Palaeontology
The Palaeontological Association
www.palass.org
[Palaeontology, Vol. 61, Part 4, 2018, pp. 631–635]

DISCUSSION

EMPIRICAL REALISM OF SIMULATED DATA IS MORE IMPORTANT THAN THE MODEL USED TO GENERATE IT: A REPLY TO GOLOBOFF *ET AL.*

by JOSEPH E. O'REILLY[1], MARK N. PUTTICK[1,2], DAVIDE PISANI[1,3] *and* PHILIP C. J. DONOGHUE[1]

[1]School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK; joe.oreilly@bristol.ac.uk, mark.puttick@bristol.ac.uk; davide.pisani@bristol.ac.uk, phil.donoghue@bristol.ac.uk
[2]Department of Biochemistry & Biological Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, UK
[3]School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK

FOSSILS provide our only direct insight into the history of life and realizing their evolutionary significance invariably requires that they are integrated into a phylogeny with their living and/or fossil relatives. There are many competing approaches to phylogeny estimation and, episodically, debate over their relative efficacy has erupted into controversy, as exemplified by the introduction of cladistics into palaeontology (Hull 1988). While there have been skirmishes on the role of stratigraphy in phylogeny estimation (Fox *et al.* 1999; Smith 2000; Fisher *et al.* 2002; Wagner 2002) parsimony has since achieved hegemony despite the introduction and implementation of a model-based approach to the analysis of morphological data (Lewis 2001). Increasingly, over the last few years, palaeontologists have performed parallel phylogenetic analyses using parsimony and model-based approaches, perhaps in a bid to integrate over the uncertainty over which method provides the most credible estimate of inter-specific relationships. Certainly, without knowledge of the true phylogeny it is not possible to reconcile the conflicting results from competing methods. Hence, Wright & Hillis (2014) took a simulation approach, generating thousands of morphology-like datasets on a known tree and then assessing the relative performance of parsimony and the Bayesian implementation of the Mk model in recovering the generating tree from the simulated data. They found that the model-based method performed best. Schooled in parsimony, we were surprised by the findings of Wright & Hillis (2014) and believed that there were aspects of their experimental design that potentially biased their analyses in favour of the probabilistic model; not least that their data were effectively generated using the Mk model. Also, we wanted to assess the performance of alternative parsimony methods, and benchmark their performance with simulated data against empirical matrices. However, even when accounting for these factors, we recovered the same result as Wright & Hillis (2014): the Bayesian implementation of the Mk model outperformed parsimony (O'Reilly *et al.* 2016). Both ourselves and others have since attempted to explore other variables influencing the estimation of phylogenetic relationships, such as tree symmetry and character design (Puttick *et al.* 2017*a*), as well as measures of clade support (Brown *et al.* 2017; O'Reilly *et al.* 2017). There are many other variables that have yet to be investigated, including character covariation, the accuracy of branch length estimates, and the impact of non-contemporaneous taxa. However, based on existing simulation approaches and the variables considered to date, the Bayesian implementation of the Mk model continues to perform with greatest accuracy, particularly when datasets are small and levels of homoplasy are high (O'Reilly *et al.* 2017).

Is parsimony dead? Goloboff *et al.* (2018) certainly do not think so, calling into question all of our results based principally on the argument that the model of evolution that we used to simulate morphology-like data, is not biologically realistic. We cannot address every point they make, not least since their critique is focused explicitly on what we did not write, rather than what we did write. However, Goloboff *et al.* (2017, 2018) object particularly to the assumption in our simulating framework of the proportionality of branch lengths among characters, which is clearly an unrealistic expectation of morphological evolution. In this we are agreed; if there were an entirely realistic model of morphological evolution available we would have used it. However, if such a model were available, we could dispense with both parsimony

doi: 10.1111/pala.12361

[The copyright line for this article was changed on June 13, 2018, after original online publication]

and the Mk model and simply apply this model to derive the true relationships among taxa.

Empirical realism of simulated data

The critique of Goloboff *et al.* (2018), focused on the biological realism of the model of evolution that we used to simulate morphology-like data, is based on a revisionist perspective. All of our original model choices were informed by the pioneering study of Wright & Hillis (2014); we employed an identical methodology to these authors, where possible, to allow for a direct comparison: we used the same generating tree (Pyron 2011); we used the same number of characters for simulation (350 and 1000 sites; we added 100 character datasets as they are representative of the size of many palaeontological studies); and we used a similar character simulation model but with modifications to violate the Mk model. We used the HKY+G model of molecular evolution on known trees to create datasets that violate assumptions underlying the Mk model. The HKY model generates data with an uneven stationary distribution of state frequencies in our simulations, violating one of the primary assumptions of the Mk model. These nucleotide datasets were converted to binary or multistate morphology-like datasets by reducing the four nucleotide states to purines and pyrimidines (R/Y coding) and recoding them as binary states, or by directly mapping the four nucleotides to integers for multistate characters. We also achieved further model misspecification by drawing a unique rate for each character from a continuous gamma distribution; the Mk model assumes all characters have an equal expected number of changes on individual branches, and the Mk+G model assumes there are n unique rates, where n is the number of discrete gamma categories. To ensure that these simulated datasets were also empirically realistic, we evaluated their overall consistency index (CI), excluding datasets that fell outside the range of CI in a published survey of empirical datasets (Sanderson & Donoghue 1989, 1996). In O'Reilly *et al.* (2016), we explored the impact of CI filtering on our results, and in subsequent papers the use of CI filtering became part of the simulation procedure (O'Reilly *et al.* 2017; Puttick *et al.* 2017*a*).

As we stated explicitly in our study, we attempted to obtain two qualities in our simulated data: (1) that the generating model violated the Mk model; and (2) that it achieved our prescribed measure of empirical realism. Our analyses using the Mk model frequently failed to recover the generating tree with precision or accuracy, demonstrating effectively that the simulated datasets are not compatible with this evolutionary model and achieve a suitable level of model misspecification. Goloboff *et al.* (2017) have already corroborated the empirical realism of the simulated datasets. Thus, Goloboff *et al.* (2017, 2018) effectively conflate the need for empirical realism in the model used to generate the data with the efficacy of the methods in analysing the data.

Alternative simulation approaches

To simulate data, Goloboff *et al.* (2018) prefer their own model, in which the rate of change for each character is completely independent on every branch of the tree. Their implicit (Goloboff *et al.* 2017) and then later explicit (Goloboff *et al.* 2018) claim that their model is more biologically realistic is no better justified than the Mk model, as neither can be supported with meaningful quantitative empirical evidence. If Goloboff *et al.* (2018) consider a model in which characters share a set of branch lengths to be biologically unrealistic, they must also accept that the assumptions of their own model are at least equally biologically unrealistic, if not potentially more so. Goloboff *et al.* (2018) argue that it is not possible to generalize based on the simulation procedure from O'Reilly *et al.* (2017) and Puttick *et al.* (2017*a*); if true, this same argument can be levelled at their own simulation procedure.

The Goloboff *et al.* (2017) simulation model effectively represents an almost polar opposite to the HKY+G simulation procedure of O'Reilly *et al.* (2017) as it allows for unique rates for each character on each branch, whereas our simulation approach reduces the number of parameters by allowing the expected number of changes on a branch to be shared among all characters, with some proportional augmentation by factors randomly sampled from a gamma distribution. Parsimony and maximum likelihood will achieve identical results if all branches are allowed a unique rate for each character (Tuffley & Steel 1997). However, this no common mechanism model is unwieldy as it employs a huge number of parameters that grows exponentially with dataset size: $(2 \times \text{number of taxa} - 3) \times \text{number of characters}$ (Huelsenbeck *et al.* 2011; Yang 2014). The simulation procedure of Goloboff *et al.* (2017) is comparable to this extremely parameter-rich model that sits at the extreme of branch-rate independence.

In reality, a more suitable model of morphological evolution probably exists somewhere on the continuum of potential models separating our simulation framework and that of Goloboff *et al.* (2017). The idiosyncrasies of morphological evolution mean that it is daunting to construct a single model of discrete character change applicable to all datasets. We possess little, if any, meaningful data regarding the manner in which rates of morphological evolution vary across characters and along the

branches of trees. Thus, if we are to assess the performance of the relatively naïve inference frameworks we have available to us, it seems logical to focus instead on the empirical realism of the structure of simulated data itself and not the biological realism of the process that generated it. Similarly, identifying a useful model that separates the simulation procedures of O'Reilly *et al.* (2017) and Goloboff *et al.* (2017) is neither straightforward nor necessary to assess the efficacy of the available phylogenetic estimation frameworks.

Simulated and empirical data

Goloboff *et al.* (2017, 2018) conclude both of their papers by observing 'the use of simulated datasets alone cannot solve that [*sic*] problem of model adequacy; empirical tests of whether morphological data fulfill the crucial assumptions of the model are required as well.' Nevertheless, they emphasize, the benefit of simulation is that it is possible to derive general patterns from statistically significant numbers of replicates. We prefer our own approach to the simulation of a set of morphological matrices through the filter of character consistency since, in our view, the approach taken by Goloboff *et al.* (2017) yielded datasets with empirically unrealistic distributions of character consistency which were frequently dominated by characters with a high CI; datasets that parsimony analysis will naturally perform well on. Implied Weights Parsimony relies upon a measure of character consistency, and is only likely to reinforce the true tree when homoplasy is low (Kluge 1997; Congreve & Lamsdell 2016).

Goloboff *et al.* (2018) question how we evaluated their simulated datasets since they did not provide any with their paper (Goloboff *et al.* 2017); we used the code provided in the supplementary materials of Goloboff *et al.* (2017) to create simulated datasets and, if the simulation strategy of Goloboff *et al.* (2017) is effective, our sample of simulated data should be statistically comparable to the data they generated and based their study on. We present the CI profile of characters within datasets simulated using their strategy in Figure 1, comparing the CI profile of empirical datasets surveyed by Goloboff *et al.* (2017; Fig. 1A), to that of 2000 datasets simulated by their protocol (Fig. 1B) versus that of O'Reilly *et al.* (2017) for 1000 replicates of 100 characters simulated on an asymmetric tree. Datasets simulated following the approach of Goloboff *et al.* (2017) always include a significant number of characters with a CI = 1.0 even though they are all comprised of multistate characters. Similarly, the simulated matrices of Goloboff *et al.* (2017) often under-represent characters with CI < 0.5 relative to the empirical matrices they surveyed. This under representation of low CI characters is particularly obvious in CI bins spanning the range 0.0–0.2, containing the most inconsistent characters. This distribution of per character CI effectively reduces the exposure of the different phylogenetic estimation methods to increasingly inconsistent characters. This bears out the point made in O'Reilly *et al.* (2017) and it is in this sense that we viewed the simulation strategy of Goloboff *et al.* (2017) to be biased in favour of parsimony.

Goloboff *et al.* (2017, 2018) ignore the empirical analyses we conducted (O'Reilly *et al.* 2016, 2017; Puttick *et al.* 2017*a*) even though model comparison using empirical data is the approach advocated by Goloboff *et al.* (2017, 2018). These analyses show that the predictions based on our simulation data are extendable to empirical datasets. Specifically, smaller datasets achieve lower precision with the Bayesian implementation of the Mk model, and larger datasets show increasing congruence in the recovered topological across all inference methods. We would not expect these predictions to be true if our simulation-based analyses were inherently invalid.

Model efficacy vs adequacy

Goloboff *et al.* (2017, 2018) conflate the issue of method efficacy and model adequacy. Our explicit aim was to evaluate the efficacy of parsimony, and both maximum likelihood and Bayesian approaches to the estimation of phylogeny. At no stage did we attempt to evaluate the adequacy of the Mk model, or its ability to effectively capture the process of morphological evolution. Similarly, at no stage did we argue that either the single parameter Markov model or the manner in which the likelihood of a topology is calculated across a dataset adequately capture the process of morphological change. Indeed, it is widely observed among proponents of statistical phylogenetic inference that the Mk model will require further development if it is to encapsulate the process of morphological change to the maximum afforded by the Markov model framework (e.g. Wright *et al.* 2016), and the potential for improvement in the Mk model can be viewed as a strength, rather than a weakness.

The future

We and others have made steps towards a simulation-based assessment of phylogenetic methods (Wright & Hillis 2014; O'Reilly *et al.* 2016, 2017; Brown *et al.* 2017; Goloboff *et al.* 2017; Puttick *et al.* 2017*a*, *b*) so far considering the impact of tree symmetry (Puttick *et al.* 2017*a*) and clade support (Brown *et al.* 2017; O'Reilly *et al.* 2017). As Goloboff *et al.* (2017, 2018) observe, there are other parameters to consider, such as non-

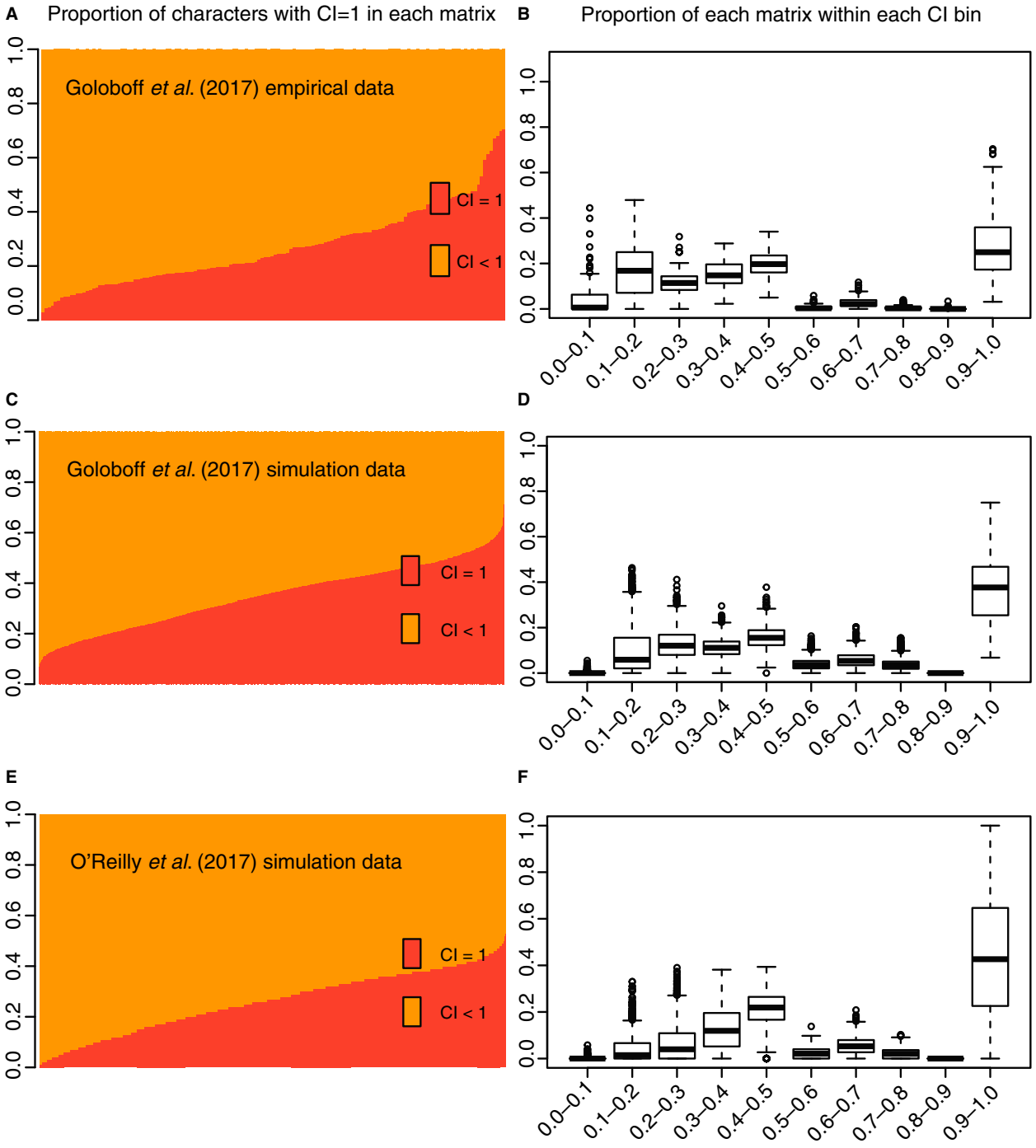

FIG. 1. Comparison of empirical and simulation datasets in terms of the consistency of the component characters. A–B, empirical datasets compiled by Goloboff *et al.* (2017). C–D, datasets simulated using the strategy of Goloboff *et al.* (2017). E–F, datasets simulated using the strategy of O'Reilly *et al.* (2017). A, C, E, the proportion of characters within each dataset that have a consistency index of 1.0. B, D, F, the proportion of characters within each dataset within each of ten consistency index bins. Colour online.

contemporaneous terminals, the accuracy of branch length estimates, character coevolution and covariation. We look forward to their exploration in turn.

In the interim, model-based phylogenetic methods appear to perform best when parsimony methods perform most poorly (when datasets are small and exhibit low character consistency) and perform at least as well as parsimony methods when they perform best (when datasets are large and exhibit high character consistency).

Acknowledgements. We would like to thank remaining members of the Bristol Palaeobiology Research Group for discussion. MNP is funded by a 1851 Research Fellowship from the Royal Commission for the Exhibition of 1851; JEO'R., DP and PCJD are funded by NERC (NE/P013678/1); PCJD is also funded by NERC (NE/N002067/1) and BBSRC (BB/N000919/1).

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.

CONGREVE, C. R. and LAMSDELL, J. C. 2016. Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Palaeontology*, **59**, 447–462.

FISHER, D. C., FOOTE, M., FOX, D. L. and LEIGHTON, L. R. 2002. Stratigraphy in phylogeny reconstruction – comment on Smith (2000). *Journal of Paleontology*, **76**, 585–586.

FOX, D. L., FISHER, D. L. and LEIGHTON, L. R. 1999. Reconstructing phylogeny with and without temporal data. *Science*, **284**, 1816–1819.

GOLOBOFF, P. A., 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.

—— —— —— 2018. Parsimony and model-based phylogenetic methods for morphological data: a comment on O'Reilly *et al.* *Palaeontology*, published online April. https://doi.org/10.1111/pala.12353

HUELSENBECK, J. P., ALFARO, M. E. and SUCHARD, M. A. 2011. Biologically inspired phylogenetic models strongly outperform the no common mechanism model. *Systematic Biology*, **60**, 225–232.

HULL, D. 1988. *Science as a process: an evolutionary account of the social and conceptual development of science*. University of Chicago Press, 538 pp.

KLUGE, A. G. 1997. Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics*, **13**, 81–96.

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. J. 2017*a*. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B*, **284**, 20162290.

—— —— OAKLEY, D., 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. J. 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.

PYRON, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Systematic Biology*, **60**, 466–481.

SANDERSON, M. J. and DONOGHUE, M. J. 1989. Patterns of variation in levels of homoplasy. *Evolution*, **43**, 1781–1795.

—— —— 1996. The relationship between homoplasy and confidence in a phylogenetic tree. 67–89. *In* SANDERSON, M. J. and HUFFORD, L. (eds). *Homoplasy: the recurrence of similarity in evolution*. Academic Press.

SMITH, A. B. 2000. Stratigraphy in phylogeny reconstruction. *Journal of Paleontology*, **74**, 763–766.

TUFFLEY, C. and STEEL, M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, **59**, 581–607.

WAGNER, P. J. 2002. Testing phylogenetic hypotheses with stratigraphy and morphology – a comment on Smith (2000). *Journal of Paleontology*, **76**, 590–593.

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

—— LLOYD, G. T. and HILLIS, D. M. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, **65**, 602–611.

YANG, Z. 2014. *Molecular evolution: a statistical approach*. Oxford University Press, 492 pp.