Palaeontology The Palaeontological Association www.palass.org

[Palaeontology, 2018, pp. 1–6]

DISCUSSION

PARSIMONY AND MODEL-BASED PHYLOGENETIC METHODS FOR MORPHOLOGICAL DATA: COMMENTS ON O'REILLY *ET AL.*

by PABLO A. GOLOBOFF, AMBROSIO TORRES GALVIS *and* J. SALVADOR ARIAS

Unidad Ejecutora Lillo, Miguel Lillo 251, 4000, S.M. de Tucumán, Argentina; pablogolo@csnat.unt.edu.ar

O'Reilly *et al.* (2017) recently published a paper recommending the use of current model-based programs for the analysis of discrete morphological data; preceded by O'Reilly *et al.* (2016) and Puttick *et al.* (2017*a*, *b*), this paper is based on the same datasets simulated for their previous work, but now analysed collapsing poorly supported groups (as recommended by Brown *et al.* 2017 and Goloboff *et al.* 2017). Goloboff *et al.* (2017) noted that the simulations in O'Reilly *et al.* (2016) and Puttick *et al.* (2017*a*), just like those of Wright & Hillis (2014) before, generated their datasets using branch lengths common for all characters; that is, with all their characters increasing their probability of change at the same branches, by the same exponential factor (with the 'length' of the branch being the combination of time and the instantaneous rate of change along the branch; see e.g. Swofford *et al.* 1996, p. 439; Felsenstein 2004, p. 147–148). This model, long known (since Felsenstein 1978) to be problematic for parsimony, is often used for DNA sequences (where it may be justified, e.g. by assuming neutrality). However, there is no evidence or theoretical argument that this model can be generally applied to morphological data, and consideration of suites of characters in different groups suggests that it cannot. It seems obvious that characters related to the forelimb evolve faster in bats than in rodents, and that the opposite is true of dental and mandibular characters; every taxonomist specialized in any group can surely think of similar examples. The homogeneous Markov model assumes (when applied to morphology) that the characters are like units that can simply switch into one or another state at any point in the tree. But the very fact that taxonomists beginning to investigate a group first need to learn the relevant characters speaks against the idea that all groups can be classified by looking at the same sets of characters randomly changing over all the tree; that is why someone who has worked extensively on spider morphology needs to learn a whole new suite of anatomical characters if starting now to work on, say, beetles. It is true that some branches of the tree of life are much longer than others (in terms of morphological distance), and that branch lengths are a crucial component of the homogeneous Markov model. But even these differences in branch lengths do not seem to follow the expectations of the model, as it is obvious that only some groups of characters change at those long branches. Consider cetaceans and chiropterans: the branch leading to each of those groups has dozens of characters changing, but no biologist would expect that the 'length' of the cetacean branch makes it more likely that some of the chiropteran synapomorphies would show up there. Goloboff *et al.* (2017, p. 27) argued that: (1) without reliable models for the evolution of discrete morphological characters, parsimony seems to be a reasonable alternative, in not striving to provide a statistical justification of results; and (2) while all methods make assumptions, the assumptions needed for non-statistical justification of a method are much less restrictive, with parsimony generally considered to be justified on the grounds of descent with modification alone (e.g. Farris 1983, 2008). Thus, Goloboff *et al.* (2017) performed alternative simulations in which the probability of character substitution does not change in concert along tree branches; the methods of phylogenetic inference that assume the concerted variation of the probability of change for all characters perform more poorly in that case, performing similarly to equally weighted parsimony, and much worse than implied weights parsimony (Goloboff 1993). None of that should be surprising, as it is in accord with theory, but O'Reilly *et al.* (2017) now make some comments questioning our results. Despite its brevity, the summary of the findings and criticisms of Goloboff *et al.* (2017) by O'Reilly *et al.* (2017) contains several

doi: 10.1111/pala.12353

inaccuracies and misleading arguments, which we discuss in this contribution.

The main shortcoming with O'Reilly *et al.*'s (2017) discussion of our paper we cannot quote, because it is in what they do *not* say. O'Reilly *et al.* (2017) say nothing about the criticism of their model which Goloboff *et al.* (2017) considered most damning: namely that *all* characters (even if in different rate categories) have their probability of change increased, in concert, at the same branches of the tree, exponentially depending on the length of the branch. Unsurprisingly, the model trees of Puttick *et al.* (2017*a*) which produce the datasets where parsimony performs most poorly relative to maximum likelihood or bayesian analyses are the asymmetric trees, which (by virtue of being ultrametric; see below) have the greatest disparity in branch lengths. Thus, the results defended by O'Reilly *et al.* (2017) strongly depend on generating data with this model; that they produced minor violations of the Mk model (Lewis 2001) used to analyse the data (e.g. recoding some purines/pyrimidines as binary) does not change the fact that all characters have their probabilities of change increased or decreased at the same branches of the tree (Goloboff *et al.* 2017, p. 3). Goloboff *et al.* (2017) showed that, if this assumption is relaxed, then the advantage of methods based on presupposing such uniformity vanishes and (for rates of homoplasy distributed as in real datasets) implied weights parsimony performs better than likelihood or bayesian methods. There is no mention of any of this in the reply of O'Reilly *et al.* (2017), who instead attribute the differences in results to the relative frequencies of characters with different amounts of homoplasy, when in fact these are quite similar in both studies (see below, and Goloboff *et al.* 2017, fig. 1a, c).

But in addition to not discussing points Goloboff *et al.* (2017) actually raised, O'Reilly *et al.* (2017) also rebut points that Goloboff *et al.* (2017) never made. O'Reilly *et al.* (2017) start by stating that:

> Goloboff *et al.* (2017) have criticized this approach to simulating morphology-like datasets on the basis that our generating trees encompass only contemporaneous taxa, assume that evolutionary rates are constant across time and the tree… However, our experiments do not attempt to simulate non-contemporary taxa…

Puttick *et al.* (2017*a*) presented as one of their most important findings the difficulty of accurately retrieving nodes closer to the root in asymmetric trees (with all nodes of symmetric model trees retrieved effectively). Figure 1 presents trees like those used by Puttick *et al.* (2017*a*) and O'Reilly *et al.* (2017) as models (for simplicity, we show 16-taxon trees; they used 32-taxon trees with the same characteristics). Branches are drawn proportionally to their length. These trees, and the resulting data, are clocklike and ultrametric. The degree of difference between the members of a clade and any taxon outside the clade is expected to be exactly the same, obeying the ultrametric inequality $d(x,y) \leq \max\{d(x,z), d(y,z)\}$. Thus, for Fig. 1, the distance between taxa *a* and *i* is expected to be the same as that between *a* and *j*−*p*, and the expected amount of evolution between the root *r* and each of the terminal taxa is identical. This is the same situation assumed by phenetic methods, so the way that O'Reilly *et al.* (2017) have modelled their data also means that phenograms produce an excellent recovery of the model tree. To show that this is indeed the case, the asymmetric tree of Figure 1 indicates the frequencies of recovery of the different partitions (simulating 100 datasets under a two-state single rate JC-Neyman model, with 100 characters, using the 'upgma' command of PAUP* for producing the phenogram); the highest frequencies are always those for phenograms, which retrieve the correct basal split much more often than Bayesian or parsimony trees. The inferred phenograms are closer to the model tree, when measured with statistics of tree-distance that are not as affected as the Robinson-Foulds (1981) distances by long moves of a single taxon, such as the distortion coefficient DC of Farris (1973) modified to be symmetric (see Goloboff *et al.* 2017, p. 5); the average distortion for the phenogram is 0.171, while for Bayesian and parsimony trees it is 0.556 and 0.343, respectively. If their model is not to be seriously considered as favouring phenetics (which few would argue for), why is it to be considered to be relevant for choosing between Bayesian analysis and parsimony? But Goloboff *et al.* (2017), in fact, never complained about 'modelling only contemporaneous taxa' or 'constant evolutionary rates'; instead they criticized the fact that conclusions resulting from such a restrictive model were presented by Puttick *et al.* (2017*a*) as if they were general. Goloboff *et al.* (2017) explicitly pointed out that:

1. The difficulty of recovering deep nodes in asymmetric trees resulted only from the restrictive clock assumption, which leads to the earliest splitting branches in asymmetric trees being much longer than the others, and thus harder to place accurately.
2. The conclusion that deep nodes of asymmetric trees are harder to recover goes against theoretical expectation in any time-reversible model, where the 'depth' of nodes is irrelevant, as the tree could be rerooted so that the 'deepest' node becomes the 'shallowest' without any changes in likelihood or parsimony. In other words, if the model tree in Figure 1A was rerooted so that *p* is now the sister of all other taxa, and all branch lengths were preserved, then the resulting data would no longer be clocklike on the model tree, but now it would be the 'shallowest' nodes (instead of the 'deepest') that would be hardest to retrieve accurately.

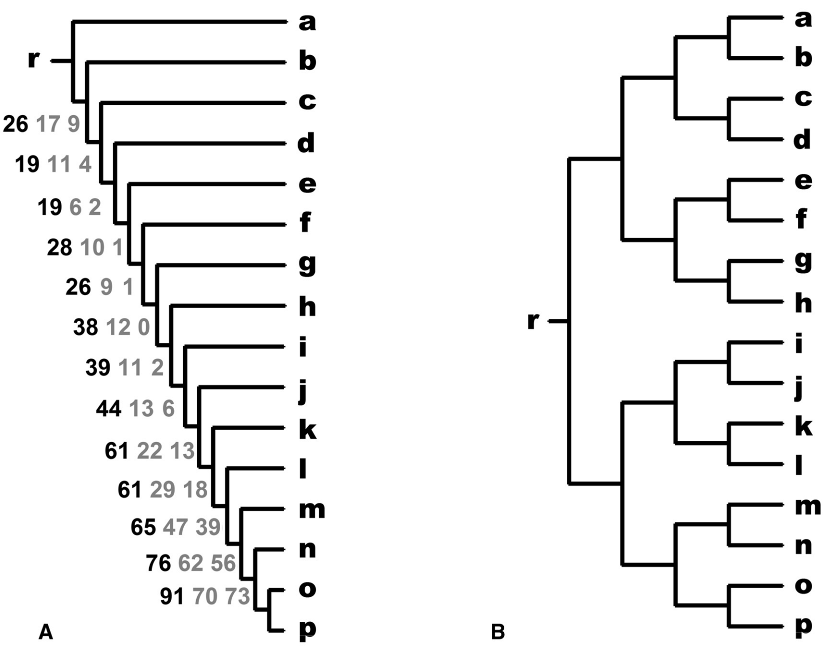

FIG. 1. Model trees, similar to those used by O'Reilly *et al.* (2017) and Puttick *et al.* (2017*a*). A, asymmetric (pectinate) tree (numbers at internal nodes indicate frequency of recovery when simulating datasets with a 2-state single rate JC/Neyman model, for phenogram, parsimony and Bayesian trees). B, symmetric (balanced) tree.

Goloboff *et al.* (2017) also mentioned that Puttick *et al.* 2017*a*) had not been explicit in their paper about the use or implications of such a strong assumption (which is evident only from their R scripts). As another thought experiment to show (*contra* Puttick *et al.* 2017*a* and O'Reilly *et al.* 2017) that asymmetric trees are harder to reconstruct, consider Figure 2. If the branch lengths for the symmetrical and asymmetrical trees had been chosen as in Figure 2 (thus no longer being clocklike), then the conclusions as to which tree shape is harder to reconstruct would have been exactly the opposite. A proper attempt to test the influence of tree shapes would need to eliminate the effect of differences in branch length in the two trees, which they never do. As Goloboff *et al.* (2017) put it: 'the right conclusion would then be that the hardest nodes to recover are those around very long branches, not the deep ones'.

In addition to overlooking the details just discussed, O'Reilly *et al.* (2017) also implied that because their 'experiments do not attempt to simulate non-contemporary

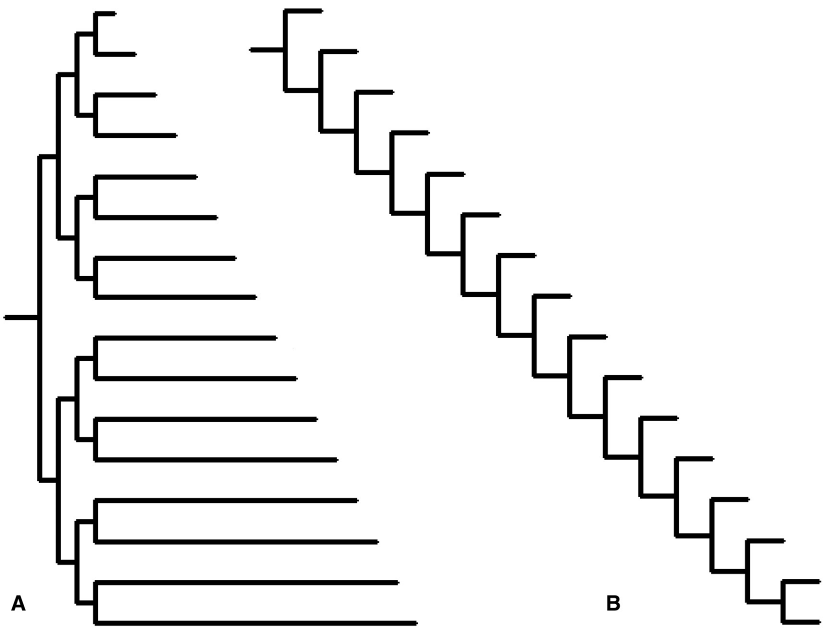

FIG. 2. Alternative branch lengths of model trees, for which (contrary to O'Reilly *et al.* 2017) symmetric trees are harder to recover than asymmetric trees. A, symmetric (balanced) tree. B, asymmetric (pectinate) tree.

taxa', then Goloboff *et al.* (2017) unfairly criticized them for not having considered fossil taxa in their simulations. But Goloboff *et al.* (2017) did not take Puttick *et al.* (2017*a*) to task for not having considered fossils; rather, Goloboff *et al.* (2017) took Puttick *et al.* (2017*a*) to task for having explicitly extended their conclusions to trees with fossil taxa, when they had clearly not modelled their data appropriately. Irrespective of whether Puttick *et al.* (2017*a*) intended their simulations to be of relevance for palaeontologists, the following quotations from their paper (italics added) lead readers of the journal *Palaeontology* to believe they may be:

> The ***fossil*** record affords the only direct insight into evolutionary history of life on the Earth, but the incomplete preservation and temporal distribution of ***fossils*** has long prompted biologists to seek alternative perspectives, such as molecular phylogenies of living species, eschewing palaeontological evidence altogether. However, there is increasing acceptance that analyses of historical diversity cannot be made without phylogenies that incorporate ***fossil*** species and calibrating molecular phylogenies to time cannot be achieved effectively without recourse to the ***fossil*** record. Integrating ***fossil*** and living species has become the grand challenge (p. 1)

> All phylogenetic methods also performed best when attempting to recover a symmetrical target tree; all methods found recovery of asymmetrical trees challenging and phylogenetic accuracy diminished from tip to root. The impact of tree topology is of particular concern since empirical phylogenetic trees are invariably asymmetric, and trees of ***fossil*** species are infamous for their asymmetry (p. 6).

O'Reilly *et al.* (2016) started their paper similarly, invoking the need to have methods for reliable phylogenetic placement of fossils. It was in regard to the second quotation above, on the difficulty of recovering deep nodes being of particular concern in the case of trees with fossil taxa, that Goloboff *et al.* (2017) noted that the condition which leads to that difficulty in asymmetric trees (i.e. having exactly the same amount of evolution from root to every terminal taxon), is especially unlikely to apply in the case of trees including taxa of different ages, and thus 'not only is Puttick *et al.*'s (2017*a*) difficulty in recovering some groups the result of very long branches instead of tree shape, it is also irrelevant for the situation they claim it affects the most: fossil phylogenies' (Goloboff *et al.* 2017, p. 14).

O'Reilly *et al.* (2017) continued by stating that Goloboff *et al.* (2017) claimed 'that [their] measure of biological realism, the spread of homoplasy exhibited by datasets, is inadequate.' It is not entirely true that Goloboff *et al.* (2017) made that criticism, or that the spread of homoplasy was the main measure of 'biological realism' of Goloboff *et al.* (2017). Goloboff *et al.* (2017) did note that O'Reilly *et al.* (2016) and Puttick *et al.* (2017*a*) had used only the consistency index CI to filter datasets by amounts of homoplasy, and that using the CI does not guarantee that the distribution of homoplasy will adjust to the distribution observed in real datasets (because different distributions can produce the same CI). But Goloboff *et al.* (2017) showed that, despite their having used only the CI to filter datasets, the distribution of homoplasy in O'Reilly *et al.* (2016) datasets was indeed quite similar to that observed in real datasets (Goloboff *et al.* 2017, p. 3–4 and fig. 1A). So, the way in which O'Reilly *et al.* (2016) and Puttick *et al.* (2017*a*) had evaluated the distribution of homoplasy in their datasets was only a minor point in Goloboff *et al.* (2017), and since this distribution is about the same in all these studies, Goloboff *et al.* (2017) had no quarrel with this aspect of their simulation. The criticism actually made is that the proportions of characters with 1, 2, . . . *n* steps over all the tree may be about the same as in Goloboff *et al.*'s (2017) datasets, but those steps will be concentrated (or diluted) on the same branches of the tree in the datasets of Puttick *et al.* (2017*a*) and O'Reilly *et al.* (2017). It is that differential concentration of changes on certain branches what causes the model-based methods to perform better than parsimony, and it is what Goloboff *et al.* (2017) considered the most unrealistic aspect of their simulations and all the simulations that employ a similar model.

One of the most paradoxical implications of the comments of O'Reilly *et al.* (2017) is that generating datasets with a realistic model may nonetheless result in unrealistic datasets:

> Our review of their datasets indicates that, while Goloboff *et al.* (2017) drew characters from an empirically realistic global distribution of homoplasy, their simulated datasets are not individually empirically realistic, with many matrices dominated by characters with very high consistency and an unrealistically small proportion of characters exhibiting high levels of homoplasy.

O'Reilly *et al.* (2017) did not specify how they 'reviewed' the datasets of Goloboff *et al.* (2017), but in any case this statement misses the point of simulations under a stochastic model. It is perfectly possible (either with Puttick *et al.*'s (2017*a*) or Goloboff *et al.*'s (2017) model) for a matrix with no homoplasy whatsoever to be generated; no doubt with a very low probability, but it could happen. The precise point of using simulations is that this will produce matrices with varying characteristics, and repeating the process many times is what allows us to study the variability under the model assumed. As Goloboff *et al.* (2017) reported having simulated ~9000 matrices, matrices which depart from the expectation at the 5% level must have been produced in 5% of the cases. Since Goloboff *et al.* (2017, fig. 1c) showed the distributions of homoplasy in the matrices resulting from simulation, it is possible that O'Reilly *et al.* (2017) based their statement above on an inspection of this figure (which in fact approaches the distribution observed in 158 empirical datasets). Their claim that many of the matrices generated by Goloboff *et al.* (2017) have unrealistically low amounts of homoplasy is without foundation.

Finally, it seems that it is O'Reilly *et al.* (2017) themselves who inadvertently provide the strongest argument in favour of implied weighting:

> The datasets simulated by Goloboff *et al.* (2017) have qualities that strongly *bias* in favour of parsimony phylogenetic inference, and implied-weights parsimony in particular, as the presence of large numbers of characters that are congruent with the tree allows implied weights to increase the power of these 'true' congruent characters. This effect will not be possible when increased levels of homoplasy are present or when the true tree is unknown (as is the case for all empirical datasets). (italics added)

The first problem with this statement is in the use of the negative word 'bias'. This usage is equivalent to saying that 'if the probability of change for all characters on each tree branch is the same, then this produces a *bias* in favour of maximum likelihood'. The proper expression would be that, under those circumstances, maximum likelihood is expected to recover the correct tree. Rephrasing in this way leads to the following deduction about implied weighting: when sampling datasets from an admittedly realistic distribution of homoplasy, implied weighting is expected to perform better than other methods.

The final claim of O'Reilly *et al.* (2017), that the beneficial effect of implied weighting will not occur 'when the true tree is unknown', making the simulations of Goloboff *et al.* (2017) irrelevant for empirical analyses, is hard to interpret. During the simulations of Goloboff *et al.* (2017) the 'true' tree was unknown to the tree-search algorithms, thus producing exactly the same situation as in empirical analyses. That is also the case for all simulation studies (including those of O'Reilly *et al.*), and thus it is hard to see how they could have thought that 'knowing the true tree' could have unfairly favoured implied weights in the simulations of Goloboff *et al.* (2017).

In summary, all of Goloboff *et al.*'s (2017) conclusions and criticisms of O'Reilly *et al.* (2016) and Puttick *et al.* (2017*a*) continue to be valid. The superior performance of model-based methods in their studies hinges on a model not proven to be generally applicable to morphological characters across taxa and anatomical systems. And, as Goloboff *et al.* (2017) noted, the use of simulated datasets alone cannot solve that problem of model adequacy; empirical tests of whether morphological data fulfill the crucial assumptions of the model are required as well.

Acknowledgements. We wish to thank comments and suggestions by Joseph Brown, Santi Catalano, Mike Pittman, Claudia Szumik, and two anonymous referees. We also acknowledge funding from CONICET (PUE 0070 to PAG), and the Research Grant Council of Hong Kong's General Research Fund (17103315 to Michael Pittman).

Editor. Andrew Smith

REFERENCES

BROWN, J. W., PARINS-FUKUCHI, C., STULL, G. W., VARGAS, O. M. and SMITH, S. A. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration. A comment on Puttick et al. *Proceedings of the Royal Society B*, **284**, 20170986.

FARRIS, J. 1973. On comparing the shapes of taxonomic trees. *Systematic Zoology*, **22**, 50–54.

—— 1983. The logical basis of phylogenetic analysis. *In* PLATNICK, N. I. and FUNK, V. A. (eds). *Advances in cladistics II*. Columbia University Press, 7–36.

—— 2008. Parsimony and explanatory power. *Cladistics*, **24**, 825–847.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**, 401–410.

—— 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA, 664 pp.

GOLOBOFF, P. 1993. Estimating character weights during tree search. *Cladistics*, **9**, 83–91.

—— TORRES, A. and ARIAS, J. S. 2017. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics*, published online 4 June. https://doi.org/10.1111/cla.12205

LEWIS, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, **50**, 913–925.

O'REILLY, J. E., PUTTICK, M. N., PARRY, L. A., TANNER, A. R., TARVER, J. E., FLEMING, J., PISANI, D. and DONOGHUE, P. C. J. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biology Letters*, **12**, 20160081.

—— —— PISANI, D. and DONOGHUE, P. C. J. 2017. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology*, **61**, 105–118.

PUTTICK, M. N., O'REILLY, J. E., TANNER, A. R., FLEMING, J. F., CLARK, J., HOLLOWAY, L., LOZANO-FERNANDEZ, J., PARRY, L. A., TARVER, J. E., PISANI, D. and DONOGHUE, P. C. 2017*a*. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B*, **284**, 20162290.

—— O'REILLY, J. O., OAKLEY, D., TANNER, A., FLEMING, J., CLARK, J., HOLLOWAY, L., LOZANO-FRENANDEZ, J., PARRY, L., TARVER, J., PISANI, D. and DONOGHUE, P. C. 2017*b*. Parsimony and maximum-likelihood phylogenetic analyses of morphology do not generally integrate uncertainty in inferring evolutionary history: a response to Brown et al. *Proceedings of the Royal Society B*, **284**, 20171636.

ROBINSON, D. and FOULDS, L. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

SWOFFORD, D., OLSEN, G., WADELL, P. and HILLIS, D. 1996. Phylogenetic inference. 407–514. *In* HILLIS, D., MORITZ, C. and MABLE, B. (eds). *Molecular systematics*, 2nd edn. Sinauer, Sunderland, MA.

WRIGHT, A. and HILLIS, D. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS One*, **9**, e109210.