Available online at www.sciencedirect.com
SCIENCE DIRECT
Molecular Phylogenetics and Evolution 39 (2006) 585–586
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Letter to the editor

On the impossibility of uniform priors on clades

Pickett and Randle (2005) showed that a uniform prior distribution on rooted binary phylogenetic trees (which assumes all these trees are equally probable) gives rise to a non-uniform probability on clades (i.e., the probability distribution on clades varies with clade size), except in the special case where the trees have four or fewer leaves. This result means that the probability that a particular strict subset S of two or more species forms a clade in the tree varies with the number of species in S. In this note, we show that this phenomenon is not specific to a uniform distribution on trees—a distribution that is also often called the 'Proportional-to-Distinguishable-Arrangements' (PDA) model (Rosen, 1978). It turns out that any 'reasonable' prior distribution on trees with more than four leaves induces a probability distribution on clades (by which we mean non-trivial clades of size between 2 and $n-1$, where n is the number of taxa in the tree) which must vary with the clade size. In the special case of four-leaf trees, the PDA model is the only 'reasonable' prior that gives a uniform distribution on clades. By 'reasonable,' we simply mean that the probability of a tree depends just on its shape, not on how the leaves are labelled (this concept, which applies to other commonly used tree priors, such as the Yule model, was referred to as 'label invariance' in Steel and Penny, 1993; and 'exchangeability' in Aldous, 1996). At the end of this letter, we briefly discuss the significance of our result, which can be stated formally as follows.

Theorem 1.

(i) *For $n > 4$ there is no label-invariant prior on binary rooted phylogenetic trees with n leaves that induces a uniform distribution on clades.*
(ii) *The PDA model is the only label invariant prior on rooted binary phylogenetic trees with four leaves that induces a uniform distribution on clades.*

Proof. In the proof, we denote the leaves (species) of the tree by the elements of the set $\{1,2,\ldots,n\}$, and we will let p_{12} be the probability that $\{1,2\}$ is a clade in a tree T randomly generated by a label-invariant prior. Similarly, we let p_{123} be the probability that $\{1,2,3\}$ is a clade in a tree T randomly generated by a label-invariant prior.

Part (i). By label invariance, p_{123} is also the probability that $\{1,2,j\}$ is a clade of T for any particular choice of j from $\{3,\ldots,n\}$. Note that $\{1,2,j\}$ and $\{1,2,k\}$ cannot both be clades in T when j is different from k (i.e., these events are mutually exclusive). Thus, the probability P^* that there exists some j from $\{3,\ldots,n\}$ for which $\{1,2,j\}$ is a clade of T is given by

$$P^* = (n-2)p_{123}. \tag{1}$$

As there are three rooted binary trees on leaf set $\{1,2,j\}$ only one of which contains the clade $\{1,2\}$, it follows by label invariance that p_{12} is at least 1/3 times the probability that there exists some leaf j from $\{3,\ldots,n\}$ for which $\{1,2,j\}$ is a clade of T (p_{12} may be strictly larger than p_{123}, as we will see below, however all we require at this stage is the inequality). That is, $p_{12} \geqslant P^*/3$. By Eq. (1) this gives

$$p_{12} \geqslant \frac{n-2}{3} \times p_{123}. \tag{2}$$

Now, if $n > 5$, then $(n-2)/3 > 1$, and so, by Eq. (2), we have $p_{12} > p_{123}$ which means that the induced prior on clades is non-uniform. This establishes part (i) for any value of $n > 5$, but does not establish the $n = 5$ case.

For this remaining case ($n = 5$), note that $p_{12} > P^*/3$ precisely if the label-invariant prior confers a positive probability that T contains a clade of size 2 that is not contained in some clade of size 3, in which case Eq. (2) is also a strict inequality. Assume there is a label-invariant prior distribution on rooted binary phylogenetic trees on five leaves that seeks to induce a uniform distribution on clades. Then this prior must assign probability 0 to any tree that has a clade of size 2 that is not contained in a clade of size 3 (otherwise, $p_{12} > p_{123}$). This property applies to two of the three possible shapes of rooted binary trees on five leaves—the only one remaining shape is the tree that has a pectinate shape, with nested clades of size 2, 3, 4, 5, and there are 60 trees with this shape. By label invariance, each of these 60 trees must have probability 1/60. There are six ways to rearrange 3, 4, and 5 beneath clade $\{1,2\}$, and

1055-7903/$ - see front matter
doi:10.1016/j.ympev.2005.10.005

there are two rearrangements of 4 and 5 subtending the three rearrangements of clade $\{1,2,3\}$. Therefore, by a straightforward counting argument, $p_{12} = p_{123} = 0.1$. However, we also have $p_{1234} = 0.2$, because 12 of the 60 trees have the clade $\{1,2,3,4\}$, contradicting the assumption that the prior induces a uniform distribution on clades. This completes the proof of part (i).

Part (ii). Let p_1, p_2 denote the probability of the prior producing a tree with a pectinate or fork shape, respectively. For $n = 4$, there are three fork-shaped trees, and 12 pectinate trees. Among the pectinate trees, same-labelled clades of 2 and 3 leaves are duplicated and triplicated, respectively. Therefore, a straightforward counting argument shows that

$$p_{12} = \tfrac{1}{6}p_1 + \tfrac{1}{3}p_2$$

and

$$p_{123} = \tfrac{1}{4}p_1 + 0p_2$$

and so, since $p_1 + p_2 = 1$, we have $p_{12} = p_{123}$ if and only if $p_1 = 4/5$, and $p_2 = 1/5$. The only label-invariant prior with this property is the PDA model (e.g., Semple and Steel, 2003). This completes the proof of part (ii). □

The above theorem does not hold for priors on trees that allow non-binary (i.e., multifurcating) trees to have positive probability; however, standard priors on trees generally assign probability zero to any non-binary tree.

The impossibility of uniform clade priors is relevant for subjective Bayesian approaches in phylogenetics because often the interest in such studies is in evaluating the support for particular clades, and so any dependence of the prior probability of any such clade on its size could influence results. Existing approaches generally invoke a label-invariant prior (such as the PDA or Yule model) as there is no reason a priori to treat any taxon differently from any other one. Our result implies that an unavoidable consequence of such an approach is that the prior probability of any specific group of taxa forming a clade must vary with the size of that clade (small clades being more likely than moderate-sized ones) and this size bias cannot be rectified by simply adjusting the (label-invariant) prior. This does not preclude the possibility that other priors could lead to a uniform distribution for clade sizes. For example, the use of empirical priors from the topologies of previously published analyses of independent data are not constrained by the result above, since such priors will not, in general, be label-invariant.

References

Aldous, D., 1996. Probability distributions on cladograms. In: Aldous, D., Pemantle, R. (Eds.), Random Discrete Structures, IMA Volumes in Mathematics and its Applications, vol. 76. Springer-Verlag, New York, pp. 1–18.

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

Rosen, D.E., 1978. Vicariant patterns and historical explanation in biogeography. Syst. Zool. 27, 159–188.

Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford.

Steel, M.A., Penny, D., 1993. Distributions of tree comparison metrics—some new results. Syst. Biol. 42 (2), 126–141.

Mike Steel
Biomathematics Research Centre,
University of Cantebury,
Private Bag 4800, Christchurch,
New Zealand
E-mail address: m.steel@math.canterbury.ac.nz

Kurt M. Pickett
Division of Invertebrate Zoology,
American Museum of Natural History,
Central Park West at 79th Street, New York, NY,
USA
E-mail address: kpickett@amnh.org

Received 4 August 2005; revised 30 September 2005;
accepted 3 October 2005
Available online 14 November 2005