

Commentary on ‘The Prior Probabilities of Phylogenetic Trees’ by Joel Velasco¹

James Justus
University of Texas, Austin
justus.phil@mail.utexas.edu

Bayesian methods have only recently been utilized within phylogenetics (Rannala and Yang 1996; Yang and Rannala 1997; Mau and Newton 1997 were some of the first). Just as there is controversy about specifying prior probabilities (“priors” hereafter) in disciplines where Bayesian methods are common, dispute about the proper specification of priors of evolutionary trees has emerged in phylogenetics (*e.g.* Pickett and Randle 2005; Brandley *et al.* 2006; Randle and Pickett 2006). On this issue, Joel argues:

- (i) so-called “uninformative” priors on tree topologies significantly bias the priors on clades, and therefore cannot be used without biasing their posterior probabilities;
- (ii) the bias in this case is defensible and desirable;²
- (iii) priors for trees should be determined from a model of taxa evolution;
- (iv) the model should be the Yule birth process.

I believe Joel has identified a clear error in the phylogenetic literature, so the following primarily consists of elaboration, extension, and squabbles.

1. Points (i) and (ii). Following Pickett and Randle (2005), Joel notes that uniform priors on tree topologies entail non-uniform priors on clades. Specifically, uniform tree topology priors for an analysis of n taxa entail clades of sizes close to 1 and n will be more probable than clades of sizes close to $\frac{n}{2}$. Uniform tree topology priors therefore bias for both small and large clade priors in Bayesian phylogenetic analyses. Joel’s analysis is thus a specific illustration of the general fact that choices of so-called “uninformative” priors depend crucially on the target parameter of the hypothesis and parameterization of the model concerned.

Although Pickett and Randle (2005) believe this is problematic,³ Joel convincingly argues that the posterior probability of a clade *should* decrease as its size increases. Without putting it explicitly in these terms, he shows that hypotheses about large clades in binary phylogenetic trees are logically stronger than hypotheses about small clades: a clade of size k ($k > 2$) entails some clade(s) of size $l < k$ exist (p. 21). The biased priors in favor of small clades reflect logical properties of clades and are therefore unproblematic.

¹ Page numbers and comments are based on an earlier draft (5-13-07).

² It is not, however, Joel’s intention to defend uniform tree topology priors on this basis.

³ Moreover, a consensus seems to be forming on this issue (see Randle *et al.* 2005).

He also explains that a sampling effect is responsible for the initially puzzling second bias: that the probability of size k clades increases as k exceeds $\frac{n}{2}$ and approaches the number of taxa being analyzed n . This follows from the fact that clade probability is inversely correlated with $\binom{n}{k}$, which decreases as $k \rightarrow n$ beyond $\frac{n}{2}$, and the assumption that all n taxa constitute a clade, which is made for all rooted trees. After noting this fact (pp. 21-22), Joel does not discuss it further, but it should be emphasized that this bias is generally *not* defensible as the other was because the assumption is generally indefensible. It is therefore justifiable, and in fact necessary, to correct for this bias in the prior probability distribution. Absent correction, the posterior probability of clades of sizes near the number of taxa being analyzed will likely be overestimated. Notice also that the severity of this bias will increase as the number of taxa analyzed does.

The explanation for this (unjustifiable) bias may also provide a response to a potential problem for his position Joel briefly considers (p. 24). The apparent problem is the high priors of large canonical clades such as Mammalia, Vertebrata, or Animalia given the sound basis for the (justifiable) bias against large clades discussed above. Joel, I believe correctly, identifies the source of the tendency towards high priors for these clades in the high degrees of statistical support, within a Bayesian framework or not, they have received in previous phylogenetic analyses.

His account of why clades close in size to the total number of taxa being analyzed have high priors, however, also furnishes a possible rationale for such high priors for these clades. If the set of taxa comprising all bacteria, for instance, constitute more than a majority of the total set of extant taxa, then its prior *should* be set higher than smaller sets of taxa. In this case, the assumption that all taxa being analyzed are a clade is true given our best theories of the origin of life, rather than an artifact of the study design. Similarly, if it is known that a particular set of n taxa share a common ancestor, if for example all mammals derive from a common ancestor, then the hypothesis that a subset of mammal taxa of size $> \frac{n}{2}$ is a clade should receive a higher probability than hypotheses for smaller subsets.

2. Points (iii) and (iv). Faced with a choice between setting priors (uniform or otherwise) for tree shapes, tree topologies, or labeled histories, Joel suggests the Yule process as the appropriate model of taxa evolution from which priors should be derived (p. 29):

In the phylogenetic case, the tree is a result of the biological process of common ancestry and descent with modification. We want to know the probability distribution that results when a tree is produced “at random” as a result of this process. Trees are the result of the sequences passing down from organism to organism via reproduction on the branches and splitting at the nodes when the organism gives rise to multiple offspring. In the abstract, it is perfectly captured by the Yule birth process in which particles reproduce with a constant probability of giving birth per particle per unit time.

The rationale for this prior choice is that the Yule process represents the simplest possible model of taxa evolution. This idea underlies Joel's claim that it constitutes a "null model of phylogeny," (p. 31) and that there is "near-universal acceptance of the Yule process being the underlying physical process for the recreation of evolutionary history," (p. 31). The Yule model entails a uniform distribution on labeled histories, rather than tree shapes or topologies. It should be recognized that uniform priors on labeled histories entail the same type of bias on clade priors as uniform priors on tree topologies. Thus, the unjustified bias discussed above (for clades of size close to the number of taxa being analyzed) should also be corrected when the Yule model is used to set priors (see §1).

There are at least three issues raised by this method of prior specification for phylogenetic trees.

- (a) Given the minimalism of this model, it is somewhat unclear that the rationale for setting priors it provides should be sharply distinguished from an indifference rationale, as Joel suggests (p. 30). After all, the Yule process represents random lineage splitting, and as such it represents a branching process that will not exhibit partiality towards any outcome tree. As Jeffreys would put it, the Yule model does not supply a "definite reason" for higher probabilities of some outcomes over others (see Kass and Wasserman 1996). If forced to specify a model of taxa evolution, for instance, someone favoring the principle of indifference could specify the Yule model precisely because it is appropriately indifferent about the branching process.
- (b) Among proposed theoretical virtues of Bayesian methods, such as that parameters can be treated as random variables or that the statistical support it provides is more easily interpretable than results of other types of statistical tests, one commonly cited virtue is that priors allow the integration of background knowledge that *should* contribute to the value of the posterior probability of a hypothesis. If background knowledge is available, it is therefore unclear that using a rather uninformative "null" model of taxa evolution, such as the Yule process, to set priors is desirable. Rannala and Yang (1996, 308) suggest, for example, that, "neither the simple birth-death process, nor a submodel of it, the Yule process, is likely to accurately describe the actual process of speciation and extinction, especially when we consider the additional effect of species sampling by biologists." Species sampling refers to the fact that most phylogenetic analyses consider only a subset of the total number of species in a clade. Joel, of course, appreciates this and is not claiming the Yule process is in any sense uniquely justified *a priori*. It is therefore important, however, to consider some modifications of and alternatives to the Yule process that improve upon it. One modification incorporates the sampling effect just mentioned. Another stems from the recognition that the so-called "molecular clock" assumption – that the rate of molecular evolution is constant across lineages – is often false. The corresponding required modification for the branching process allows branching rates to vary for different lineages (Thorne *et al.* 1998). Represented formally, this requires abandoning the constant rate of evolution λ of the Yule process, and instead

allowing rates of different values across lineages. Given that “close” branches are more likely to exhibit similar evolution rates, a positive correlation constraint on their rates is usually imposed. Yet another modification is to allow the rate of evolution itself to evolve over time along different lineages (Thorne *et al.* 1998). Specifically, it is biologically plausible that species that evolve quickly will give rise to species with similarly high speciation rates. Notice that these improvements on the Yule model do not depend upon information about the particular taxa being analyzed. The prior probability distributions they entail are therefore not taxon-specific and can be utilized in any Bayesian phylogenetic analysis.

- (c) Similar to (b), if there is information about the specific taxa being analyzed that is relevant to determining their phylogeny, it can be incorporated into the analysis with priors. Joel, however, claims that, “the interpersonal Bayesian ideal is not to build such evidence into the priors but rather, to operate ‘as if’ we were ignorant” (p. 13). But as long as the information is objective – for example, agreed to by all or almost all phylogeneticists – rather than merely subjective, the objective Bayesian phylogeneticist can argue that it *should* be used to set priors, in this case informative priors. Results from morphological, geographical, and fossil studies (whether Bayesian or not) can therefore be utilized to set informative priors (Randle and Pickett 2006). Joel suggests it is inappropriate to build this information into priors because, “If we wanted to know what we should believe based on the molecular and geographical evidence combined, both would be part of our data and for that we need some further procedure for combining different types of data (which Bayesian analysis is again ideally suited for)” (p. 13). Without further argument, however, it is unclear why codification into the prior is not a legitimate way of incorporating other types of data and study results into a Bayesian analysis. After all, the ability to incorporate background knowledge into a Bayesian analysis via priors is one of the merits commonly cited for Bayesianism over other methods, such as maximum likelihood methods.

References

- Bradley, M. C.; Leaché, A. D.; Warren, D. L.; McGuire, J. A. 2006. "Are Unequal Clade Priors Problematic for Bayesian Phylogenetics?" *Systematic Biology* **55**: 138-146.
- Kass, R. E. and Wasserman, L. 1996. "Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association* **91**: 1343-1370.
- Mau, B. and Newton, M. A. 1997. "Phylogenetic Inference for Binary Data on Dendrograms Using Markov Chain Monte Carlo." *Journal of Computational and Graphical Statistics* **6**: 122-131.
- Pickett, K. M. and Randle, C. P. 2005. "Strange Bayes Indeed: Uniform Topological Priors Imply Non-Uniform Clade Priors." *Molecular Phylogenetics and Evolution* **34**: 203-211.
- Randle, C. P.; Mort, M. E.; and Crawford, D. J. 2005. "Bayesian Inference of Phylogenetics Revisited: Developments and Concerns." *Taxon* **54**: 9-15.
- Randle, C. P., and Pickett, K. M. 2006. "Are Non-Uniform Clade Priors Important in Bayesian Phylogenetic Analysis? A Response to Bradley et al." *Systematic Biology* **55**: 147-151.
- Rannala, B. and Yang, Z. 1996. "Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference." *Journal of Molecular Evolution* **43**: 304-311.
- Thorne, J. L.; Kishino, H.; and Painter, I. S. 1998. "Estimating the Rate of Evolution of the Rate of Molecular Evolution." *Molecular Biology and Evolution* **15**: 1647-1657.
- Yang, Z. and Rannala, B. 1997. "Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method." *Molecular Biology and Evolution* **14**: 717-724.