BIOLOGY LETTERS

royalsocietypublishing.org/journal/rsbl

Research

Cite this article: Smith MR. 2019 Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. *Biol. Lett.* **15**: 20180632.
http://dx.doi.org/10.1098/rsbl.2018.0632

Received: 6 September 2018
Accepted: 11 January 2019

Subject Areas:
evolution, taxonomy and systematics, palaeontology

Keywords:
phylogenetic inference, parsimony analysis, equal weights, implied weighting, Bayesian phylogenetic methods, information content

Author for correspondence:
Martin R. Smith
e-mail: martin.smith@durham.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4381736.

THE ROYAL SOCIETY PUBLISHING

Evolutionary biology

Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets

Martin R. Smith

Department of Earth Sciences, Lower Mount Joy, Durham University, Durham DH1 3LE, UK

MRS, 0000-0001-5660-1727

Phylogenetic analysis aims to establish the true relationships between taxa. Different analytical methods, however, can reach different conclusions. In order to establish which approach best reconstructs true relationships, previous studies have simulated datasets from known tree topologies, and identified the method that reconstructs the generative tree most accurately. On this basis, researchers have argued that morphological datasets should be analysed by Bayesian approaches, which employ an explicit probabilistic model of evolution, rather than parsimony methods—with implied weights parsimony sometimes identified as particularly inaccurate. Accuracy alone, however, is an inadequate measure of a tree's utility: a fully unresolved tree is perfectly accurate, yet contains no phylogenetic information. The highly resolved trees recovered by implied weights parsimony in fact contain as much useful information as the more accurate, but less resolved, trees recovered by Bayesian methods. By collapsing poorly supported groups, this superior resolution can be traded for accuracy, resulting in trees as accurate as those obtained by a Bayesian approach. By contrast, equally weighted parsimony analysis produces trees that are less resolved and less accurate, leading to less reliable evolutionary conclusions.

1. Introduction

Evolutionary history can be reconstructed using parsimony-based or probabilistic approaches. Because models used with molecular datasets generally share a common probabilistic construction, statistical methods can be used to determine the most appropriate model [1]. With morphological datasets, however, it is more difficult to establish whether probabilistic models or parsimony better reconstruct phylogenetic relationships (which are typically unknown).

A pragmatic approach to this question is to simulate data from a known tree. With the important caveat that generative trees and simulated morphological datasets may be unrealistic [2,3], probabilistic approaches typically reconstruct the generative tree most accurately (i.e. with least conflict), followed by parsimony under equal and implied weights in turn [4–9].

Previous studies have advocated accuracy as the sole criterion by which to select a method [5–11]. Congreve & Lamsdell [9] (problematically [2]) define the most accurate tree as the one that bears the fewest incorrect splits. Other authors [5–8,11] use the Robinson-Foulds (RF) distance as a proxy for accuracy (even though the RF distance is also influenced by precision; a pair of trees can be made two units more similar by replacing an incorrect partition with a correct one, or by collapsing two incorrect partitions). Goloboff *et al*. [2] propose alternative tree similarity metrics as proxies for accuracy.

Accuracy alone, however, is not the only goal when reconstructing trees [11]. No tree shows less conflict than a single polytomy, for a total absence of relationship information guarantees that no relationship is incorrectly resolved.

An emphasis on accuracy therefore disadvantages methods that produce highly resolved trees [11] (and *vice versa*). This trade-off has been acknowledged by collapsing some poorly supported groups before calculating accuracy (even if accuracy is still equated with 'performance') [2,6,8,11]. Naturally [12], methods that yield less resolution are consistently more accurate [2,5,7,8,11].

We should be seeking not the most accurate method, but the method that recovers as much *information* as possible about the true tree, striking a balance between the complementary quantities [12] of accuracy and resolution. For example, a tree that resolves 20 relationships conveys much information about the correct tree, even if one of those relationships is incorrect; a tree that resolves just one relationship conveys less information, even if that single relationship is correct. If two trees are equally accurate, we should prefer the more precise. Here I explore the impact on previous studies of evaluating trees according to their total shared information content, rather than 'accuracy' alone.

2. Material and methods

Congreve & Lamsdell ([9]; CL hereafter) simulated 55-character matrices from a bifurcating 22-tip tree using a Markov *k*-state (Mk) 1 parameter model with rates sampled from a discretized gamma distribution [13]. Their generative tree is the single most parsimonious tree obtained from a study of Ordovician trilobites; its edges were assigned a unit length.

O'Reilly *et al.* ([5]; OR hereafter) simulated matrices containing 100, 350 and 1000 characters from a bifurcating 75-tip tree using a modified HKY85 model [14]; they followed a previous simulation study [4] in selecting a single bifurcating tree from a morphological + molecular analysis of Lissamphibia.

I used TNT [15] to conduct parsimony searches on each of these matrices under equal and implied weights, using the parsimony ratchet and sectorial search heuristics (search options: xmult:hits 20 level 4 chklevel 5 rat10 drift10). I took a strict consensus of all optimal trees obtained under equal weights, and under implied weights [16] at the concavity constants used in each respective study (CL: $k = 1$, 2, 3, 5 and 10; OR: $k = 2$, 3, 5, 10, 20 and 200). For each dataset, I generated a further strict consensus of all trees that were optimal under any of the concavity constants, excluding the unreasonable value of $k = 1$, which inadequately penalizes extra steps beyond the first, and thus exhibits undesirable properties of clique analysis [17] (see electronic supplementary material).

I also generated majority-rule consensus trees in MrBayes 3.2.2 [18] using an Mk model, with rates distributed according to a gamma parameter. I combined results from four independent runs, each of which employed four Metropolis-coupled Markov chains. After a burn-in period of 4 000 000 generations, the cold chain in each run was sampled every 10 000 generations for 6 000 000 generations. The sampled topologies faithfully reflected the posterior distribution for each dataset ($0.999 <$ potential scale reduction factor < 1.001; estimated sample size > 400).

To explore the relationship between resolution and accuracy, I generated further trees for each analysis by collapsing poorly supported groups. Under the Mk model, I collapsed groups whose posterior probability was less than 95%, 90%, 85%, … 50%. In parsimony analyses, I compared different measures of node support. Under jackknife and bootstrap resampling, I collapsed groups with (i) absolute frequency supports of less than 0%, 2%, 4% … 100%; (ii) relative frequency (GC) support of less than −100%, −95%, … 95%, 100%. Under Bremer support, I collapsed groups with Bremer support values less than 1, 2, 3, … 20 with equally weighted trees (TNT command subopt `x`; `bbreak;`); under implied weighting, Bremer support values were drawn from a logarithmic distribution ($0.73^{0\ldots19}$, $2.5 \times 10^{-3} \rightarrow 1 \times 10^{0}$), reflecting the fractional nature of tree scores under implied weights [16].

Symmetric difference metrics calculate how much information two trees hold in common [19]—that is, how much information a generated tree contains about the generative tree. Where the generative tree is bifurcating, a particular relationship may be resolved the same way (s) or a different way (d) on each tree, or resolved in the comparison tree only (r) [20,21]. The symmetric difference ('SD', also termed the Robinson-Foulds distance) is given by $2d + r$. The symmetric difference is conventionally normalized against the total information present (TIP) in the two trees, $2d + 2s + r$ [21]. Undesirably, this assigns a fully unresolved tree the same score as a tree that is perfectly resolved and completely incorrect (figure 1*a*). In the present context, therefore, it is more appropriate to normalize against the maximum information (MaxI) that could potentially have been resolved, $2(d + s + r)$.

The unit of relationship information may be a quartet (a four-taxon statement) [20–22] or a bipartition split [23–25]. (Each clade in a tree corresponds to a bipartition that splits taxa into 'members' and 'non-members'.) Partitions offer a simple but incomplete measure of the topological information accommodated in a tree. The trees ((A, (**X**, B)), (C, D)) and ((A, B), ((C, **X**), D)) both contain the same information regarding the relationships between (A, B) and (C, D), yet have no partitions in common. As a consequence, the partition difference (= Robinson-Foulds distance) suffers four essential shortcomings [23]. First, it is imprecise; the number of unique values that the metric can take is two fewer than the number of taxa. (Simply put, a precise method can allocate distinct difference values to two trees that an imprecise method would assign an identical score.) Second, it is rapidly saturated; relatively minor differences can result in the maximum distance value. Third, its value can be counterintuitive; for example, moving a single tip to a particular location can generate a higher difference value than moving both that tip and its immediate neighbour to the same point (electronic supplementary material). Fourth, balanced trees contain proportionally more uneven partitions, and thus attract lower average distances than asymmetric trees (electronic supplementary material).

Quartets, by contrast, represent all topological information within a tree. The quartet dissimilarity measure is precise, does not rapidly reach saturation, generates a meaningful value for random trees, is robust to the placement of wildcard taxa and consistently increases in value as trees become more different; and every quartet represents an equal quantity of information. I consider it to represent a more useful, meaningful and interpretable indicator of tree similarity.

I calculated quartet distances using the tqDist algorithm [26] via the `QuartetStatus` function in the new R package `Quartet` [27]. Partition distances were calculated using the `Quartet` function `SplitStatus`. To summarize results, s, d and r were calculated for each individual tree relative to the generative tree, and the mean of each parameter was calculated at each resolution in each analysis.

Previous studies (e.g. [5,6]) have plotted unnormalized symmetric difference against the resolution. The unnormalized symmetric difference, however, is a function of both resolution and accuracy: a change in resolution (x) necessarily influences the value, and the range of possible values, of the symmetric difference (y). Because the axes are not independent, this is analogous to plotting x against y/x; the inherent correlation between the axes makes it difficult to interpret the relative contributions of x and y to the plotted function. I instead plotted the proportion of quartets or partitions that are the same in both trees (s), different

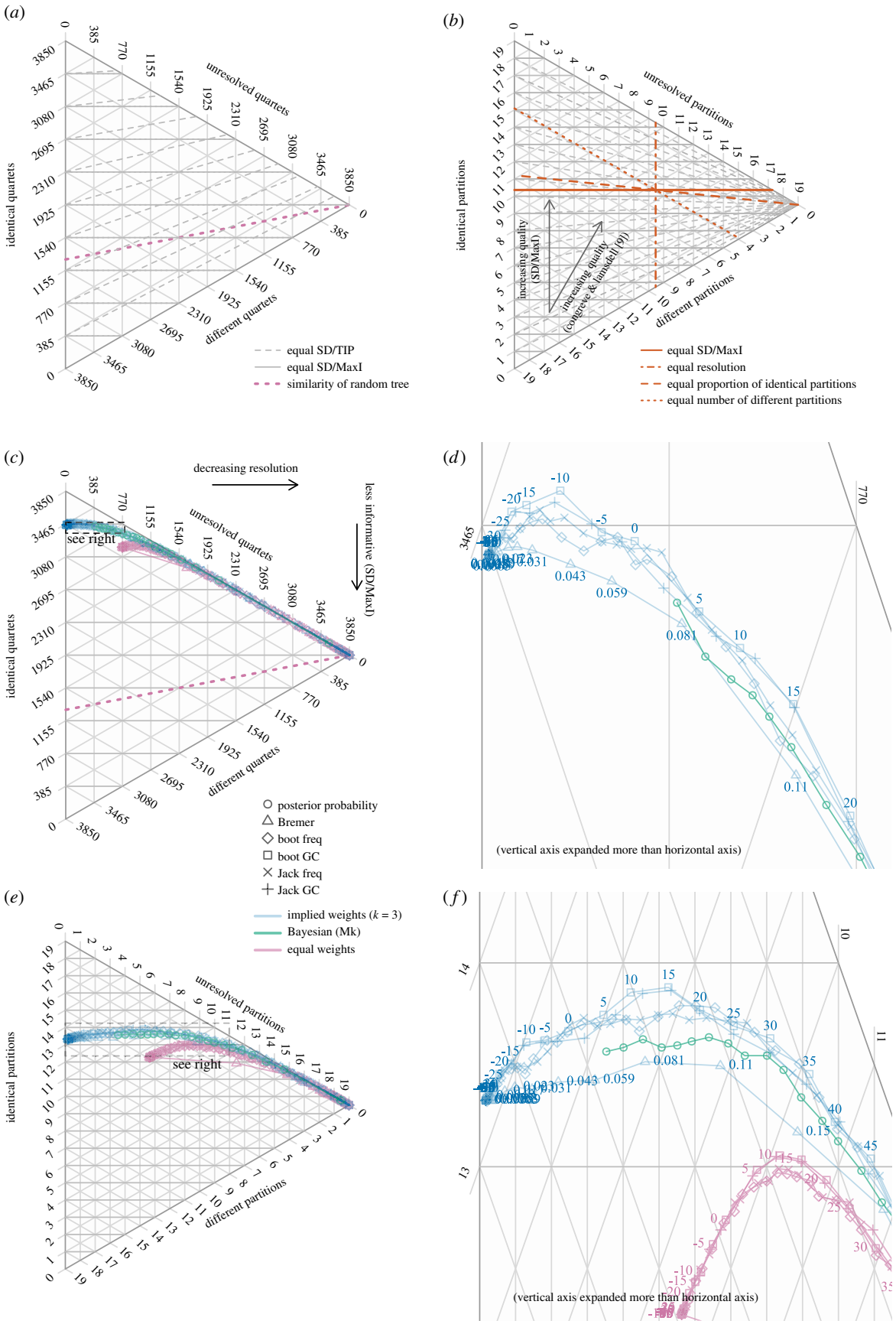

Figure 1. Method selection. (*a*) Normalizing symmetric difference against the total information present in two trees (SD/TIP, dashed lines) scores a completely incorrect bifurcating tree (all relationships resolved differently; bottom corner) no worse than a polytomy (all relationships unresolved; rightmost corner). Random trees (coloured line) with more relationships resolved receive better scores, as some relationships will by chance be resolved correctly. Normalizing against the maximum possible relationship information (SD/MaxI, solid lines) penalizes misinformation over non-information; random trees with more relationships resolved (which thus contain more misinformation) consequently receive worse scores. (*b*) Four measures of tree quality. (*c*–*f*) Impact on tree quality when least-supported groups are collapsed: (*c*–*d*) Counting quartets; (*e*–*f*) counting partitions.

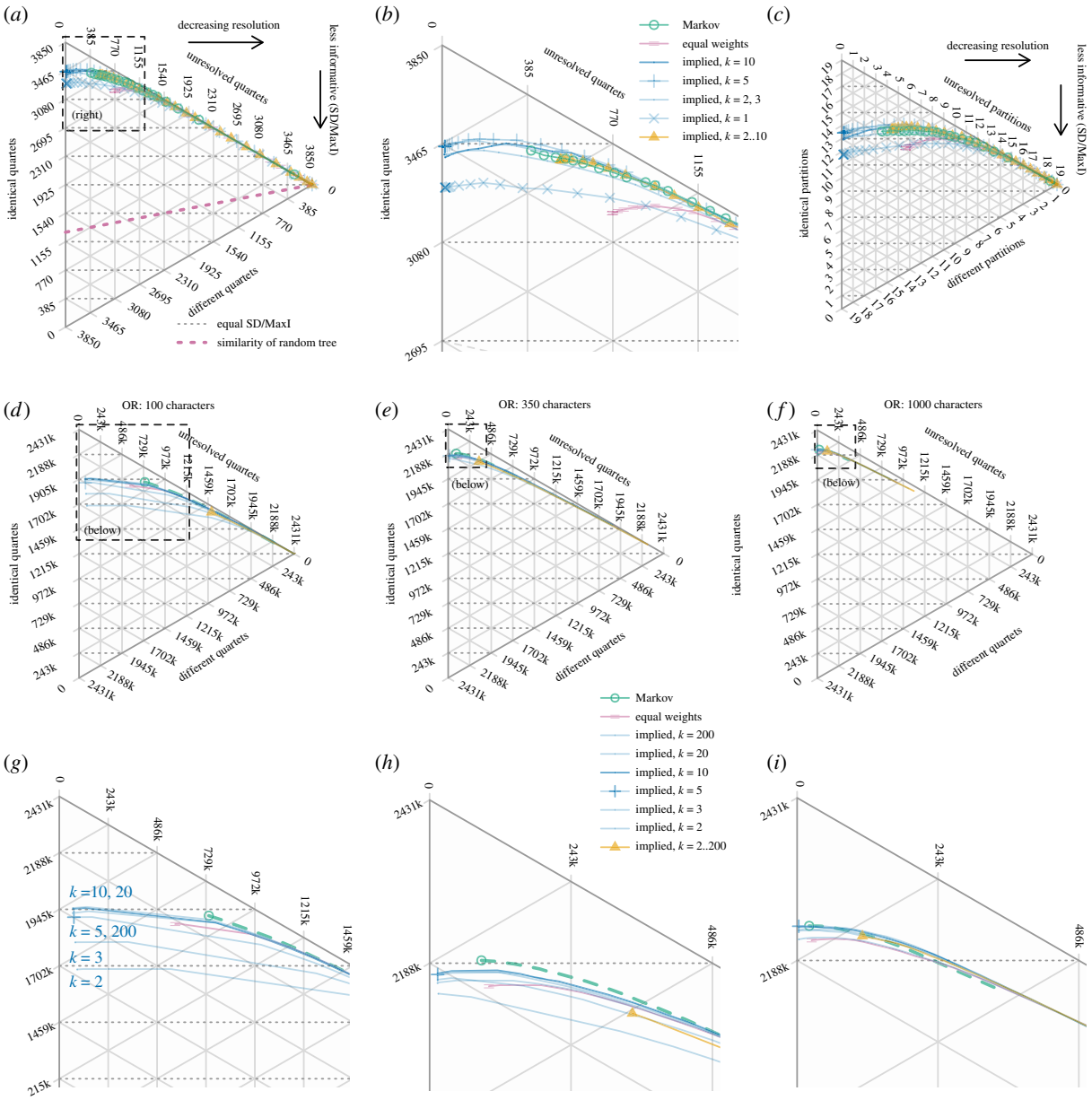

Figure 2. Status of quartets and bipartitions in trees recovered from simulated datasets. Points denote the average number of quartets (*a*–*b*,*d*–*i*) or partitions (*c*) that are the same as the generative tree, resolved differently to the generative tree, or not resolved. Each series indicates the effect of progressively collapsing the least-supported groups in trees generated by analysis of CL (*a*–*c*) and OR datasets (*d*,*g*: 100; *e h*: 350; *f*,*i*: 1000 characters) under the specified analytical parameters. The vertical direction corresponds to similarity (i.e. more informative trees); the horizontal direction corresponds to resolution.

in both trees (*d*) and only resolved in the generative tree (*r*) on ternary plots using the `Ternary` R package [28], oriented such that SD/MaxI decreases vertically and resolution decreases horizontally (figure 1*a*). This plotting configuration distinguishes the relative contributions of resolution and accuracy to overall similarity (figure 1*b*).

Data, scripts and analyses used in this study are archived on GitHub [29,30].

3. Results

Ideally, measures of node support would assign incorrect nodes low support values. With the CL datasets (55 characters, 22 tips), resampling methods accomplished this more effectively than Bremer support (figure 1, *c*–*f*), a metric that has attracted criticism [31,32]. The groups contradicted/supported (GC) metric outperformed group frequency (as anticipated by [33]), whereas bootstrap resampling outperformed the jackknife approach (contra [34]); subsequent analyses thus employed the bootstrap GC metric. Differences between methods were not statistically significant (electronic supplementary material).

With the CL datasets, there is no significant difference (at $p = 0.01$) between the MaxI-normalized quartet symmetric difference of the best trees generated by the Mk model or implied weights ($k \in \{2, 3, 5, 10\}$)—but the best trees generated by equal weights, implied weights with $k = 1$, and the consensus of k values are significantly worse than those produced by the other methods (figure 2*a*,*b*; electronic supplementary material).

Collapsing the least-supported groups initially increases the overall accuracy (as predicted in [2,35]), leading to a slight increase in the overall informativeness of the tree (figure 2*a*,*b*). Beyond a GC score of *ca* −15, the gain in accuracy no longer offsets the resolution lost; collapsing further groups thus removes 'correct' information and reduces the similarity between the tree and the reference tree. Indeed,

the optimal tree is only perfectly resolved in a minority of cases (CL, 18%; OR: less than 0.2%). Because a Bayesian approach results in less resolution, its most resolved trees cannot generally be improved by collapsing groups (figures 1*c*,*d* and 2).

These results hold even if the (problematic) partition difference metric is employed (figure 1*e*–*f*), though relatively more groups must be collapsed (those with a GC score of less than 10) to maximize this metric. The results do not meaningfully change when datasets with low consistency indices are excluded.

Similar results are observed in the OR datasets (figure 2*d*–*i*): at any given level of resolution, the best trees obtained by the Mk model are similar in accuracy to those obtained under implied weights (except with very small values of *k*) but are more accurate than those obtained using equal weights.

These datasets also demonstrate the impact of dataset size on tree quality. With larger ratios of characters to taxa (1000 or 350 characters, 75 tips), all methods produced reasonably accurate, well-resolved trees (figure 2*e*–*f*, *h*–*i*). With the smallest (100 character) datasets (figure 2*d*,*g*), trees were much more different from the generative tree, and the choice of method influenced results more strongly: the Bayesian approach could obtain substantially less resolution, and implied weights recovered poor trees at low values of *k*. No existing method can overcome the inherent limitation of a low character to taxon ratio.

4. Discussion

When accuracy and resolution are recognized as complementary aspects of information [12], parsimony and probabilistic analyses generate equally informative reconstructions of evolutionary history in the simulation studies analysed herein. Parsimony results are most informative when groups with a bootstrap GC value of less than −15 are collapsed, and are as accurate as Bayesian results if nodes are collapsed until trees exhibit an equal resolution. As an important caveat, parsimony analysis must employ a moderate weighting scheme. At low values of the concavity constant ($k < 2$, say), implied weights begin to exhibit the undesirable properties of clique analysis, whereas at high values (as $k \to \infty$), it converges to the inferior equally weighted parsimony (electronic supplementary material). Each of these extremes yields results that are less accurate and less resolved, making them more different from the generative tree and consequently less informative about evolutionary history; results encountered only under such parameters do not merit biological interpretation.

Quite aside from issues with the validity of data simulation protocol [2,3], previous results that favour Bayesian methods over parsimony [5–8,10], or equal weights over implied weights [9], have arisen because accuracy has been considered the sole measure of a method's performance. Future simulation studies should evaluate methods based on normalized tree similarity metrics that reflect the total *information* contained within two trees—a quantity that reflects both resolution and accuracy. In the analyses examined herein, neither Bayesian nor parsimony analyses generate consistently superior results. Of course, other factors may influence a researcher's choice of methods: Bayesian models, for instance, can readily integrate non-morphological data [36,37] and allow probabilistic hypothesis testing using Bayes Factors [38]. Such considerations notwithstanding, researchers may wish to explicitly compare the results of both Bayesian and implied weights analyses when conducting phylogenetic analysis; observations common to both approaches and receiving strong node support values are particularly likely to be well supported by underlying data.

Data accessibility. Data and analyses are available on GitHub and archived with Zenodo. CL: https://github.com/ms609/CongreveLamsdell2016 [29]; OR: https://github.com/ms609/OReillyEtAl2016 [30]. The datasets are available in the electronic supplementary material.

Competing interests. I declare that I have no competing interests.

Funding. I received no funding for this study.

Acknowledgements. The TNT software is supported by the Willi Hennig Society. Detailed comments from two anonymous referees substantially improved the manuscript.

References

1. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589. (doi:10.1038/nmeth.4285)
2. Goloboff PA, Torres A, Arias JS. 2018 Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics* **34**, 407–437. (doi:10.1111/cla.12205)
3. Goloboff PA, Torres Galvis A, Arias JS. 2018 Parsimony and model-based phylogenetic methods for morphological data: comments on O'Reilly *et al*. *Palaeontology* **61**, 625–630. (doi:10.1111/pala.12353)
4. Wright AM, Hillis DM. 2014 Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* **9**, e109210. (doi:10.1371/journal.pone.0109210)
5. O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016 Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett.* **12**, 20160081. (doi:10.1098/rsbl.2016.0081)
6. Puttick MN, O'Reilly JE, Pisani D, Donoghue PCJ. 2019 Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. *Palaeontology* **62**, 1–17. (doi:10.1111/pala.12388)
7. Puttick MN *et al*. 2017 Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proc. R. Soc. B* **284**, 20162290. (doi:10.1098/rspb.2016.2290)
8. O'Reilly JE, Puttick MN, Pisani D, Donoghue PCJ. 2018 Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology* **61**, 105–118. (doi:10.1111/pala.12330)
9. Congreve CR, Lamsdell JC. 2016 Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Palaeontology* **59**, 447–462. (doi:10.1111/pala.12236)
10. Puttick MN *et al*. 2017 Parsimony and maximum-likelihood phylogenetic analyses of morphology do not generally integrate uncertainty in inferring evolutionary history: a response to Brown *et al*. *Proc. R. Soc. B* **284**, 20171636. (doi:10.1098/rspb.2017.1636)
11. Brown JW, Parins-Fukuchi C, Stull GW, Vargas OM, Smith SA. 2017 Bayesian and likelihood

phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick *et al*. *Proc. R. Soc. B* **284**, 20170986. (doi:10.1098/rspb.2017.0986)

12. Mackay DM. 1953 Quantal aspects of scientific information. *IEEE Trans. Inf. Theory* **1**, 60–80. (doi:10.1109/TIT.1953.1188569)
13. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925. (doi:10.1080/106351501753462876)
14. Hasegawa M, Kishino H & Yano T-A. 1985 Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174. (doi:10.1007/BF02101694)
15. Goloboff PA, Farris JS, Nixon KC. 2008 TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786. (doi:10.1111/j.1096-0031.2008.00217.x)
16. Goloboff PA. 1993 Estimating character weights during tree search. *Cladistics* **9**, 83–91. (doi:10.1111/j.1096-0031.1993.tb00209.x)
17. Wilkinson M. 1994 Three-taxon statements: when is a parsimony analysis also a clique analysis? *Cladistics* **10**, 221–223. (doi:10.1111/j.1096-0031.1994.tb00174.x)
18. Huelsenbeck JP, Ronquist F. 2001 MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755. (doi:10.1093/bioinformatics/17.8.754)
19. Gaudesi M, Squillero G, Tonda A. 2014 Universal information distance for genetic programming. *Proc. 2014 Conf. Genet. Evol. Comput*, pp. 137–138. (doi:10.1145/2598394.2598440)
20. Estabrook GF, McMorris FR, Meacham CA. 1985 Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst. Zool.* **34**, 193–200. (doi:10.2307/sysbio/34.2.193)
21. Day WHE. 1986 Analysis of quartet dissimilarity measures between undirected phylogenetic trees. *Syst. Biol.* **35**, 325–333. (doi:10.1093/sysbio/35.3.325)
22. Bandelt HJ, Dress A. 1986 Reconstructing the shape of a tree from observed dissimilarity data. *Adv. Appl. Math.* **7**, 309–343. (doi:10.1016/0196-8858(86)90038-2)
23. Steel MA, Penny D. 1993 Distributions of tree comparison metrics—some new results. *Syst. Biol.* **42**, 126–141. (doi:10.1093/sysbio/42.2.126)
24. Penny D, Hendy M. 1985 The use of tree comparison metrics. *Syst. Zool.* **34**, 75–82. (doi:10.2307/2413347)
25. Robinson DF, Foulds LR. 1981 Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
26. Sand A, Holt MK, Johansen J, Brodal GS, Mailund T, Pedersen CN. S. 2014 tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics* **30**, 2079–2080. (doi:10.1093/bioinformatics/btu157)
27. Smith MR. 2019 Quartet: comparison of phylogenetic trees using quartet and bipartition measures. *Zenodo* (doi:10.5281/zenodo.2536318)
28. Smith MR. 2017 Ternary: an R package to generate ternary plots. *Zenodo* (doi:10.5281/zenodo.1068997).
29. Smith MR. 2019 Distance metrics for trees generated by Congreve and Lamsdell (2016). ms609.github.io/CongreveLamsdell2016. (doi:10.5281/zenodo.2536874)
30. Smith MR. 2019 Distance metrics for trees generated by O'Reilly *et al*. (2016). ms609.github.io/OReillyEtAl2016. (doi:10.5281/zenodo.2536935)
31. Wilkinson M, Thorley JL, Upchurch P. 2000 A chain is no stronger than its weakest link: double decay analysis of phylogenetic hypotheses. *Syst. Biol.* **49**, 754–776. (doi:10.1080/106351500750049815)
32. DeBry RW. 2001 Improving interpretation of the decay index for DNA sequence data. *Syst. Biol.* **50**, 742–752. (doi:10.1080/106351501753328866)
33. Goloboff PA, Farris JS, Källersjö, M., Oxelman B, Ramírez MJ, Szumik CA. 2003 Improvements to resampling measures of group support. *Cladistics* **19**, 324–332. (doi:10.1016/S0748-3007(03)00060-4)
34. Kopuchian C, Ramírez MJ. 2010 Behaviour of resampling methods under different weighting schemes, measures and variable resampling strengths. *Cladistics* **26**, 86–97. (doi:10.1111/j.1096-0031.2009.00269.x)
35. Goloboff PA, Szumik CA. 2015 Identifying unstable taxa: efficient implementation of triplet-based measures of stability, and comparison with Phyutility and RogueNaRok. *Mol. Phylo. Evol.* **88**, 93–104. (doi:10.1016/j.ympev.2015.04.003)
36. Lee MSY, Soubrier J, Edgecombe GD. 2013 Rates of phenotypic and genomic evolution during the Cambrian explosion. *Curr. Biol.* **23**, 1889–1895. (doi:10.1016/j.cub.2013.07.055)
37. Zhang C, Stadler T, Klopfstein S, Heath TA, Ronquist F. 2016 Total-evidence dating under the fossilized birth–death process. *Syst. Biol.* **65**, 228–249. (doi:10.1093/sysbio/syv080)
38. Kass RE, Raftery AE. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. (doi:10.1080/01621459.1995.10476572)