Available online at www.sciencedirect.com

SCIENCE DIRECT®

Molecular Phylogenetics and Evolution 34 (2005) 203–211

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.elsevier.com/locate/ympev

Strange bayes indeed: uniform topological priors imply non-uniform clade priors

Kurt M. Pickett[a], Christopher P. Randle[b,*,1]

[a] Division of Invertebrate Zoology, American Museum of Natural History, 79th Street at Central Park West, New York, NY 10024, United States
[b] Department of Ecology and Evolutionary Biology and Natural History Museum, University of Kansas, 1200 Sunnyside Ave.
Lawrence, KS 66045, United States

Received 4 June 2004; revised 3 September 2004

Abstract

While Bayesian analysis has become common in phylogenetics, the effects of topological prior probabilities on tree inference have not been investigated. In Bayesian analyses, the prior probability of topologies is almost always considered equal for all possible trees, and clade support is calculated from the majority rule consensus of the approximated posterior distribution of topologies. These uniform priors on tree topologies imply non-uniform prior probabilities of clades, which are dependent on the number of taxa in a clade as well as the number of taxa in the analysis. As such, uniform topological priors do not model ignorance with respect to clades. Here, we demonstrate that Bayesian clade support, bootstrap support, and jackknife support from 17 empirical studies are significantly and positively correlated with non-uniform clade priors resulting from uniform topological priors. Further, we demonstrate that this effect disappears for bootstrap and jackknife when data sets are free from character conflict, but remains pronounced for Bayesian clade supports, regardless of tree shape. Finally, we propose the use of a Bayes factor to account for the fact that uniform topological priors do not model ignorance with respect to clade probability.

Keywords: Bayesian phylogenetics; Clade support; Prior probability; Jackknife; Bootstrap

1. Introduction

Over the past few years, a method of complex problem-solving known as Markov Chain Monte Carlo (MCMC) has been gaining popularity among phylogeneticists (see reviews in Holder and Lewis, 2003; Huelsenbeck et al., 2002; Lewis, 2001). MCMC itself is not new, dating from Metropolis et al. (1953), and its Bayesian character—its ability to sample a posterior distribution—is well established (Tierney, 1994). But the implementation of the MCMC algorithm as an application in phylogenetics is fairly new, originating with the doctoral dissertation work of Li (1996) and Mau (1996). Others have discussed Bayesian interpretations of phylogenetic problems (Farris, 1973; Harper, 1979; Wheeler, 1991), but these did not involve MCMC, and so we do not treat those interpretations here.

Considering only explicitly statistical methods of phylogenetics, a Bayesian approach is, in some ways, more appealing than the likelihood approach. As is well known, the likelihood of a hypothesis (here, the tree) given the data is proportional to the probability of the data given the hypothesis (see Edwards, 1992, p. 9). Phylogeneticists—whose primary investigation usually relates to tree selection—are more concerned with the probability of the tree, conditional on the model and the data (rather than the probability of

[*] Corresponding author. Fax: +1 785 864 5860.
E-mail address: randle@ku.edu (C.P. Randle).
[1] Authors contributed equally and are listed alphabetically by convention.

1055-7903/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2004.09.001

the data), and this is what Bayes' formula provides. But to accomplish this inversion (see Farris, 1973), information regarding the prior probabilities of the trees is needed. Many advocates of Bayesian phylogenetics have commented on the importance of this prior assessment of tree probabilities (Huelsenbeck et al., 2001, 2002; Lewis, 2001) and have mentioned that prior selection can be problematic (Holder and Lewis, 2003; Lewis, 2001).

Because of these issues, some questions remain open. What constitutes a reasonable topological prior probability, and how does one arrive at such a distribution? The estimation of prior probabilities is difficult when little is known about the phylogeny of a group of organisms beforehand, which is most often, or arguably *always*, the case for systematic studies.

Proponents of the new Bayesian approach to phylogeny have advocated the use of uniform topological priors to reflect ignorance (see review in Huelsenbeck et al., 2002; Lewis, 2001). When nothing is known regarding the relationships of taxa prior to analysis, all tree hypotheses are considered to be equally probable. This may be valid given the "principle of insufficient reason" (LaPlace 1820, as cited in Kass and Wasserman, 1996). For example, Farris (1973, p. 251) argued that, in a Bayesian framework, "... P {E}, the probability of evolutionary hypothesis [tree] E not conditional upon any data, may be treated as if equal for all E." While it may seem paradoxical on the one hand to claim the superiority of a method due to its ability to incorporate prior knowledge, and, on the other, to claim that ignorance should be modeled, this concern is not unique to Bayesian *phylogenetics*, and forms the kernel of the schism between the "empirical" and "subjective" schools of Bayesian statistics. This debate is beyond the scope of the present study.

Other than modeling ignorance, justifications for the use of uniform topological priors are that the likelihood function will overwhelm any information in the topological priors anyway (see review in Lewis, 2001), and that the topological prior information is unimportant to the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) because it is identical in both the numerator and denominator, when all topologies are considered equally probable a priori. The conditions under which the former will be true in phylogenetic analyses have not been established. The latter, however, is necessarily true when tree topology is the hypothesis being evaluated. Because of this property of the Metropolis-Hastings algorithm, and because every tree is given an equal probability a priori, uniform priors are said to model ignorance effectively. However, it has long been established that no prior can be devised that models ignorance for all hypotheses simultaneously (Franklin, 2001; Kass and Wasserman, 1996; Walley, 1996). This applies to phylogenetics, when the hypothesis being evaluated is not the entire topology, but the presence of individual clades.

We will show that when uniform topological priors are stipulated, clade probabilities are not equal a priori. Specifically, the number of taxa in a clade, given the number of taxa in the entire analysis, affects the prior probability of that clade in a predictable way. Because of this, if the hypothesis being investigated is monophyly (i.e., the probability of the clade), uniform topological priors do not model ignorance, an undesirable property of a prior distribution when little is known a priori. While we do not argue that clade priors must be uniform, the clade priors that result from uniform topological priors are difficult to justify as reasonable in any case. Under these conditions, the claim that uniform topological priors do not influence results in a Bayesian framework is false.

2. Uniform topological priors and clade priors

Considering a pool of fully bifurcating, equiprobable, rooted trees for n taxa, the probability of a given clade of T taxa is equivalent to the probability of randomly choosing a tree containing that clade, or, considered another way, the sum of the (equal) probabilities of all trees containing that clade. Here, the probability of a clade is obtained by multiplying the number of rearrangements of that clade by the number of rearrangements of all taxa not in that clade, divided by the number of possible rooted trees for n taxa (see Eq. 1). This means that if all trees are considered equally probable, the probability of a clade is dependent on the number of taxa it contains, T, and the number of total taxa in the analysis, n.

Because a monophyletic group of T taxa is rooted, the number of rearrangements is the same as the number of rooted trees for T taxa (Felsenstein, 1978; see review in Swofford et al., 1996). This value is multiplied by the number of possible $n - T$ rearrangements (that do not compromise the monophyly of T). The denominator is simply the number of possible labeled trees for n taxa (as in Felsenstein, 1978). Therefore,

$$\frac{\left[\prod_{i=2}^{T} 2i-3\right]\left[\prod_{i=T+1}^{n} 2i-2T-1\right]}{\prod_{i=2}^{n} 2i-3}. \tag{1}$$

Eq. (1) calculates the probability of monophyly for T taxa, given that all possible rooted topologies are equally probable (see Formula 12 of Brown (1994) for a similar, independently derived formula. However, Brown's formula results in somewhat different values than those obtained here [see reported values therein]).

To demonstrate this more intuitively, consider a set of $n = 5$ taxa, A–E, for which there are 105 bifurcating, rooted trees. Considering the monophyly of $T = 3$ taxa,