posterior probability distribution of a Bayesian phylogenetic analysis, which they refer to as “topology-Bayes” and “clade-Bayes.” They claim that the clade-Bayes approach leads to problems such as “exaggerated clade support, inconsistently biased priors, and the impossibility of topology hypothesis testing,” which are not problems for the topology-Bayes approach. However, their argument for topology-Bayes over clade-Bayes is based on errors in the interpretation of summary statistics associated with Bayesian phylogenetic analysis. Although there is a well-documented difference between the maximum posterior probability topology and the majority-rule consensus topology (the established terms for topology-Bayes and clade-Bayes summaries, respectively), both have a place in phylogenetic analysis. Choice of summarization strategy should be driven by choice of parameters that need to be estimated versus those to be marginalized given the evolutionary questions being asked or hypotheses being tested.

Wheeler and Pickett (2008) discuss two alternate approaches to summarizing the posterior distribution resulting from a Bayesian phylogenetic analysis, the maximum posterior probability topology, and the majority-rule consensus topology. The maximum posterior probability topology is a summary that estimates the topology with the highest marginal posterior probability, whereas the majority-rule consensus topology is a summary that provides the consensus topology of splits with a marginal posterior probability greater than 0.5 (Larget and Simon 1999; Felsenstein 2004; Yang 2006). Wheeler and Pickett (2008) refer to these summaries as topology-Bayes and clade-Bayes, respectively. They assert that topology-Bayes (i.e., the maximum posterior probability topology) provides a topology selected by an optimality criterion (the maximum a posteriori tree) and thus represents a proper phylogenetic hypothesis of relationships. On the other hand, Wheeler and Pickett (2008) argue that clade-Bayes (i.e., the majority-rule consensus topology) is an inappropriate summary of the posterior probability of a phylogeny to be used in phylogenetic analysis and is flawed by three major problems: exaggerated clade support, inconsistently biased priors, and the impossibility of topology hypothesis testing. We disagree with this interpretation and characterization of differences between these two different summarization strategies. In this paper, we first show that there is a fundamental error in the probability formulation of the clade-Bayes presented by Wheeler and Pickett (2008). Although the correction of this error might not be germane to the essential core of the arguments of Wheeler and Pickett (2008), we feel that such a substantive statistical error should not stand uncorrected in the literature of the community, to avoid misleading current and future workers. We then show that the three “problems” that Wheeler and Pickett (2008) argue result from the usage of clade-Bayes are not so much problems but rather natural and predictable properties of this summarization strategy under probability theory and, moreover, in some cases

apply to the topology-Bayes estimator as well. We finally suggest that the characterization of phylogenetic hypothesis testing in the Bayesian context as given by Wheeler and Pickett (2008) is misleading in its emphasis on the optimal topology as the primary parameter to be estimated. We argue that one of the strengths of a Bayesian analysis is the ability to estimate a particular parameter or parameters of interest while simultaneously accommodating for uncertainty in other parameters, and, depending on the evolutionary questions being considered by an investigator, the optimal topology may not even be a parameter of interest.

Equation (3) of Wheeler and Pickett (2008) purports to represent the posterior probability of a clade given the data. Unfortunately, this equation violates basic probability theory: The denominator of the right-hand term does not denote an exhaustive set of mutually exclusive points in the sample space as required by the law of total probability (O’Hagan and Forster 2004). Thus, the equation is invalid. The correct posterior probability of a clade is actually given by the sum of the posterior probabilities of the trees in which the split subtending the clade occurs (Larget and Simon 1999; Yang 2006), and this is indeed how this is implemented in a number of different Bayesian phylogenetic programs such as MrBayes (Huelsenbeck and Ronquist 2001), as acknowledged by Wheeler and Pickett (2008).

To demonstrate exaggerated clade support, Wheeler and Pickett (2008) describe a hypothetical scenario in which support for a relationship increases as data are duplicated. Although their calculations for the posterior probability of clades here is correct, their statement that “evidence in favor of AB is equal to that against” in their example is not. In a statistical framework, evidence for a hypothesis is the statistical support for that hypothesis. In their example, the topology ((A,B),C,D) is better supported by the data than any other topology (e.g., yielding a tree length of 6 as opposed to 7 if we were to adopt a parsimony criterion). As such, it is not surprising that support for this split increases as the data supporting this split increase. Wheeler and Pickett (2008) state that the maximum posterior probability topology is not subject to this increase in support as data increase. It is unfortunate that they make this claim without any discussion or proof as their own example demonstrates the opposite. In fact, in the four-taxon example that they present, the posterior probability of the topology

Key words: Bayesian phylogenetics, summaries, topology estimators, support.

E-mail: jeet@ku.edu.

Mol. Biol. Evol. 26(1):1–3. 2009

doi:10.1093/molbev/msn250

Advance Access publication November 4, 2008

is identical to the clade posterior probability that one would focus on in the “clade-Bayes” approach, thus in this case making clade-Bayes and topology-Bayes approaches equal. This fact is not limited to the four-taxon case but is true of all topologies when data are duplicated; clade posterior probability and topology posterior probability will both increase by the same factor. Therefore, the statement by Wheeler and Pickett (2008) “if the T_1 approach is adopted, the support inflation seen in T_c becomes moot” is incorrect because support “inflation” will increase in both approaches equally. Furthermore, it should be noted that this is not (over) inflation but the expected result when data in support of a clade and topology are increased. Thus, because this result is expected in a statistical framework, we therefore argue that calling this increase “exaggerated” is misleading.

Wheeler and Pickett (2008) criticize the clade-Bayes approach as problematic due to “inconsistently biased” priors induced upon clades when uniform priors are placed on topologies. The issue has been addressed thoroughly elsewhere (Brandley et al. 2006; Velasco 2007), and so we refrain from reproducing the same arguments in their entirety here. As discussed by Velasco (2007), this problem is actually a fundamental property of probabilities on trees. In the case of complex, hierarchical models (such as those used in phylogenetic inference), it is impossible to decompose the problem into many different parameterizations simultaneously. We need to decide a priori the parameterization that allows us to best describe our prior knowledge in terms of a probability statement and accept the way the probability gets distributed to other parameters in alternate parameterizations of the hierarchical model. If we believe that all topologies are equally likely before analyzing the data, then we are making implicit statements regarding our beliefs on the probability of clades of different sizes (i.e., that extreme-sized clades—either larger or smaller—are more probable a priori than moderate-sized ones). Thus, the problem of nonuniform clade priors given uniform priors on topologies is not a problem as such but really a case of incorrect intuition about probability statements on trees, leading in turn to the erroneous expectation that all clades should be equally probable regardless of size (Velasco 2007).

Wheeler and Pickett (2008) also suggest that the clade-Bayes approach leads to the “impossibility of hypothesis testing” because the topology resulting from such a summary does not represent an actual phylogenetic hypothesis (i.e., an actual topology preferred under a particular optimality criteria) but rather a consensus of clades with greater than 50% posterior probability. One of the strengths of a Bayesian analysis is the ability to estimate a particular parameter or parameters of interest while simultaneously accommodating for uncertainty in other parameters (nuisance parameters). Different summarization strategies marginalize over different nuisance parameters. For example, if we were interested in comparing two (or more) fully specified trees, that is, identifying the hypothesis of topology that best describes the relationships of a group of organisms, then the maximum posterior probability topology is indeed an appropriate summary. The Bayesian approach to select among the different hypotheses in this context would be to simply compare the posterior probabilities of those topologies. Alternatively, if we were interested

in the labeled membership of a particular clade (i.e., the monophyly of a group), the majority-rule consensus topology is a more appropriate summary. In some cases, the hypothesis in question involves not the tree topology itself but rather an aspect of the evolutionary process that can only be estimated with some knowledge of a phylogeny. The work of Lutzoni et al. (2001) is an example of this approach—they calculate an estimate of an evolutionary parameter over every tree from their approximation of the posterior distribution rather than first summarizing the posterior probability distribution with one tree topology and then answering their question from that single tree. For the hypothesis of interest to them, this approach is superior to using one estimate of tree—regardless of whether that estimate of the tree is the maximum posterior probability tree or a consensus tree—as it allows them to provide the best estimate of the parameters of interest to them while incorporating uncertainty in topologies and other parameters.

Wheeler and Pickett (2008) portray clade-Bayes and topology-Bayes as very different analyses, or analyses under different optimality criteria, when they are in fact different ways of summarizing the same posterior distribution. Choice of summarization strategy does not reflect choice of different optimality criteria in the way presented by Wheeler and Pickett (2008) but rather a choice of parameters to estimate versus parameters to integrate over. Thus, the appropriateness of a particular summary over others is contingent on what the parameters of interest are given the evolutionary questions being asked or hypotheses being tested. The notion that these summaries are not necessarily equivalent, that is, the maximum posterior probability tree \neq majority-rule tree, is uncontroversial. It is up to the investigator to select the appropriate topology for his or her application.

Acknowledgments

We would like to thank the University of Kansas Systematics Discussion Group, Mark Holder, Rafe Brown, and John Kelly for discussions, ideas, comments, and reviews. J.S. acknowledges funding support from National Science Foundation’s Cyberinfrastructure for Phylogenetic Research (CIPRES) project (award number 0715370, PI: Tandy Warnow), while C.W.L. was supported by DEB 0640737 to Rafe Brown during the preparation of this manuscript. We also would like to thank the MBE chief editor, Marcy Uyenoyama, the Associate Editor, Scott Edwards, as well as four anonymous reviewers for various comments and critiques that greatly strengthened this letter.

Literature Cited

- Brandley MC, Leaché A, Warren D, McGuire J. 2006. Are unequal clade priors problematic for Bayesian phylogenetics? *Syst Biol.* 55:138–146.
- Felsenstein J. 2004. *Inferring phylogenies*. Sunderland (MA): Sinauer Associates, Inc.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 17:754–755.
- Larget B, Simon DL. 1999. Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol Biol Evol.* 16:750–759.

- Lutzoni F, Pagel M, Reeb V. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. *Nature*. 411:937–940.
- O'Hagan A, Forster J. 2004. Kendall's advanced theory of statistics. Volume 2B: Bayesian inference, 2nd ed. London: Arnold Publishers.
- Velasco JD. 2007. Why non-uniform priors on clades are both unavoidable and unobjectionable. *Mol Phylogenet Evol*. 45:748–749.
- Wheeler WC, Pickett KM. 2008. Topology-Bayes versus clade-Bayes in phylogenetic analysis. *Mol Biol Evol*. 25:447–453.
- Yang Z. 2006. *Computational molecular evolution*. Oxford: Oxford University Press.

Scott Edwards, Associate Editor

Accepted October 29, 2008