Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution 45 (2007) 748–749
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Short Communication

Why non-uniform priors on clades are both unavoidable and unobjectionable

Joel D. Velasco

University of Wisconsin, Department of Philosophy, 5185 Helen C. White Hall, Madison, WI 53706, USA

Received 10 February 2007; revised 27 June 2007; accepted 10 August 2007
Available online 26 August 2007

Pickett and Randle (2005) showed that for $n > 4$ taxa, any prior probability distribution that is uniform with respect to tree topologies induces a non-uniform distribution on clades. Steel and Pickett (2006) then strengthened this result by proving that any label invariant distribution on trees induces a non-uniform prior on clades, but left open the possibility of a uniform distribution on clades that is not label invariant. The point of this letter is to show that uniform priors on clades are impossible on any probability distribution on topologies (with five or more leaf taxa) and that this fact is in no way a bad consequence for Bayesian phylogenetic practice.

Any arbitrary group of taxa is a possible clade and Pickett and Randle (2005) contend that all such groups regardless of their size should have the same prior probability of forming an actual clade on the true tree. Recall that the number of possible clades of size x is just the number of possible ways of choosing a group of size x from the collection of n taxa which is just n choose $x = \frac{n!}{x!(n-x)!}$. The fact it is impossible for each of these possible clades to have the same probability of forming an actual clade can easily be seen to follow from these two elementary facts: (1) since smaller clades are nested inside larger clades, on any tree (and therefore on the true tree), there are at least as many actual clades of size two as there are of size three and (2) there are many more possible clades of size three than of size two. So not all possible clades of size two or three could be equally probable and *a fortiori* not all possible clades can be equally probable. This result is perhaps more easily seen by attending to the following two graphs in Fig. 1.

The first graph depicts how the size of a clade determines the number of possible clades of that size. I have used 50 taxa as an example, but the shape of the curve is the same for any number of taxa. Notice that the scale is logarithmic meaning that there are vastly more possible clades of size 25 than, say, size 10. The second graph depicts how the size of a clade determines the expected number of clades of that size on a uniform prior distribution on topologies. Since the probability of a clade is just the expected number of clades of that size divided by the number of possible clades of that size (assuming all clades of the same size have the same probability), if the probability of a clade is to be the same for every size, these two curves must have the exact same shape (one should be the other multiplied by a constant—the probability). Notice that the "expected clades" curve is calculated under a uniform prior on topologies (as in Pickett and Randle, 2005)—for other topology distributions the curve varies in shape slightly, but a few aspects remain constant such as the fact that its peak must be at size 2. Since no distribution on topologies gives it the same shape as the "possible clades" curve, the probabilities of all possible clades can never be identical.

Formally, the theorem is the following: for $n > 4$ taxa, there is no probability distribution over rooted binary phylogenetic trees which assigns equal probability to each non-trivial clade (clades of size 2 through size $n - 1$).

Proof. For n taxa, there are a total of n choose $2 = \frac{n!}{2!(n-2)!}$ possible clades of size two while there are n choose $3 = \frac{n!}{3!(n-3)!}$ possible clades of size three. Therefore, for $n > 5$, there are more possible clades of size three than of size two. If each possible clade of size three or size two had the same probability of being an actual clade, there would be more expected clades of size three than of size two which is impossible on any bifurcating tree and therefore on any distribution over trees (since each clade of size three entails the existence of a distinct clade of size two).

For the special case of $n = 5$, there are 10 possible clades of size two and 10 of size three. If they were each assigned equal probability, the expected number of clades of size two and size three would have to be equal. This is possible

E-mail address: jdvelasc@wisc.edu

1055-7903/$ - see front matter
doi:10.1016/j.ympev.2007.08.003

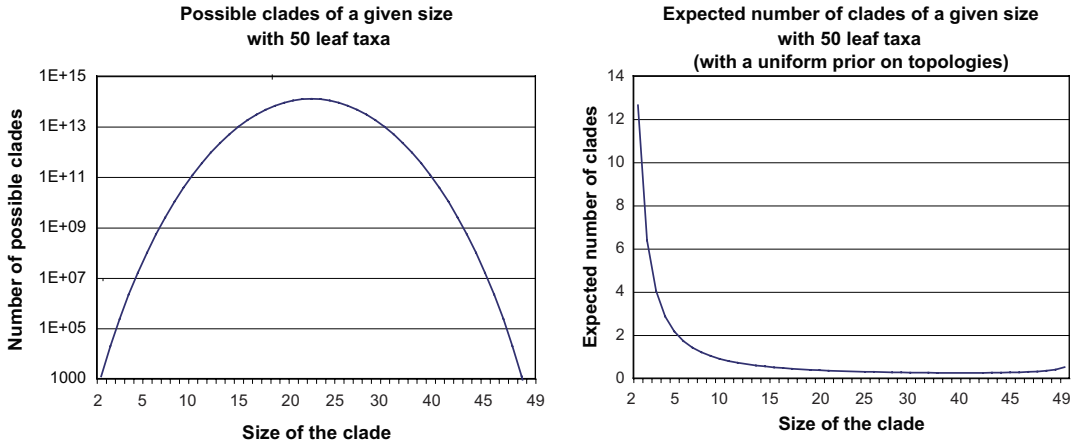

Fig. 1. Two graphs comparing the number of possible clades of a given size to the expected number of clades of that size. The expected number of clades (the value on the right) divided by the number of possible clades (the value on the left) is the probability.

only if the only tree topologies with positive probability have the pectinate shape, which has nested clades of size 2, 3, and 4. On this distribution, there are equally many clades of size two as size three, but there are also equally many clades of size four. Since there are only 5 possible clades of size four, on this distribution they cannot have the same probability as the clades of size two and three.

As Steel and Pickett (2006) point out, this theorem does not hold if we allow non-binary branching trees. But by thinking about this graphically, one can see that to accept the claim that all possible clades are equally probable, not only would multifurcating trees have to be possible; in addition, they would have to be extremely probable. There would have to be many more clades of larger sizes than smaller sizes. Clades of size two would have to be the least common of all. For 50 taxa, a polytomy of size 25 would be over 100 trillion times as probable as a resolved clade of that size.

It would be a mistake to believe that this argument shows that prior probability distributions cannot capture ignorance with respect to clades and therefore cannot be used. There is no such ignorance to capture; we are not ignorant of important logical and biological facts connecting clade sizes. Under extremely weak assumptions about the relative frequency of bifurcations versus multifurcations, it is actually inconsistent to hold that all clades are equally probable regardless of size. Our background knowledge suffices to tell us that the number of taxa in a group must be relevant to its probability of being monophyletic.

References

Pickett, K.M., Randle, C.P., 2005. Strange bayes indeed: uniform topological priors imply non-uniform clade priors. Mol. Phylogenet. Evol. 34, 203–211.

Steel, M., Pickett, K.M., 2006. On the impossibility of uniform priors on clades. Mol. Phylogenet. Evol. 39, 585–586.