WILEY-BLACKWELL

Cladistics (2019) 1–22

Cladistics

10.1111/cla.12380

Likelihood approximations of implied weights parsimony can be selected over the Mk model by the Akaike information criterion

Pablo A. Goloboff* and J. Salvador Arias

Unidad Ejecutora Lillo, Consejo Nacional de Investigaciones Científicas y Técnicas, Fundación Miguel Lillo, Miguel Lillo 251, 4000 S.M. de Tucumán, Argentina

Accepted 26 February 2019

Abstract

A likelihood method that approximates the behaviour of implied weighting is described. This approach provides a likelihood perspective on several aspects of implied weighting, such as guidance for the choice of concavity values, a justification to use different concavities for different numbers of taxa, and a natural basis for extended implied weighting. In this approach, the number of free parameters in the estimation depends on *C*, the number of characters (in contrast to the standard Mk model, which estimates *2T–3* parameters for *T* taxa). Depending on the characteristics of the dataset, the likelihood obtained with this approach may in some cases be similar or superior to that of the Mk model, but with fewer parameters being adjusted. Because of that tradeoff, testing against the Mk model by means of the Akaike information criterion on a set of 182 morphological datasets indicated many cases (36) in which the likelihood approximation to implied weighting is the best method, from an information-theoretic point of view. Given that it is expected to produce (almost) the same results as this maximum-likelihood approximation, implied weighting can therefore be seen as a valid alternative to the Mk model often used for morphological datasets, on the basis of a criterion for model fit widely advocated by likelihoodists.

Introduction

Although methods of phylogenetic inference (maximum-likelihood, ML) based on likelihood and detailed models of evolution assuming a common mechanism are currently widespread, the arguments that have been advanced for a universal preference of these methods over parsimony (maximum-parsimony, MP) are not entirely satisfactory. This is especially true in the case of phylogenetic analyses of morphological characters (where the Mk model of Lewis, 2001 is often used in ML or Bayesian analyses), but a case could be made as well for MP methods in the case of molecular sequences (e.g. Goloboff et al., 2017: 433).

Several lines of argument have been advanced against MP by proponents of models. One is based on simulations (e.g. Wright and Hillis, 2014; O'Reilly et al., 2016; Puttick et al., 2017), using a known model tree to generate datasets and studying the degree to which the trees inferred by different methods approximate the model tree. Goloboff et al. (2017, 2018a) replied to these arguments, showing that (depending on the model used to generate the data) either parsimony or model-based methods may perform best. And, in addition to generating their datasets under a common mechanism (which favours model-based inference), Puttick et al. (2017) only reported comparisons for the worst options (*k* values) of implied weighting parsimony they examined. Goloboff et al. (2018b) and Smith (2019) report that, even for datasets generated under the Mk model with a common mechanism, the results of parsimony and model-based methods are of almost the same quality when comparisons are performed more carefully.

A second line dismisses MP on the grounds that the assumptions made by model-based methods are biologically (or biochemically) justified and make these

*Corresponding author:
E-mail address: pablogolo@csnat.unt.edu.ar

methods superior (e.g. Steel, 2005; Huelsenbeck et al., 2011). However, Goloboff et al. (2018b) showed, with a hypothesis testing approach, that empirical datasets reject the common mechanism assumed by standard ML methods (particularly for morphology, but also in a significant fraction of molecular datasets). That finding of Goloboff et al. (2018b) both weakens the idea that the Mk model reflects a biological reality and shows that simulations based on assuming a common mechanism are of little empirical relevance.

A third line of reasoning to defend ML methods is that the estimations based on MP are too highly parameterized (Lewis, 2001; Steel, 2002). Some early authors had considered that ML is preferable over MP, because MP is *too simplistic a model* (e.g. Yang, 1996). However, after Tuffley and Steel (1997) proposed the no-common-mechanism (NCM) model, some authors continued to favour ML, now on the grounds that MP is *too complex a model* (e.g. Lewis, 2001; Huelsenbeck et al., 2008, 2011). In the NCM model, the length of every branch of the tree is independently set for every character, causing ML to select the exact same trees as MP, but requiring estimation of a large number of parameters. The Akaike information criterion (AIC; Akaike, 1973; Posada and Buckley, 2004) is widely used to evaluate models, based on a tradeoff between model fit and complexity. Holder et al. (2010) showed that, given the large number of parameters in NCM, it is numerically impossible for the AIC to select NCM over standard ML models (as a result of the interplay between possible increases in likelihood by freeing branch lengths to vary for each character, and the number of parameters added to the estimation).

While it is true that the NCM model of Tuffley and Steel (1997) is very parameter-rich, alternative formulations of MP involving simpler estimations are in fact possible (e.g. Goloboff, 2003). Huelsenbeck et al. (2008: 406), for example, noted that "the Tuffley and Steel (1997) model is just one of several that gives a correspondence between the parsimony and maximum likelihood methods." Holder et al. (2010: 478–479) were also aware of this, and they explicitly noted that their result should not be interpreted as indicating that the AIC will always prefer ML over *any* formulation of MP—just over the NCM formulation of MP. However, despite the warning of Holder et al. (2010), the idea that MP always requires estimating too many parameters has been overgeneralized by supporters of model-based methods. For example, O'Reilly et al. (2018: 632) mistakenly claimed that Goloboff et al.'s (2017) simulations assumed the NCM model and were thus vitiated by using a large number of parameters. This claim conflates the number of parameters used in a *simulation* with that used in an *estimation* process, but it is incorrect even from that perspective: Goloboff et al. (2017) used MP to infer trees, but their data were neither generated nor analysed with NCM.[1] O'Reilly et al. (2018) clearly paid no heed to Huelsenbeck et al.'s (2008) or Holder et al.'s (2010) warnings about MP not being a synonym of NCM.

Despite the fact that alternative implementations of MP might be more favourably compared to standard ML by different model-selection methods, no actual comparison has ever been published. Such comparison is made difficult by the scarcity of implementations that produce results similar to parsimony. The aim of this paper is to conduct that comparison.

First, we consider less highly parameterized Poisson models that behave similarly to MP—both equal weights parsimony (EWP) and implied weights parsimony (IWP; Goloboff, 1993). IWP used to be viewed rather favourably by model-based phylogeneticists (e.g. Ronquist et al., 1999; Nylander et al., 2004), but more recently has been questioned (Congreve and Lamsdell, 2016; O'Reilly et al., 2016; Puttick et al., 2017). The model proposed here shows that the basic premises of IWP can be justified from a likelihood perspective, and helps illuminate several aspects of implied weighting: the choice of weighting strength, differences in weighting strengths that may be required by different numbers of taxa (discussed by Goloboff et al., 2008a), and a natural justification to collectively weight partitions (as in "extended" implied weighting; Goloboff, 2013).

In the second part of the paper, we discuss the minor differences to be expected between the results of these approximate methods and EWP and IWP. In particular, we discuss the apparent inability of Poisson-based models to directly take into account unexplained similarity (which IWP does; see De Laet, 2005), and the sensitivity to the size of the state space assumed for all methods applicable to morphology. We demonstrate that these differences are minor and, within certain circumstances, the results of EWP and IWP are expected to closely resemble those of their Poisson counterparts. This second part of the paper is more technical, and readers less interested in details can skip it to go directly to the third section.

In the third part, we apply these implementations to a sample of empirical datasets, and show that (contrary to the results of Holder et al., 2010 with NCM) the model that approximates IWP is often selected for morphological datasets, and a model that collectively

[1] Ironically, the method used by Goloboff et al. (2017) for generating their data (i.e. independently allocating character changes equiprobably on any branch) is *identical* to that used by Puttick et al. (2018), the same set of authors as O'Reilly et al. (2018). Either Puttick et al. (2018) failed to realize the similarity, or they believe that generating data in such a way implies a large number of parameters only when the authors do not profess a preference for model-based methods.

weights characters by partition is sometimes selected for molecular datasets.

Models approaching equal weights and implied weights parsimony

Generalities

The different ML models proposed in this paper can approach the results of EWP and IWP with varying degrees of precision. The models that are based on using the likelihood from the most likely individual reconstruction approximate the results of their parsimony counterparts more closely; these single-reconstruction methods are denoted by prepending the letter *s* (i.e. sMP_{lik} and *simplik*). The variants not so prefixed (i.e. MP_{lik} and *implik*), based on summing the likelihood for all reconstructions for each character (as in standard ML methods), produce results that are very close, but not identical, to those of their parsimony counterparts. But their likelihood values are more precise, and therefore it is these variants that are used for the model selection tests. The crucial difference between standard models, and the models proposed here to approach EWP and IWP, is that the latter use the same length for all the branches of the tree.

Basic assumptions

Only the basic aspects are covered here; for a more in-depth coverage of methods used in ML, see Swofford et al. (1996) and Felsenstein (2004). The models used here are, just as in standard likelihood models, homogeneous Poisson models; they are also Markov processes (as lineages become independent after splitting). The idea is to emulate EWP and IWP for binary and non-additive characters, and then the probabilities of transformation between all states must be the same (parsimony allows for differential costs of transformations, but that case is not considered here). This is the same approach as in the Mk model (Lewis, 2001; Mkv is just a variant of Mk that takes into account ascertainment bias), as well as in JC69 (Jukes and Cantor, 1969) or the more general Neyman model (Neyman, 1971). That is, the probability of starting at state *j* and ending in state *k* in a character with *s* states along a given branch of length *b* is:

$$\begin{aligned} P_{jk} &= \frac{1}{s} + \frac{s-1}{s}\, e^{-b} \quad (\text{if } j = k, \text{ stasis}) \\ P_{jk} &= \frac{1}{s} - \frac{1}{s}\, e^{-b} \quad (\text{if } j \neq k, \text{ change}) \end{aligned} \qquad \text{[Formula1]}$$

The length of a branch is the expected number of substitutions per character along the branch (the product of instantaneous rate and time; Swofford et al., 1996; Lewis, 2001), taking into account that multiple substitutions may nonetheless produce the same final state. Note that these probabilities depend on the number of states; as *s* is larger, it is less likely that multiple substitutions (along the same branch, or different branches) produce the same final state.

Standard models: heterogeneous branch lengths and average likelihood

The models with heterogeneous branch lengths are derived from Felsenstein (1981a): they consider that the length *b* is common to all the characters in the dataset (or partition) and differs among branches of the tree, and they obtain the likelihood by summing over all reconstructions for each character (i.e. all possible paths to the data; this is often referred to as integrated likelihood, as in Huelsenbeck et al., 2008). We refer to this combination, in the remainder of the paper, as *standard models*. The branch lengths of the tree are chosen so that the likelihood is maximized. Given that the likelihood for the dataset is the product of the likelihoods of the individual characters, which may become rather small when many characters are present, the logarithms are normally used for each character, so that:

$$\ln L = \sum \ln l_i$$

where L is the likelihood of the entire tree, and l_i is the likelihood of character *i*. As likelihoods vary between 0 and 1, log-likelihoods are negative numbers, so the negative likelihood is used as a number to be minimized (just as the parsimony score).

These standard models allow for some characters to evolve at a faster (or slower) rate, but all characters are sped up (or slowed down) at the same branches by the same exponential factor (the branch length). The most common way to take into account rate heterogeneity is the discretized gamma distribution (Yang, 1994), which allows a variety of shapes with a single parameter. The probabilities of transformation between two states *j*,*k* for a character with rate *r* are then:

$$\begin{aligned} P_{jk} &= \frac{1}{s} + \frac{s-1}{s}\, e^{-rb} \quad (\text{if } j = k, \text{ stasis}) \\ P_{jk} &= \frac{1}{s} - \frac{1}{s}\, e^{-rb} \quad (\text{if } j \neq k, \text{ change}) \end{aligned}$$

The likelihood for each of the rate categories is obtained by multiplying the probability of the observed pattern by the probability that a character belongs to the given category. The gamma parameter is set to the value that maximizes the likelihood summing over all rate categories (clearly, when all characters evolve with exactly the same rate, the best

likelihood will be obtained with a gamma that concentrates all the distribution at a single value).

A likelihood justification for a different form of character weighting (i.e. cliques and threshold methods) has been proposed by Felsenstein (1981b). The assumptions of Felsenstein (1981b) have been strongly criticized by Farris (1983: 32–34) and are substantially different from those presented here. The main difference with the present approach is that Felsenstein (1981b) did not invoke uniformity of branch lengths, and expected some fraction of the characters to exhibit very high homoplasy (then effectively giving some characters zero weight).

Models for equal weights parsimony: sMP_{lik} and MP_{lik}

sMP_{lik}. One model for EWP results from Goldman's (1990) work. Goldman (1990) showed that, for probabilities of change (p) and stasis (q) between two states fixed for all the branches of a given tree, the single reconstruction that maximizes the overall probability of the observed state distribution for each character is (as long as $q > p$) the one that minimizes the number of steps; therefore, the most parsimonious tree maximizes this probability over all characters. For a given reconstruction, with n branches where the character is transformed, and u branches where it is not, the individual likelihood of character i is then $l_i = q^u p^n$, and l_i is maximum when n (the number of transformations) is minimum. Goldman (1990) was interested mostly in establishing conditions of equivalence between parsimony and likelihood, so he did not discuss the values of q, p that (for a given value of n) maximize l_i. For characters with s states, these probabilities are simply

$$q = u/(n+u)$$
$$p = (1-q)/(s-1) \qquad \text{[Formula 2]}$$

(note that the formula for p reduces to $p = 1 - q$ when there are only two states). The branch length that produces these values can be obtained by solving the equation of probability of stasis (in Formula 1),

$$b = \ln\left(\frac{s-1}{s}\right) - \ln\left(q - \frac{1}{s}\right) \qquad \text{[Formula 3]}$$

This method estimates a single parameter for all characters and all branches; the method is highly constrained, not allowing for longer or shorter branches, or faster and slower characters [i.e. in agreement with Yang's (1996) characterization of EWP as a very simplistic model]. We thus refer to this method (identical to Goldman's, except that the branch length parameter is optimized) as sMP_{lik} (for *single-reconstruction MP under likelihood*) and, as long as the state space (see next section, under "State space and homoplasy") is the same for all characters, it is exactly equivalent to EWP, producing the same ordering for any set of trees.

MP_{lik}. The likelihood for individual characters, in standard Poisson models for phylogenetics, is not calculated from the single reconstruction with highest likelihood, but instead from the sum of likelihoods of all possible reconstructions. We refer to this as the MP_{lik} method (for *MP under likelihood*): calculate the likelihood under a single length common to all branches (and characters) in the tree, chosen so as to maximize the likelihood, summing the likelihood for all possible reconstructions. Note that (even if both approaches use the likelihood that results from summing all reconstructions), MP_{lik} differs from the standard likelihood model in all the branches of the tree having the same length. For large numbers of characters with a common state space, when the length of all branches is constrained to be the same, the tree of highest likelihood will converge to the EWP tree and (if the data have been generated by a model tree with identical branch lengths, below a certain threshold value) this will be identical to the model tree. Steel (1989) and Kim (1996) showed that even with large numbers of characters and constant branch lengths, the EWP tree may be different from the MP_{lik} tree. In simulation experiments, we have confirmed that (for large numbers of characters) model trees with constant branch lengths are consistently recovered by both EWP and MP_{lik} for very long branches (even a length of 2, i.e. an average of two changes per character per branch, consistently produces the model tree from sufficiently large numbers of characters). Note that this does not mean that EWP and MP_{lik} produce the same results for *any* possible input; they converge when the number of characters is large, but may produce different trees for some specific data inputs (see examples in next section, under "More steps, higher likelihood?"). The correspondence between EWP and MP_{lik} is therefore close, but only approximate.

In standard models, where different branches of a tree can have different lengths, branch lengths are chosen so that the sum of likelihoods for all reconstructions is maximized (via Formula 1). This is a heuristic process that needs to be performed iteratively on all the branches of the tree, which can be made efficiently thanks to the "pruning" and "pulley" algorithms of Felsenstein (1981a). In the case of MP_{lik}, where branch lengths are the same for all the branches of the tree, the pulley algorithm is unnecessary (i.e. the maximization does not involve "pulling" down the tree to different rootings to optimize branch lengths one at a time, hence requiring simpler computations). Thus, in MP_{lik}, the pruning algorithm is simply used repeatedly to

calculate total likelihoods for different branch lengths. The branch lengths calculated with Formula 3 provide an initial approximation, and the best value is found heuristically (with any desired degree of approximation, as the function is continuous and has a single optimum); in the implementation used here, we have used 10^{-5} for the difference in probability of stasis as the limit.

Models for Implied Weights: simplik and *implik*

The preceding section considered the use of a single branch length common to all characters. Using the same length for all branches of the tree amounts to a model where transformations for each individual character can be equiprobably located on any branch of the tree. Parsimony has traditionally been justified on the basis of making few assumptions (e.g. Farris, 1983), and that idea is compatible with considering, a priori, that character changes could equally well occur on any given branch. As discussed by Goloboff et al. (2018b), the notion that transformations within a given character can be equiprobably located on any branch is a statement on the product of evolution, not a statement on a specific process; and this product may result from different processes. A specific process is needed to assign actual likelihoods to the result, and this is here achieved by assuming a Poisson process and constraining lengths of all branches to be equal. The same equiprobable distribution, however, could be achieved by generating data from a model tree where the lengths vary independently for each branch and character (as in NCM; Tuffley and Steel, 1997), or by just randomly distributing changes over branches (as done by Goloboff et al., 2017, the same "model" used by Puttick et al., 2018).

One of the widely used variants of parsimony is implied weighting (Goloboff, 1993), where trees are compared according to the implied reliability of the characters (with reliability considered as a decreasing function of homoplasy); this leads to a scoring function that increases by smaller differences for each of the successive steps of a character. Although parsimony seems to require that transformations for a given character are placed equiprobably on any *branch* of the tree, the requirement that transformations are equally likely to occur in any of the *characters* can be eliminated, and (from the likelihood perspective) this automatically produces a method that behaves almost exactly as IWP.

Simplik. Consider first the case of likelihood evaluated using the optimal reconstruction, as in Goldman's (1990) approach. If some characters are more likely to change than others, then the branch length should be separately adjusted for each character, and the obvious criterion to select branch lengths is maximizing the likelihood for each individual character. For invariant characters, the optimal branch length is 0, so that $P_{(\text{stasis})} = 1$; the individual likelihood for an invariant character is thus a constant for the different trees, therefore having no effect on tree choice. As variable characters with different numbers of steps are considered, the branches have to be adjusted (following Formulae 2–3) independently for each character. The increase in log-likelihood for an additional step becomes smaller and smaller as the character has more steps, as branches become longer (to accommodate for additional steps). When (estimated) branch lengths are short, then a transformation is improbable, and another transformation strongly decreases the likelihood (the product of the probabilities along all branches), and vice versa. Thus, the first steps of homoplasy are more costly than subsequent ones. This is exactly how IWP operates, and so adjusting branch lengths individually for each character (while constraining all tree branches to have the same length, and using the single reconstruction of highest likelihood) is an ML equivalent of IWP. We refer to this method as *simplik* (for *single reconstruction implied weighting under likelihood*).

By default, TNT (Goloboff et al., 2008b; Goloboff and Catalano, 2016), the main program for IWP, uses the complement of the formula originally proposed by Goloboff (1993), choosing the tree that minimizes $\sum h_i/(k + h_i)$ (where h_i is the homoplasy of character i, and k is a constant of concavity, with larger values weighting less strongly against characters with homoplasy; see Goloboff, 1995). Figure 1 compares the weights derived from *simplik*, for different numbers of taxa, and those resulting from the original formula for IWP with different values of k. Goloboff et al. (2008a) had proposed a possible way to set the weighting strength for different numbers of taxa (a possibility first raised by Goloboff, 1993); *simplik* provides another way to naturally rescale the weighting function, based on the Poisson substitution process assumed by likelihood models. As can be seen from Fig. 1, the *simplik* weighting function down-weights extra steps much faster for fewer taxa. The default formula for IWP produces weights that are very similar to those of *simplik*, when mild concavity values are used. As a quick approximation, in Fig. 1 we have used a value of k equal to half the number of taxa; this is a much milder weighting function than normally used. The similarity between the two methods is quite close, even if using the default formula for weighting. Figure 2a shows (for a dataset with 60 taxa and 500 characters generated under the Mk model, with very unequal branch lengths, as in Goloboff et al., 2018b) the correlation between the default weighting function of TNT (with $k = 30$), and the *simplik* score, in 15 000

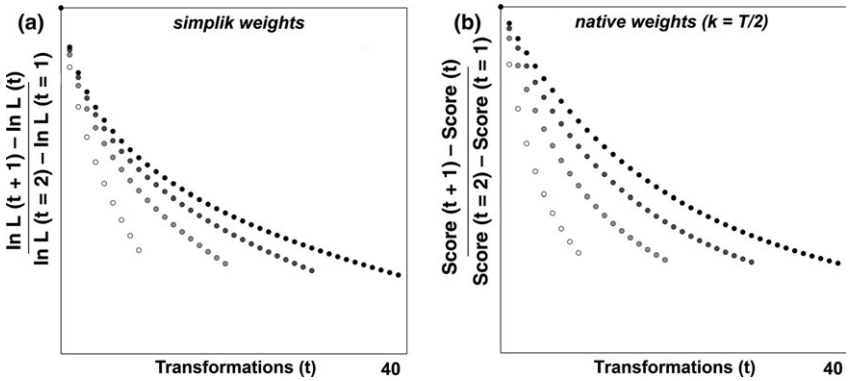

Fig. 1. Comparison of relative weights determined by (a) *simplik*, with (b) the relative weights determined with the default formula of TNT (with a value of k equal to half the number of taxa), for 20, 40, 60 and 80 taxa (darker shading indicates more taxa). The x-axis displays number of steps or transformations (in a two-state character), while the y-axis displays the relative weights (normalized so that the first step of homoplasy costs unity).

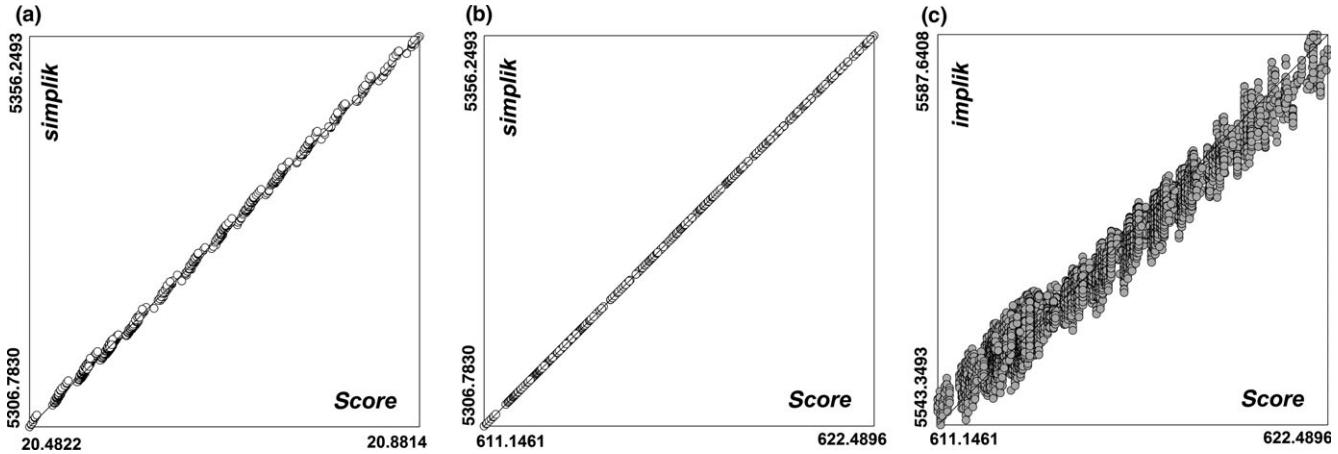

Fig. 2. Plots of different approximations to implied weights. The scores were calculated on 15 000 trees (optimal and near optimal under implied weights), for a simulated dataset with 60 taxa and 500 characters (with two states, generated under the Mk model, with branch length determined as $x^2/3000$, where x is a random number taken from the interval 1–20). This leads to model trees with significant differences in branch lengths and makes for a lower correspondence between the values of implied weighting and the *implik* approximation. Values on the y-axis are log likelihoods. (a) Values of *simplik* scores (y-axis) plotted against implied-weighting scores with the default weighting function with $k = 30$ (x-axis). The correlation (R) is 0.9995. (b) Values of *simplik* scores (y-axis) plotted against implied weighting scores with relative weights as in *simplik* (i.e. as in Fig. 1a); the correspondence is exact. (c) Values of *implik* scores (y-axis) plotted against implied weighting scores with relative weights determined from *simplik* (i.e. as in Fig. 1a); the correlation (R) is 0.9825.

trees (optimal and near optimal under IWP). Given the similarity in the curve shapes observable in Fig. 1, both IWP (with $k = 30$) and the *simplik* method order the trees in almost exactly the same sequence. Given that TNT allows the user to define any weighting function for IWP, the values from *simplik* can then be used in TNT as relative weights, producing an exact correspondence between *simplik* and IWP when all characters have the same number of states (Fig. 2b) and no missing entries are present (see below). The correspondence is only approximate when missing entries are present or different characters have different numbers of states (this is so even if the state space assumed is the same for all characters, as the treatment of homoplasy is different in both methods; see next section for examples and discussion).

Implik. The method described in the preceding subsection uses the likelihood of the optimal reconstruction to evaluate the trees. The standard approach in model-based phylogenetics uses the *sum* of likelihoods of all reconstructions instead of the maximum. We refer to the approach of optimizing branch length independently for each character, using the likelihood from all reconstructions, as *implik* (for *implied weighting under likelihood*). For possible character patterns, the correct probabilities under the assumed model (i.e. the probability of actually

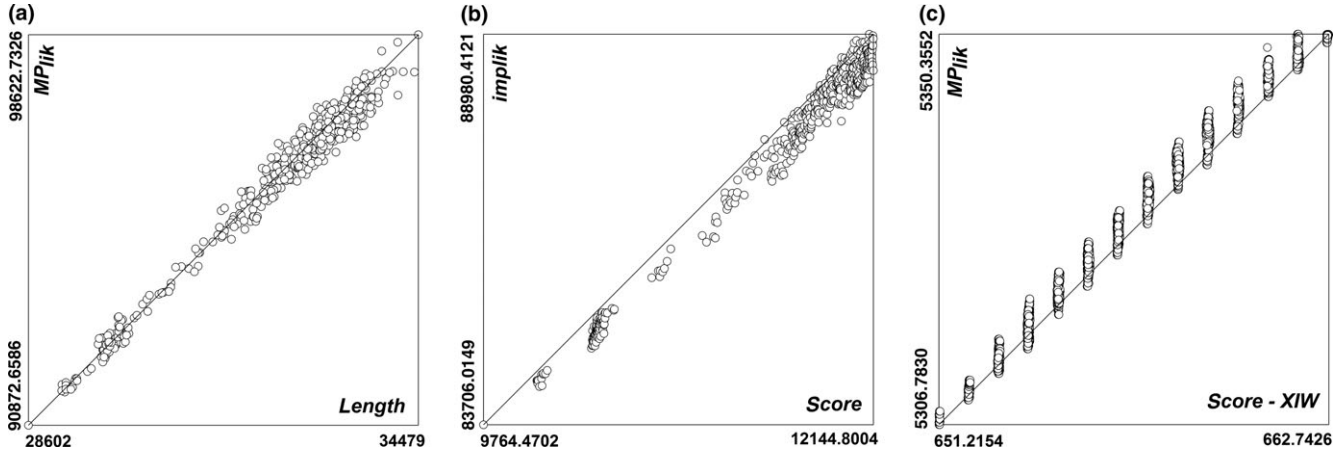

Fig. 3. Plots of different approximation to equal weights parsimony and implied-weights parsimony. Values on the *y*-axis are log likelihoods. For (a) and (b), the scores were calculated on 500 trees (optimal and near optimal under implied weights), for a simulated dataset with 10 taxa and 15 000 characters (with two states, generated using an exponential function with $\lambda = 0.15$, which leads to a better correspondence between the values of parsimony and their likelihood approximations). For (c), the scores were calculated on the same dataset and trees as in Fig. 2. (a) Values of MP_{lik} scores (*y*-axis) plotted against parsimony length (*x*-axis); correlation (*R*) is 0.9912. (b) Values of *implik* scores (*y*-axis) plotted against implied-weighting scores with relative weights as in *simplik* (i.e. as in Fig. 1a); correlation (*R*) is 0.9932. (c) Values of MP_{lik} scores calculated collectively for four groups of characters (*y*-axis) plotted against extended implied weighting scores (*x*-axis) for the same character groups.

generating each pattern under the model) are produced only by this sum (Goloboff, 2003: 100), which considers all possible paths to the data. The correspondence between IWP and the method based on summing up the likelihood of all paths to the data is not exact, but it is quite close (Fig. 2c), and both methods are based on the idea that (on average) it costs less to add homoplasy to characters with more steps. Evaluating different models requires us to compare the likelihoods they produce, as well as the number of estimated parameters. Thus, comparisons for model selection are better carried out using similar (and more accurate, under the Poisson model assumed) methods to calculate the likelihoods. Just as in the case of EWP, the results of using the likelihood from a single optimal reconstruction, or the sum from all possible reconstructions, converge to ranking the trees in the same sequence when all the branch lengths of the model tree are uniform and there are large numbers of characters (e.g. as in Fig. 3a), the results of using a single or all possible reconstructions converge as well in the case of branch lengths optimized independently for each character (Fig. 3b) when the model tree has uniform branch lengths and there are many characters. For *implik*, the ordering of the 15 000 trees is not exactly the same as IWP, but the results are strongly correlated (Fig. 3b). Therefore, implied weighting provides a close approximation to the results of *implik*.

The approximate correspondence between the results for *implik* or *simplik* on the one hand, and IWP on the other, is interesting from several points of view. Not only does this correspondence provide a natural likelihood justification for IWP (for those inclined to think that only model-based justifications are legitimate), but also provides additional rationale for aspects of IWP that have so far been justified only from the perspective of parsimony. The guidance for the choice of k values has been discussed above. This correspondence also provides justification for the idea of applying IWP to blocks of data, a method proposed by Goloboff (2013) as extended implied weighting, which is potentially useful for molecular sequences (where the original formulation of IWP may not be the best method; for discussion see Goloboff et al., 2008a; Goloboff, 2013). While block weighting requires special calculations under IWP, achieving this in the present context only requires that the likelihoods are calculated by finding a single optimal branch length for each block, as in sMP_{lik} or MP_{lik}, then multiplying the likelihoods for each block (or summing the log-likelihoods), to obtain the total likelihood (or log-likelihood). Transformations in the block of data with longer optimized branch lengths will be less costly (i.e. more probable), simply as a by-product of the application of sMP_{lik} or MP_{lik}. Figure 3c shows the result of applying both extended IWP and sMP_{lik}, for the same four blocks of data, to a set of trees for the dataset of Fig. 2, showing again a close correspondence.

Additional properties of likelihood approximations of EWP and IWP

Number of parameters

The number of parameters that need to be estimated under the single-reconstruction approach used in parsimony and the sMP_{lik} and *simplik* variants is

controversial. Felsenstein (1978) first showed that parsimony may be inconsistent when all characters evolve with common rates and the length of tree branches differs substantially. He attributed (Felsenstein, 1978: 408–409) the inconsistency of parsimony to the fact that it estimates the cost from the best possible individual reconstruction, arguing that (for a tree of T taxa) this amounts to assigning a specific ancestral state to each of the T–2 internal nodes of the tree for each of the C characters, in his view requiring estimation of a number of parameters that keeps increasing as taxa or characters are added to the dataset.

The viewpoint of Felsenstein (1978) was contested by Farris (1986: 22–23), and Felsenstein (1987: 208) agreed that "it is not obvious" whether estimating C (T–2) parameters is required for approaches that estimate likelihood from the best individual reconstruction. Goldman (1990: 350) likewise recognized that ancestral states "are not parameters of the evolutionary process, but random variables: particular realizations of parts of the process", but they could nonetheless be treated "as though they were" parameters of the model. However, the hesitation in considering the use of optimal ancestral reconstructions as requiring estimation of an inordinate number of parameters magically evaporates when the argument can be used to criticize parsimony, and the argument is presented with absolute certitude (e.g. Lewis, 2001: 914; Holder et al., 2010: 479–480; Huelsenbeck et al., 2011: 225).

Goloboff (2003: 99–101) argued that using the best possible likelihood of all reconstructions does not amount to a specific reconstruction of ancestral states. The inconsistency when using the score from the best reconstruction (be it likelihood, or parsimony) results from the fact that the best reconstruction alone does not provide the correct probabilities of the observed distribution of states among the terminals for the model in question; only the sum of all likelihoods (possible paths) provides this. If a reconstruction was a parameter, then there would be a reconstruction that would produce the correct value for the probability of generating the observed state distribution under the model, and only the sum for all reconstructions provides that value. The likelihood of all possible reconstructions is also (implicitly) considered both in the case of parsimony and in standard likelihood. These values are all summed up in standard likelihood, and just one value is selected in the case of parsimony. Selecting the value of likelihood of the best possible reconstruction is not the same thing as selecting the best reconstruction as such (much like an apparent synapomorphy in two sister taxa being in fact due to independent parallel acquisition does not make the tree false; see Farris, 1983: 13–14). Interestingly, using the likelihood of the best possible reconstruction causes inconsistency only when the data are generated under certain models (and, as noted by Goloboff, 2003, when the specification of a reconstruction has to select from among fewer alternatives, given that increasing the number of possible states makes parsimony consistent). Otherwise, the identification of ancestral conditions does not cause any estimation problems. If the inconsistency was indeed caused by reconstructing specific ancestral states for each character and node, then evaluating trees with the likelihood of the best reconstruction should produce inconsistency always, not only under some specific models.

In standard likelihood calculations, the likelihood of each individual reconstruction is multiplied by the prior probability ($1/s$, in the case of the symmetrical models used for morphology) that one of the s states at the root actually occurs there (this calculation is easily incorporated into the pruning algorithm, see Swofford et al., 1996). Considering the effect of such multiplication by the prior in the case of single-reconstruction algorithms also supports the idea that the number of free parameters is really not increasing when the likelihood is estimated from the best possible reconstruction. With such multiplication, the values of likelihood obtained from the best reconstruction (*simplik*) will always be smaller than those obtained when summing up reconstructions (*implik*). The two values will converge only when branch lengths are very short, because in that case most of the likelihood contribution will come exclusively from the MP reconstructions (Felsenstein, 1981b). If the number of parameters was really larger in the single-reconstruction case, one would expect (this method being based on the same model of character change as the multiple reconstruction one) that the likelihood be *increased* as more parameters can be tuned to increase fit—yet the opposite happens.

Then, we consider that the number of parameters in the single-reconstruction approaches does not increase with the number of nodes in the tree. The problem with using the single reconstruction approaches for testing the fit of different models is not the number of parameters, but instead that the resulting values of likelihood are only approximate. The values obtained from the single-reconstruction approach are near the (correct) values only when the likelihood of the optimal reconstruction is not divided by the prior, and even then, only approximately (Goloboff, 2003). Thus, to obtain accurate comparisons using model-selection methods on empirical datasets, it seems better to use the likelihoods obtained from MP_{lik} and *implik* (instead of sMP_{lik} and *simplik*), and the number of estimated parameters is in this case uncontroversial. A general comparison of the approaches examined here, and the numbers of parameters to adjust for each, is shown in Table 1. The number of parameters

Table 1
Comparison between basic aspects of the different approaches explored in this paper. In the "Parameter no." column, T = the number of taxa, and C = the number of characters. For sMP_{lik} and *simplik*, the number of parameters estimated is controversial; for each of those, the number proposed in this paper is shown first, followed by the alternative (less preferred) number of parameters in italics

Method	Branch lengths	Emulates	Reconstructions	Parameter no.	Reference
Mk	Different for each branch, same for all characters	JC69	Average	$2T - 3$	Lewis (2001)
Mk + Γ	Different for each branch, same for all characters	JC69 + Γ	Average	$2T - 2$	Lewis (2001)
sMP_{lik}	Same for all branches and characters	Equal weights parsimony	Optimal	1 *C(T – 2)?*	Goldman (1990)
MP_{lik}	Same for all branches and characters	Equal weights parsimony	Average	1	Yang (1996)
Simplik	Different for each character, same for all branches	Implied weights	Optimal	C *C(T – 2)?*	This paper
Implik	Different for each character, same for all branches	Implied weights	Average	C	This paper
Partition MP_{lik}	Same for all branches and characters within each partition	Extended implied weights	Average	Number of partitions	This paper

estimated by the standard Mk or JC69 models for a tree of T taxa is then *2T–3* (i.e. the number of branches in the tree); in the case of gamma-distributed rate heterogeneity, one additional parameter is included (the value of α). In the case of *implik*, the number of parameters equals the number of characters, C. In the case of MP_{lik}, a single parameter (the unique branch length) needs to be estimated; this highly constrained model will naturally tend to produce low likelihoods, but the low number of parameters might make it a viable candidate in some cases.

The numbers of parameters to estimate under each approach indicate that *implik*, the likelihood analogue of IWP, may (other things being equal) tend to be selected over the Mk model when the dataset has few characters relative to the number of taxa (i.e. when $C < 2T$–3). Indeed, the application of IWP to matrices with few taxa (e.g. fewer than 15–20 taxa) rarely produces results different from EWP (as noted by Goloboff, 1997: 237–238), thus indicating that the application of IWP is in that case superfluous (as expected from considering the scant opportunity for accurately assessing homoplasy that such small trees provide). This is why the analyses of Goloboff et al. (2008a, 2017) only examined datasets where $T \geq 50$. General consideration of the number of parameters estimated by the different approaches indicates the same general picture. In the typical molecular dataset, with large numbers of characters relative to numbers of taxa, the opposite is true: the analogue of IWP, *implik*, is less likely to be selected as an appropriate model. When $C > 2T$–3, *implik* will be selected as an appropriate model only if the likelihood is increased by a very large factor over that of standard models with a common mechanism for all characters.

State space

The number of possible states a character can take, or *state space*, needs to be carefully considered in the case of likelihood methods for morphology. If the state space is different for the different characters in the matrix, an additional element of discordance between the likelihood approximations used here and their parsimony counterparts is introduced: the probability of transformation between states then changes with the number of states, implying different prior weights for the characters with different numbers of states. The size of the state space is naturally fixed in the case of molecular sequences, but it is problematic in the case of morphology. For the Mk model, both PAUP* (Swofford, 2002) and MrBayes (Ronquist et al., 2012) use the same state space for all characters, based on the largest observed state in the matrix. This is probably the best course of action, because it avoids

different prior weights solely on the basis of observed numbers of states, but there seems to be no reason for this, other than making the relative contributions of each character equal a priori.

For morphology, it seems hard to justify any specific choice of a state space on the grounds of an empirical reality, instead of purely methodological considerations. Why should the state space of a character known to vary among only two possible conditions be considered as comprising many more options? During the actual evaluation of possible reconstructions for a character "spine", with states "0, present" and "1, absent", assuming a larger state space (s) would imply that conditions 2, 3, …, s–1 are considered as well as possible ancestral assignments—even when those assignments would be undefined or impossible. This reflects, in our view, a dilemma (so far unsolved) in the choice of state space for morphology in model-based phylogenetics. Using a state space that varies among characters may be more meaningful in biological terms, but this would make some characters more influential than others, on the basis of properties (such as the mere number of observed or possible conditions) that are not obviously related to the reliability of the characters.

Methodological considerations therefore indicate that the best choice probably is (as done by PAUP* and MrBayes) a uniform state space for all characters. However, a problem remains: why should the state space be determined by the largest observed state in the *matrix*? Why not use a much larger state space? After all, it is possible that either (i) characters not yet included in the matrix have a much larger number of conditions, or that (ii) taxa not yet included in the matrix have additional states not present in the matrix with the current taxon selection. In addition, the number of alternative forms many morphological characters could take is potentially very large (e.g. in characters related to shape rather than presence/absence), even if many of those forms have never been realized in the course of evolution. These considerations suggest that in the case of morphology perhaps a *large*, *common* state space is the most appropriate choice.

As the size of the state space increases, it is well known that the differences between EWP and standard likelihood narrow (Felsenstein, 1978; Farris, 1983; Steel and Penny, 2000, 2004; Goloboff et al., 2017). Figure 4 shows this for EWP and the Mk model. For two states assumed as possible conditions, the optimal trees under EWP—white arrow–differ from the optimal trees under Mk—grey arrow–but for five or more states both Mk and EWP select the same trees as optimal, with all the trees located closer to the diagonal indicating identity between the two approaches as the number of states increases. More relevant for the present paper, a similar phenomenon also occurs when a single length is used for all the branches of the tree, so that EWP converges more closely to MP_{lik} (Fig. 5), and IWP converges more closely to *implik* (Fig. 6). Increasing the state space s for likelihood calculations has the disadvantage that the times needed for scoring trees increase with s^2 (as in Sankoff parsimony; Sankoff and Rousseau, 1975), but in that case the quality of the approximations EWP≈MP_{lik} and IWP≈*implik* increases as well, and the times needed for calculating trees under EWP or IWP do not increase with the numbers of states.[2]

Note that Wheeler (2016: 224) considered that the NCM approach of Tuffley and Steel (1997) "can be viewed as a likelihood-based character weighting scheme in parsimony analyses". He was referring to prior weighting on the basis of different number of states, not on the basis of homoplasy or different numbers of steps. Under NCM, the likelihood of a tree directly depends on the sum of the number of steps (n_i) for each character i of the total C (Tuffley and Steel, 1997):

$$L = \prod_{i=1}^{C} (1/s)^{n_i+1}$$

The exponent adds 1 to n_i to take into account the prior probability of the root state. Applying logarithms to both sides of the equation,

$$\ln L = \ln(1/s) \left(C + \sum_{i=1}^{C} n_i \right)$$

the resulting equation makes it evident that the tree score depends exclusively on the sum of equally weighted steps ($\sum n_i$). In other words, the difference in log-likelihoods for a given step difference d between two characters is the same, regardless of whether it is a difference between 1 and 1 + d, or between 20 and 20 + d steps. Thus, when the size of the state space is constant over all characters, NCM does not mimic IWP, but instead EWP (as intended by Tuffley and Steel, 1997). This identity between NCM and EWP breaks down when the state space varies among characters.

Steps and homoplasy

Using a common state space for all characters makes their relative prior weights more even. However, even if

[2] Assumed numbers of states, that is. The number of actually observed states has some effect on the times needed for parsimony calculation (e.g. multicharacter algorithms for optimization can pack fewer characters together for more observed states), although still much less than the s^2 needed in the case of likelihood.

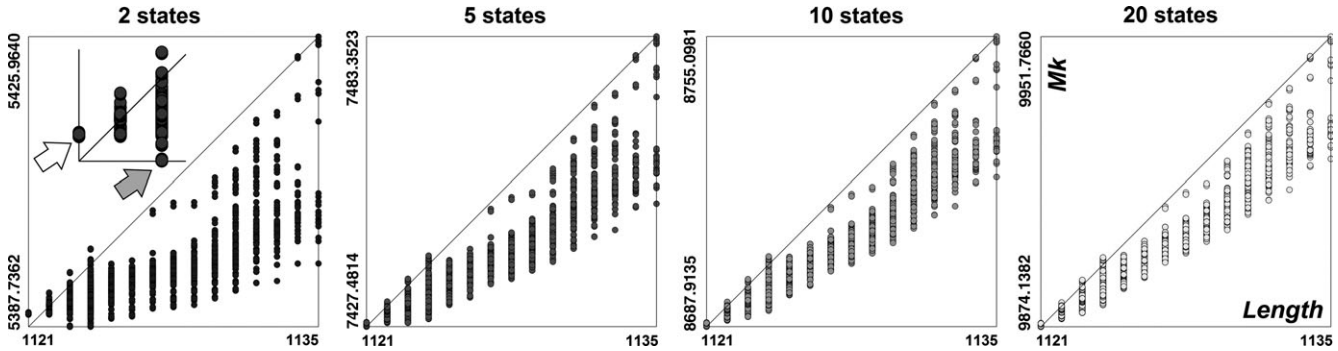

Fig. 4. Effect of increasing the size of the character-state space, for the same dataset and trees as in Fig. 1, on the correlation between the scores under the Mk model (*y*-axis, log likelihoods) and the scores under equal weights parsimony (*x*-axis). The arrows in the leftmost diagram indicate trees that are optimal under the Mk model (grey) or under parsimony (white); the same trees become optimal under both criteria as the size of the character-state space increases.

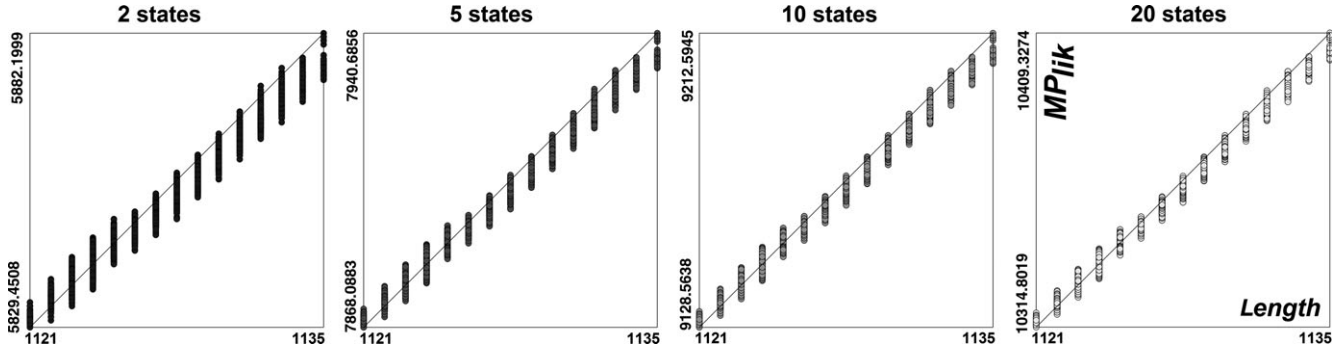

Fig. 5. Effect of increasing the size of the character-state space, for the same dataset and trees as in Fig. 1, on the correlation between the scores of the MP_{lik} approximation (*y*-axis) and the parsimony score under equal weights (*x*-axis). Values on the *y*-axis are log likelihoods.

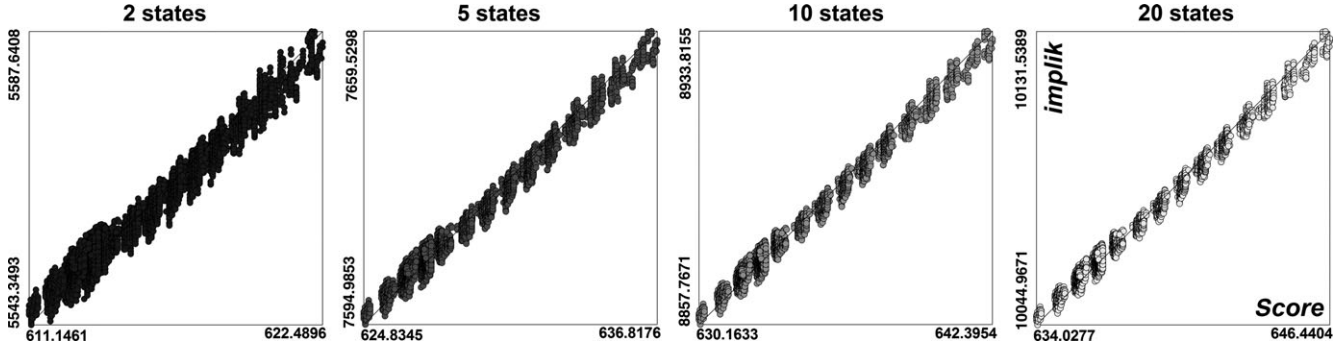

Fig. 6. Effect of increasing the size of the character-state space, for the same dataset and trees as in Fig. 1, on the correlation between the scores of the *implik* approximation (*y*-axis) and the score under implied weights (*x*-axis, user-defined relative weights as in *simplik* for the corresponding state space). Values on the *y*-axis are log likelihoods.

the size of the state space is fixed at a relatively large number, some differences remain between IWP and its likelihood analogues *simplik* and *implik*. This is because the differences in log-likelihoods, in the case of *simplik* and *implik*, decrease with the number of *steps*, regardless of whether those steps are *homoplastic* or not. Implied weighting, instead, was designed to down-weight characters only as their homoplasy increases, not their number of steps, so as to avoid penalizing multistate characters for no reason other than displaying more variability (see Goloboff, 1993). An example of this difference is shown in Fig. 7, a dataset where most of the tree structure is determined by the black characters (3–6), but with the first two characters in the matrix in conflict with each other, determining two alternative positions for taxon *e*. The first character is

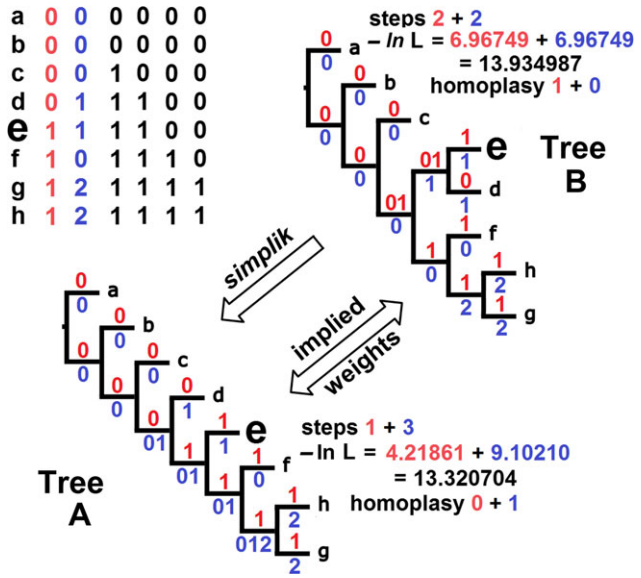

Fig. 7. A case where differences in the number of states in two characters results in a difference in prior weights under the likelihood approximation to implied weights. The first and second characters in the matrix are in conflict and have different numbers of states. Even with an assumed common state space of 3 (the largest number of states in any character), the tree that saves steps in the character with fewer states (tree A) is preferred under *simplik*, because the approach does not distinguish between steps (required for observed states) and homoplastic steps (required for origination of similar features). Both trees, A and B, are considered identical under implied weights, because each requires one step of homoplasy in one of the characters.

binary, and can have 1 step on tree A (where *e* is the sister group of *fgh*) or 2 steps on tree B (where *e* is sister to *d*). The second character has 3 states, with 3 steps on tree A, or 2 steps on tree B. Because the difference in log-likelihood between 1 and 2 steps (for the first, binary character) is larger than the difference between 2 and 3 (for the second, multistate character), the tree that saves steps in the binary character is preferred by *simplik*, at the expense of postulating more steps for the multistate character. However, each of the two characters can have either 0 or 1 steps of homoplasy on the alternative trees, so the two trees are viewed as exactly the same under implied weighting. At least from the perspective of morphological analysis, the fact that the second character has taxa *gh* with a third state does not seem to count against its reliability; what should count against the reliability of a character is *unexplained similarity*, and all the similarities in the states for the second character can be accounted for by common ancestry in tree B. The logic of the Mk model, based on counting only transformations, does not seem to have a direct way to distinguish between steps and *homoplastic* steps (i.e. independent transformations into the same condition). Under the Poisson process assumed by the Mk model, homoplasy is simply a byproduct of the number of transformations, and so trees are evaluated under that model only on the basis of their numbers of steps, not on the basis of homoplasy. Given that difference, we consider the treatment under IWP to be preferable. From that perspective, each of the trees, A and B, has a single instance of a similarity not attributable to common ancestry (i.e. the similarity in state 1 of *d* and *e* in the second character for tree A, or the similarity in state 1 of *e* and *fgh* in the first character for tree B), and so both trees are considered equivalent under IWP.

More steps, higher likelihood?

Although the methods based on both optimal and integrated reconstructions will converge to the same results when large amounts of data are generated from a model tree with all branches of the same length, they need not produce the same result for every input. The methods based on summing up reconstructions may produce results different from EWP or IWP in some cases. Given that the likelihood from all reconstructions is considered, this may produce differences in the likelihood for trees that (from the perspective of parsimony) would seem to be identical. This can happen even in the case of trees with a single transformation. An example is presented in Fig. 8, which shows that depending on the location of a unique transformation (for a tree with 10 taxa, with a state space of 2, in black), the log-likelihood for the character can vary by 0.06749 units (i.e. 4.42705–4.35956, with the change located on the branches marked A or B in Fig. 8). As the state space is increased to 20 (grey), the differences in log-likelihood become much smaller, 0.03500 (i.e. 9.3988–9.3638). Of course, in the standard Mk model, where different branches of the tree have different lengths, the cost of a transformation will also depend on where that transformation is located, but that is the intention of the model. The evaluation constraining all branch lengths to be identical shown in Fig. 8 is meant to mimic parsimony, so the differences depending on where the change is located are undesirable (and they certainly cannot be justified on the same grounds as in the standard Mk model). The differences, in this case, arise from the summing up of reconstructions, not from differences in branch lengths (as they do under the Mk model).

Summing up the likelihood of reconstructions can thus produce different likelihoods for trees of the same numbers of steps. This is why *implik*, despite converging to the results of implied weighting for large numbers of characters, can produce different results for particular datasets. The differences in likelihood between different locations for the same number of steps can even be large enough to make the worst likelihood for a given number of steps lower than the best

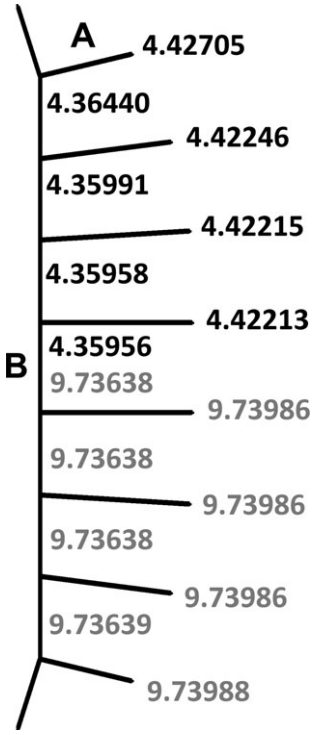

Fig. 8. Example showing that the score obtained for single-step (two-state) characters under a likelihood approach that uses the same length for all the branches of the tree may differ, depending on where the change is located. The black numbers indicate the log likelihoods for a state space of 2; the positions A and B indicate the branch with a change for the characters with the worst and best likelihoods, respectively. The grey numbers indicate the log likelihoods for a state space of 20.

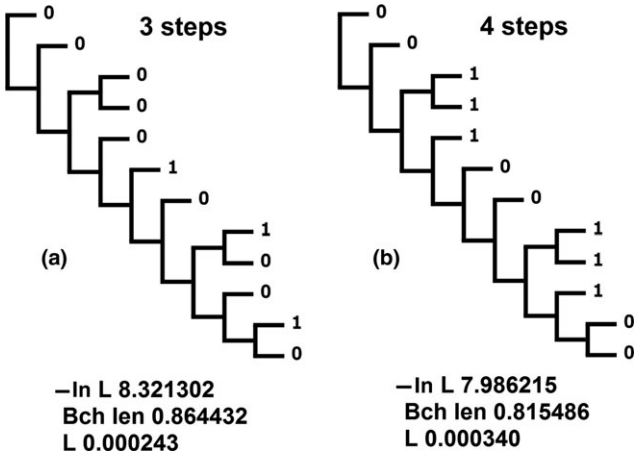

Fig. 9. (a) A character with a minimum of three changes has a worse individual likelihood (under *implik*) than a character with (b) a minimum of four changes. See text for additional discussion.

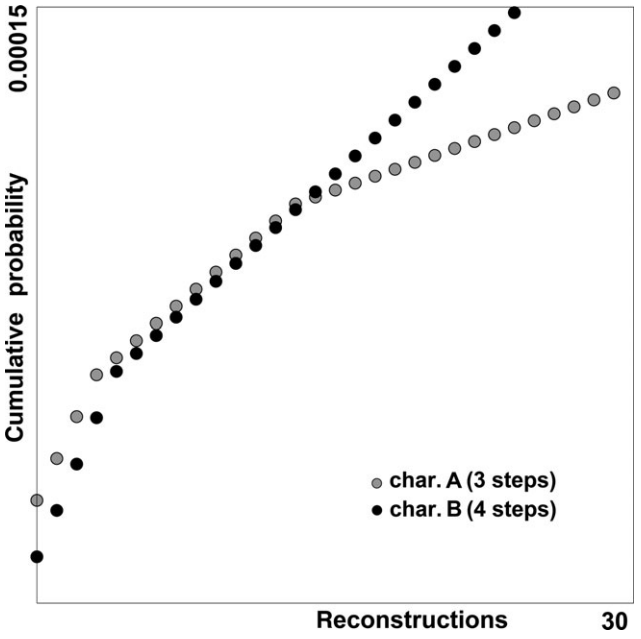

Fig. 10. Cumulative likelihoods for the characters in Fig. 9, for the best 30 reconstructions in each character. The best reconstruction of character A has a better likelihood than the best reconstruction for character B, but there are more reconstructions with a relatively high likelihood in character B, so that its overall likelihood is better than that of character A.

likelihood for a larger number of steps. In other words, a character with more homoplasy may nonetheless have a higher likelihood. An example is shown in Fig. 9. Character A has 3 steps, and an *implik* log-likelihood of −8.3213; character B cannot be reconstructed to have fewer than 4 steps, but it has an *implik* log-likelihood of −7.9862.

The main cause of the difference is that (even if the individual reconstruction of 3 steps in character A has a higher probability than any of the individual reconstructions of 4 or more steps in character B) there are more alternative reconstructions of 4, 5 or 6 steps for character B than for character A, which, summed together, contribute more to the likelihood. Figure 10 shows the cumulative probabilities for the first 30 reconstructions (ordered from highest to lowest individual likelihood). For character A (grey, with a minimum of 3 steps), the first reconstruction has a higher likelihood than the best reconstruction for character B (black, with a minimum of 4 steps), but as subsequent reconstructions are considered, the likelihood of character B matches and eventually exceeds that of character A. In other words, even when the reconstruction with fewest steps has fewer steps in character A than in character B, character B has more reconstructions with relatively few steps; there are more reconstructions with 9–12 steps in character A, and more reconstructions with 4–8 steps in character B. This distribution of numbers of steps among possible reconstructions is shown in Fig. 11a. The main contribution to the likelihood of character B then comes from the (many) alternative reconstructions with 4–6 steps, as

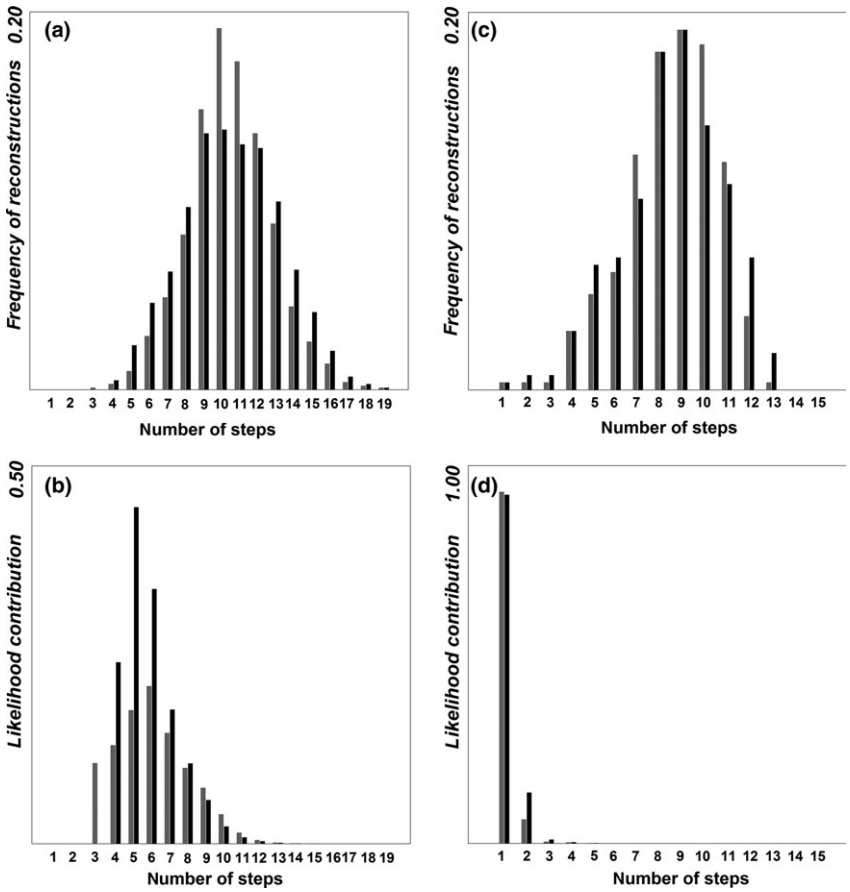

Fig. 11. (a, b) Histograms showing the relative frequencies of reconstructions with different numbers of steps (a), and the relative likelihood contributions of reconstructions with different numbers of steps (b) for the two characters in Fig. 9 (grey, for character A, and black, for character B, both drawn on a common scale). (c, d) The same as in (a) and (b), but for the characters indicated as A (grey) and B (black) in Fig. 8. See text for additional discussion.

shown in Fig. 11b; the likelihood of each individual reconstruction with 9–12 steps in character A is too low to substantially contribute to the overall likelihood of this character. The net result of this distribution of steps among possible reconstructions for each character is then that character B has a higher likelihood, even if the best possible reconstruction (the one with fewest steps and highest probability) is better for character A. These inversions will not occur when branches are very short (because in that case, as noted above, the majority of the likelihood contribution will be given by the most parsimonious reconstructions, as the probability of stasis is much larger than the probability of change). Therefore, these inversions can occur at relatively low numbers of steps for few taxa, but only at larger numbers of steps for larger numbers of taxa (and then, they can occur only in characters with significant amounts of homoplasy, thus making it less likely that those characters have a substantial influence on the final tree choice).

A similar consideration of the reconstructions with different numbers of steps helps explain why the single-step characters in Fig. 8 have different likelihoods. Figure 11c shows the step distribution of the two characters with lowest and highest likelihoods shown in Fig. 8 (indicated as "A" and "B", respectively). In the character changing at branch B, there are 2 reconstructions with 2 steps (instead of the single one for the character changing at branch A). The main difference in overall likelihood between these first two characters is caused by this difference (Fig. 11d).

A branch-length conundrum?

Consideration of the branch lengths shown in Fig. 9 indicates a strange behaviour of the calculation of likelihoods by summing reconstructions when all branches of the tree are constrained to have the same length. In Fig. 9, character A can be reconstructed to have only 3 steps, while character B cannot be reconstructed to have fewer than 4. One would expect from this that the branch lengths optimal for character A are shorter (i.e. with a lower probability of change) than those for character B. Yet the opposite is observed, which is

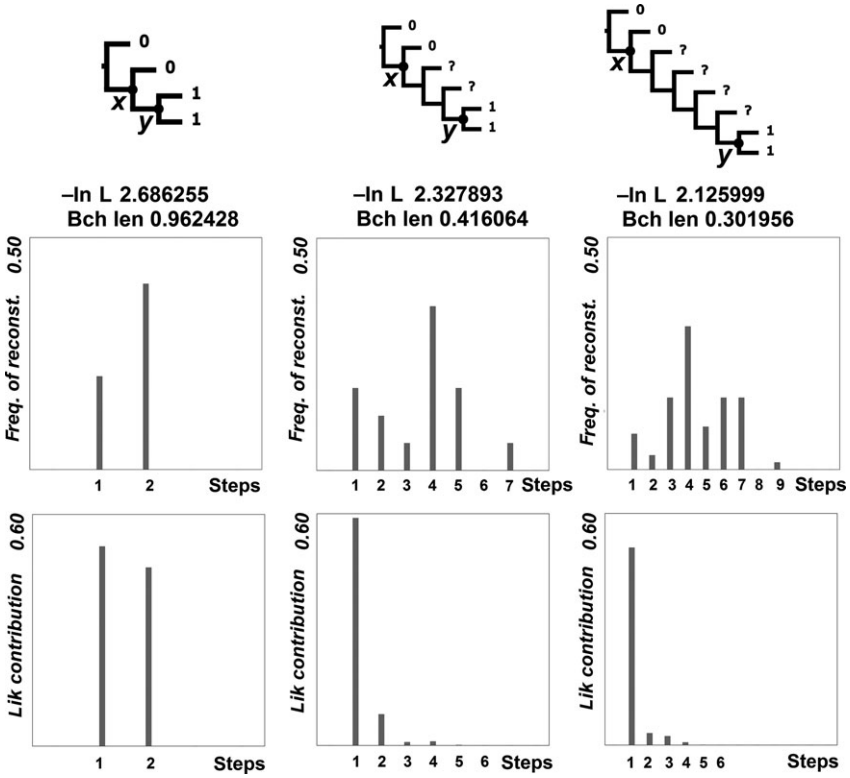

Fig. 12. Example showing how the insertion of taxa with missing entries between groups of taxa with different states can increase the likelihood under the *implik* or MP_{lik} approaches. The histograms show (for each case) the frequency of reconstructions with one or two steps (upper diagrams) and the relative likelihood contribution from reconstructions with different numbers of steps (lower diagrams).

counterintuitive. The reason for this difference is that, for character B, the reconstructions that contribute the bulk of the total likelihood are those with 4–7 steps (as shown in Fig. 11b, black bars). For character A, there are many more reconstructions with 9–12 steps than for character B, and even if each of these does not contribute much to the likelihood, they influence the final result because of outnumbering the reconstructions with fewer steps. Thus, the branch length for character A is adjusted to best fit the many reconstructions with 9–12 steps, while that for character B is adjusted to best fit reconstructions with 4–7 steps.

More taxa, higher likelihood?

With standard likelihood models, or with parsimony, adding taxa can never increase the fit of the data to a tree, even if the data have missing entries. An unusual behaviour of the likelihood calculation with fixed branch lengths and summation of all possible reconstructions is that adding taxa with missing entries can *increase* the fit. An example is shown in Fig. 12. The four-taxon tree to the left has a relatively low likelihood; the middle branch (from *x* to *y*) is quite long, but making it longer would decrease the probability of stasis along the branches leading to the terminals. When two taxa with missing entries are added in the middle of the tree, the length of each tree branch can be decreased, but there are now three branches (instead of a single one) between the nodes *xy*, so that even if the nominal branch length is less than before, the *xy* path is longer, thus making the probability of change along the *xy* path higher. Therefore, the middle tree has a higher likelihood than the left tree, made possible by the addition of the new taxa. This effect is even more pronounced on the tree to the right, with five branches along the *xy* path and the best likelihood of the three trees. Note that this effect of taxa with missing entries increasing the likelihood occurs only when the missing entries are located along branches that would (in a most parsimonious reconstruction) have character-state changes; placing the taxa with missing entries among several taxa with an identical observed state does not have that effect. The single reconstruction versions (sMP_{lik} and *simplik*) can also have changes in likelihood when taxa with missing entries are added to the tree, but the likelihood always becomes worse. The decrease in likelihood when adding taxa with missing entries is not identical to the behaviour of parsimony, but is not as unusual as the *increase* in likelihood with the addition of taxa observable in MP_{lik} and *implik*.

In the case of MP_{lik} and *implik*, one of the computational consequences of the possibility that the likelihood increases with the addition of taxa is that it becomes impossible to calculate trees by implicit enumeration, as that approach is based on discarding large proportions of the possible complete solutions using the likelihoods of incomplete solutions. Should one be interested in finding the optimal trees under any of those two criteria, for datasets containing missing entries, searches would have to be based either on algorithms that exhaustively enumerate all trees, or on heuristics.

In the standard Mk model, the same effect of taxon addition can be achieved, but only for the likelihood of individual characters, not for the entire dataset. For example, if other characters without missing entries in the added taxa frequently change along the *xy* path in the rightmost tree (thus making the path between those two nodes longer), then the likelihood for the character with missing entries may be higher relative to the leftmost tree; but in this case the overall likelihood for all characters will decrease. Predicting the influence of missing entries in likelihood approaches which sum up all possible reconstructions is, in general, quite difficult. In many cases, adding missing entries to a dataset produces unexpected behaviour for standard likelihood models. Goloboff and Wilkinson (2018: fig. 2) showed an example where JC69 applied to a set of perfectly compatible characters could produce trees that require some homoplasy, as a side result of differences in the branch lengths of subtrees that are themselves compatible. This type of effect has been observed only in cases of missing entries. Lemmon et al. (2009), Simmons (2012, 2014), and Simmons and Goloboff (2013) have showed other cases where missing entries seem to produce unusual or undesirable behaviour in standard likelihood.

Implementation

As our interest in the various alternative approaches to evaluate likelihood is only exploratory, a simple implementation has been done using TNT scripts. The scripting language of TNT allows us to calculate the log-likelihood values for any given (binary) tree, with expressions that are the namesakes of each of these methods. This approach allows handling multiple trees or, if desired, carrying out rudimentary searches (e.g. by using TNT options that create SPR or TBR loops; see Goloboff et al., 2008b: 784). Given that the implementation of the Mk model in PAUP* only handles characters with a common state space, and the state space in that program cannot be defined as an arbitrary number, the Mk approach (without estimating the proportion of invariants, see Lewis, 2001: 917–918, under a single rate) was also implemented in TNT (with the scripting expression *mklik*). To explore the effect of the size of the state space, in our implementation the size of the state space can be defined as any arbitrary number (with $2 \leq s \leq 64$, with the *lset* command), regardless of the number of states actually present in the matrix (clearly, s must be greater than or equal to the actual largest state in the matrix). In the case of MP_{lik} and sMP_{lik}, the state space can be set by the user, but it is always the same for all characters; in the case of *implik*, *simplik* or the Mk model, it is optionally possible to use the number of observed states as the state space for each character. The results of our Mk implementation were verified to be similar to those in PAUP* (version 4.0a build 164, 1 November 2018, with the options *condvar = no rates = equal*, to make results comparable).

Comparing models

Having described approaches that produce results similar to those of EWP and IWP, the next step is using model-selection methods to determine whether some of those models are viable alternatives to the standard Mk or JC69. The models tested here are fully non-nested, and given that Holder et al. (2010) used the AIC to compare models, the same comparison (appropriate for non-nested models) is carried out here. The AIC (Akaike, 1973) is defined as

$$\mathrm{AIC} = 2\mathrm{NP} - 2 \ln L$$

where NP is the number of parameters in the model. Models with lower AIC values are preferred; the strength S of preference for the model with minimum AIC ($\mathrm{AIC_{min}}$), relative to model x with AIC_x, is given (e.g. Burnham and Anderson, 2002: 74) by

$$S = e^{(\mathrm{AIC_{min}} - \mathrm{AIC_x})/2}$$

For the present comparisons, we considered the model with the lowest AIC to be strongly supported over the competitor model when $S < 0.15$ (i.e. the losing model will minimize information loss with a probability of 0.15 or less relative to the winning model). However, given the large values in AIC for most of the cases examined (average over 10 000), the differences in AIC are usually much larger than required to cross the threshold of 0.15 (1.897), so that using different thresholds produces no appreciable change in overall results.

Table 2 summarizes all the AIC comparisons done; the raw results (together with the datasets and scripts) are included in the Supplementary Material.

Table 2
Results of applying AIC tests to different datasets. In the case of simulated datasets, a result matching the expectation for the corresponding data-generation model is in bold type. In the case of molecular datasets, the results for XIW outperforming Mk in parentheses correspond to division into three blocks (four cases for the 1st, 2nd and 3rd position) and division into 25 groups of contiguous positions (five cases); no case is shared for these two options

Type of dataset	Taxa	Characters	Number of cases	*Implik* better than Mk	*Implik* better than MP_{lik}	MP_{lik} better than Mk	XIW better than Mk
Simulated; all branch lengths equal for all characters	60	50–150	100	62	0	**100**	–
Simulated; all tree branches the same length, varying for characters	60	50–150	100	**100**	**100**	23	–
Simulated; branch length common to all characters, varying over tree	60	50–150	100	**2**	0	**3**	–
Morphological	35–170	22–1844	182	36	86	21	–
Molecular	60–208	305–2218	37	0	32	0	9 (4+5)

Simulated datasets

To test whether the comparisons behave as expected, we initially tested three different types of simulated dataset (with 100 datasets of each type, all with 60 taxa). For these datasets, the number of characters was a random number between 50 and 150, randomly choosing between two and four as the maximum number of states in the matrix. Although the model tree (a different random tree for each dataset) is included in all datasets, the tree used for testing the fit of the different models was simply found with a quick heuristic search under EWP (with the *xmult* command with default parameters, just as we do for the empirical datasets). The first type of dataset used a model tree where all branch lengths were identical (a random number between 0.10 and 0.50), and all characters had the same rates. The model expected to best fit the data generated under such conditions is EWP, as both Mk and IWP would fit a superfluous number of additional parameters. As expected, the AIC chose MP_{lik}/EWP over both *implik*/IWP and the Mk model, in all 100 cases. The second type of dataset was generated with the exponential function of Goloboff et al. (2017) and character-state changes equiprobably allocated to any tree branch (we used a λ randomly chosen in the range 0.05–0.25). Given that different characters have different amounts of homoplasy, and that within every character changes are equiprobably located over all tree branches, the expected best method is *implik*/IWP. The AIC in this case chose *implik*/IWP over both MP_{lik}/EWP and the Mk model in all 100 cases; note that the Mk model used gamma (with four discrete categories) to estimate rate heterogeneity, but this clearly did not function as well as the analogue of implied weights, *implik*, for these simulated datasets. The third and last type of simulated dataset used branch lengths b common for all characters, but different for the different branches of the tree (determined as $b = x^2/3000$, where x is a random integer in the range 1–20, so that branch lengths vary between 3.33×10^{-3} and 0.133, as in Goloboff et al., 2018b). In this case, the Mk model was (as expected) preferred over *implik*/IWP or MP_{lik}/EWP in the vast majority of cases (98 and 92 cases, respectively). These comparisons suggest indeed that the methods approximating EWP and IWP perform as expected, and that the parameterizations considered here for each method are appropriate.

Morphological datasets

Having shown that the alternative likelihood approaches behave as expected for simulated data, we now apply the AIC test to empirical morphological datasets. A set of 182 datasets was used, combining

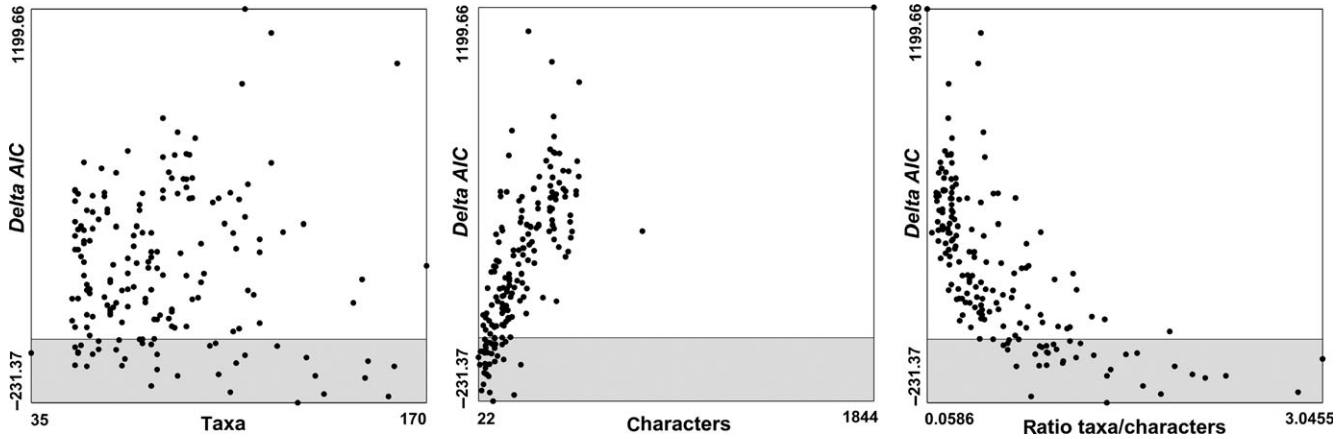

Fig. 13. Plots of differences in AIC values (*y*-axis) as a function of the number of taxa, characters, and the ratio between taxa and characters, on 182 empirical morphological datasets. Differences were calculated as $AIC_{Mk} - AIC_{implik}$. Negative values (indicated in the grey region of the diagram) thus correspond to cases where the likelihood approximation of implied weighting is preferred over the Mk model.

the datasets with 50 or more taxa from Goloboff et al. (2018b) with the datasets from Goloboff et al. (2017). Given that the chances of characters having significantly different amounts of homoplasy are much lower in datasets with low numbers of taxa, and that IWP for empirical datasets with up to 15–20 taxa normally produces the same trees produced by EWP (or at most, a subset), we decided to study only datasets with larger numbers of taxa. As some of these were originally matrices combining molecules and morphology, the morphology part had only missing entries for some of the taxa, and these were subsequently removed (this left only six datasets with fewer than 50 taxa). All the characters were treated as non-additive. The likelihoods were measured on the tree found by a round of *xmult* under EWP with default settings (this is intended to favour neither the Mk nor the *implik* models). The AIC was calculated (with PAUP*) for the Mk model with gamma rate heterogeneity (with the default four discrete categories), without estimating invariant characters (*lscore condvar = no genfreq = equal rates = gamma shape = estimate*).

As shown in Table 2, *implik*/IWP was the preferred model in about 20% of the cases (see Supplementary Material for the complete results of the tests). Holder et al. (2010) demonstrated that the NCM approximation to EWP would never be chosen by the AIC, but the situation is clearly different with other approximations to weighted parsimony. The discussion above (see *Numbers of parameters*) showed that (other things being equal), *implik* will tend to be preferred over the Mk model more commonly for cases with more taxa and fewer characters. Implicit in the discussion of Goloboff et al. (2008a: 760) is the idea that larger numbers of taxa will enable more accurate evaluation of the relative amounts of homoplasy for each character, which would make the results of IWP better justified for larger numbers of taxa. Plotting (Fig. 13) the difference in AIC as a function of the number of taxa, characters, or the ratio between taxa and characters shows that *implik* indeed tends to be preferred more often when the number of taxa is larger relative to the number of characters. The influence of the number of taxa is less clear, although there seems to be a weak trend for *implik* to be selected (or at least, to be closer to the fit of the Mk model) more often for larger numbers of taxa.

Molecular datasets

In our comparisons, we used the set of molecular datasets from Goloboff et al. (2018b), excluding the Zilla dataset (due to its size), leaving a total of 37 molecular datasets. The comparisons used the JC69 model, which is the equivalent of the Mk model. Empirical applications of ML normally use more complex models of character state transformations (such as K2P or GTR). Those would be equivalent to Sankoff parsimony. TNT implements both Sankoff parsimony and an implied weights option for weighting individual character state transformations (Goloboff, 1997). However, we currently have no implementation for a likelihood approximation for either of these approaches. Thus, the appropriate comparison in this case is the simpler JC69 model. A gamma distribution with four rate categories was used for the JC69 calculations.

The Supplementary Material contains the complete results for the tests; a brief summary is presented here. Almost since its inception, the idea of IWP was proposed to be more relevant in the case of morphology than in the case of molecular sequences (e.g. Goloboff, 1997: 225; Goloboff et al., 2008a: 769). The comparisons based on AIC prefer IWP over EWP in 84% of the cases, but never prefer IWP over JC69. This may

in part be due to molecular datasets typically having many more characters than morphological datasets (cf. Fig. 13). Goloboff (2013) proposed some extended weighting methods (XIW) that could be used for sequence data, such as collectively weighting blocks of characters, instead of the original approach to implied weighting. In our comparisons, we tried two alternative approaches: (i) three groups of characters within each dataset, with groups created by choosing every third character, starting from the first, second and third character in the matrix (codon division), and (ii) 25 groups of contiguous positions (contiguity division), with the size of each group varying depending on the total number of characters in the dataset. For the groups created with codon divisions, XIW was selected over JC69 for four datasets. The pool of datasets includes some rRNA (for the spider datasets, see Supplementary Material), where the division in codon positions has no biological meaning and is not expected to improve the estimation (and did not), but those cases were tested for comparability with the other datasets. For the 25 groups created on the basis of contiguity, five cases (including two of the rRNA datasets) selected XIW over JC69. Although the gamma distribution generally improved the likelihood of JC69 significantly, the likelihood produced by the approximation to XIW (even if constrained to have a single branch length within each block) was high enough to produce AIC values selecting XIW (with one or the other division in groups) over JC69+gamma in 24% of the cases, or nine of 37 datasets. Of course, our comparison is an oversimplification, considering only substitutions; the analysis of sequence data needs to consider other variables, such as indels, or chromosome rearrangements. Such repertoire of transformation is implemented under both parsimony and likelihood criteria (Wheeler et al., 2014), but it is not clear to us whether the models assumed are specified with enough precision to make the likelihood values truly comparable. Even if they were comparable, the decision of whether an approach considering those alternative types of transformations is better or worse than an approach considering only substitutions cannot, obviously, rest only on comparisons of AIC values or other methods for model selection. Regardless, it is clear that when it comes to considering only substitutions and prealigned sequences (i.e. the most widely used type of data), likelihood approximations to weighted parsimony can produce models that are as effective as standard likelihood methods.

Conclusions

The approaches described here, based on fixing all branches of the tree to have the same length, are not presented with the aim of replacing any existing method. Instead they are of interest (just like the NCM of Tuffley and Steel, 1997) as exploratory tools, to help better understand the behaviour and properties of parsimony and likelihood methods. Despite some unusual behaviour in specific cases (which prevents a recommendation for general applied use), the correspondence between the methods is close enough to warrant considering EWP and IWP as approximations to likelihood methods with constant branch lengths. If the state space that can be assumed under likelihood is large (which is reasonable in the case of morphology), the correspondence becomes even closer.

The uniformity of branch lengths used in MP_{lik} and *implik* is not defended here on the grounds that evolution must be extremely uniform, but instead on the interrelated premises that (i) character-state changes in a given character should a priori be considered as possibly occurring with the same probability on any of the branches of the tree, and (ii) the probability of change in a given character along a branch of the tree does not depend on whether other characters change along that branch. The uniformity of branch lengths, more than describing a very simple and constrained process, is intended as the best a priori expectation for a process so complex that no effective prediction can be easily made as to which branches of the tree are more likely to have changes in any given character. This "simplicity of complex systems" (more properly, simplicity of description and analysis of *some* of the aspects of the system) is also observed in other fields (e.g. statistical physics).

The uniformity of branch lengths has some implications that can perhaps be viewed as disadvantages: the approach assumes a perspective based on splitting points (i.e. speciation or cladogenesis), where the number of expected changes depends on the branching points between two nodes (Goloboff, 2003: 99), instead of the gradualistic perspective embodied in standard ML models (where time is one of the two main factors behind the expected number of changes). That split-based perspective also implies that dating nodes becomes more difficult and imprecise: accurate dating can only be accomplished in a gradualistic model where change is at least partly proportional to time. The belief that the Mk model allows more precise dating is one of the reasons why it has become so popular for the analysis of morphological data (Wright, 2017). However, there is a big difference in expecting morphological characters to evolve as assumed by the Mk and standard ML models, and morphological characters actually having done so. The recent empirical analysis of Goloboff et al. (2018b) strongly indicates that morphological datasets are very far from following the rules assumed by the Mk model. Some dating programs (e.g. Pathd8; Britton et al., 2007) use input trees where

amounts of evolution between nodes as well as calibration points are specified by the user; the amounts of evolution may well be based on most parsimonious reconstructions, but this still adds another layer of assumptions over parsimony itself. Or, alternatively, a dating with minimum ages can be accomplished by placing fossils of known age in the cladogram (e.g. Sterli et al., 2013), which properly acknowledges that the fossils only place a lower bound on dates.

The two premises used here to justify the uniform branch lengths (change for different characters uncorrelated, change located equiprobably on any tree branch) are also shared by the "episodic" model presented by Goloboff et al. (2018b), a model that seems to better fit the characteristics of many morphological datasets. The episodic model uses those two premises, but only within certain regions of the tree, with the characters being invariable in the rest of the tree. Goloboff et al. (2018b) suggested that (with certain restrictions) the accuracy of parsimony will be comparable for datasets generated under the episodic model, or under a model where characters are free to vary over the entire tree. Goloboff et al. (2018b) presented the episodic model only as a means to generate datasets; no likelihood implementation of the episodic model, or any reasonable approximation, has been proposed (the covarion model of Fitch and Markowitz 1970 comes closest, but does not incorporate either of the premises mentioned above). Thus, our comparison of AIC values includes all the alternatives that can be effectively tested at this time.

The similarity in results for IWP and a method that selects optimal values of branch lengths, uniform across all the branches of the tree and different for the different characters, helps illuminate several aspects of IWP that seem to puzzle critics. For example, the idea of maximizing the sum of weights under a concave function of the homoplasy is one of the aspects of implied weighting that Congreve and Lamsdell (2016) found objectionable, but maximizing sums of weights directly results from the present approach. The idea that the cost of adding a step should be smaller as the number of steps increases can then be justified explicitly from a likelihood perspective, as is the choice of values for the concavity constant. We note that a likelihood approach leads to much weaker weighting functions (i.e. larger values of k) than normally used in empirical analyses (e.g. Goloboff et al., 2008a used k values up to 16, and Goloboff et al., 2017, up to 12). The weighting curve resulting from the likelihood approximation corresponds (for two-state characters, and datasets with realistic numbers of taxa) to the one resulting from a k value of roughly half the number of taxa. This does not mean that the likelihood approximation provides absolute grounds for selecting exact values of k. The degree to which the cost of an extra step decreases with number of steps also changes with numbers of states, and it depends on a specific model of equiprobability of transformations between all states, and a given relationship between the overall probabilities of change and branch lengths. In addition, other likelihood approaches (e.g. Felsenstein, 1981b) can be used to justify alternative ways to downweight characters with homoplasy. However, the present results do suggest that it may be advisable to explore weaker concavities than is usually done.

Finally, our findings demonstrate that bombastic claims on the superiority of standard likelihood methods over parsimony lack serious justification, and that much remains to be explored on the differences (and similarities) between alternative phylogenetic approaches. A clear example of a novel approach is Samson et al.'s (2018) analysis, showing that parsimony (with characters weighted either equally or differentially) produces better stratigraphic fit than Bayesian methods. While the importance of likelihood methods is undeniable, and the use and discussion of those methods has greatly contributed to clarify many aspects of phylogenetic inference, this hardly means that there is no longer a place for parsimony methods in phylogenetics. The issue of model choice in phylogenetics is a complex one, and a simple number such as the AIC statistic as used here cannot capture all the aspects of the problem. However, most proponents of model-based phylogenetics (e.g. Holder et al., 2010) agree that AIC is one of the most important tools for model choice. The main finding of this paper is that methods behaving essentially like weighted parsimony can be preferred over the Mk model, according to a criterion that (following Posada and Buckley, 2004) is widely espoused by likelihoodists themselves. Thus, declarations like O'Reilly et al.'s (2018), pretending to have shown that "parsimony is dead" (p. 631), have no basis other than the authors' own prejudice.

Acknowledgments

The research for this paper was facilitated by a grant (PUE 0070) from Consejo Nacional de Investigaciones Científicas y Técnicas to PAG. We thank S. Catalano, M. I. and N. P. Giannini, M. Mirande, M. Pittman, D. Pol, C. Szumik, and A. Torres-Galvis for comments and advice. Associate Editor Mark Simmons also provided help above and beyond the call of duty. Comments from two reviewers (W. Wheeler, and Anonymous) helped clarify the manuscript. The Supplementary Material for this paper (scripts, datasets and test results for all individual datasets) can be found at http://www.lillo.org.ar/phylogeny/published/Goloboff_Arias_2019_IMPLIK.zip.

References

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado. Academic Press, New York, NY, pp. 267–281.

Britton, T., Anderson, C., Jacquet, D., Lundqvist, S., Bremer, K., 2007. Estimating divergence times in large phylogenetic trees. Syst. Biol. 56, 741–752.

Burnham, K., Anderson, D., 2002. Model Selection and Multimodel Inference – A Practical Information-theoretic Approach. Springer Verlag, New York, NY.

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

De Laet, J., 2005. Parsimony and the problem of inapplicables in sequence data. In: Albert, V. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, UK, pp. 81–116.

Farris, J., 1983. The logical basis of phylogenetic analysis. In: Platnick, N., Funk, V. (Eds.), Advances in Cladistics II. Columbia University Press, New York, NY, pp. 7–36.

Farris, J., 1986. On the boundaries of phylogenetic systematics. Cladistics 2, 14–27.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

Felsenstein, J., 1981a. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., 1981b. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. J. Linn. Soc. 16, 183–196.

Felsenstein, J., 1987. [Statistical Analysis of Hominoid Molecular Evolution]: Comment. Stat. Sc. 2, 208–209.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Fitch W, Markowitz E., 1970. An improved method for determining codon variability in a gene and its applications to the rate of fixation of mutations in evolution. Biochem. Genet. 1970;4:579–593.

Goldman, N., 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst. Zool. 39, 345–361.

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

Goloboff, P., 1995. Parsimony and weighting: a reply to Turner and Zandee. Cladistics 10, 91–104.

Goloboff, P., 1997. Self-weighted optimization: character state reconstructions and tree searches under implied transformation costs. Cladistics 13, 225–245.

Goloboff, P., 2003. Parsimony, likelihood, and simplicity. Cladistics 19, 91–103.

Goloboff, P., 2013. Extended implied weighting. Cladistics 30, 260–272.

Goloboff, P., Catalano, S., 2016. TNT version 1.5, including a full implementation of geometric morphometrics. Cladistics 32, 221–238.

Goloboff, P., Wilkinson, M., 2018. On defining a unique phylogenetic tree with homoplastic characters. Mol. Phylogenet. Evol. 122, 95–101.

Goloboff, P., Carpenter, J., Arias, J., Miranda-Esquivel, D., 2008a. Weighting against homoplasy improves phylogenetic analysis of morphological data sets. Cladistics 24, 758–773.

Goloboff, P., Farris, J., Nixon, K., 2008b. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

Goloboff, P., Torres, A., Arias, J., 2017. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. Cladistics 34, 407–437.

Goloboff, P., Torres, A., Arias, J., 2018a. Parsimony and model-based phylogenetic methods for morphological data: a response to O'Reilly et al. (2017). Palaeontology, 61, 625–630.

Goloboff, P., Pittman, M., Pol, D., Xu, X. 2018b. Morphological datasets fit a Common Mechanism much more poorly than DNA sequences and call into question the Mkv model. Syst. Biol.(in press). https://doi.org/10.1093/sysbio/syy077.

Holder, M., Lewis, P., Swofford, D., 2010. The Akaike Information Criterion will not choose the no common mechanism model. Syst. Biol. 59, 477–485.

Huelsenbeck, J., Ané, C., Larget, B., Ronquist, F., 2008. A Bayesian perspective on a non-parsimonious parsimony model. Syst. Biol. 57, 406–419.

Huelsenbeck, J., Alfaro, M., Suchard, M., 2011. Biologically inspired phylogenetic models strongly outperform the no common mechanism model. Syst. Biol. 60, 225–232.

Jukes, T., Cantor, C. 1969. Evolution of protein molecules. In Munro, N. (Ed.), Mammalian Protein Metabolism. Vol. 3. Academic Press, New York, NY, pp. 21–132.

Kim, J., 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45, 363–374.

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58, 130–145.

Lewis, P., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50, 913–925.

Neyman, J., 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta, S., Yackel, J. (Eds.), Statistical Decision Theory and Related Topics. Academic Press, New York, NY, pp. 1–17.

Nylander, J., Ronquist, F., Huelsenbeck, J., Nieves-Aldrey, J., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47–67.

O'Reilly, J., Puttick, M., Parry, L., Tanner, A., Tarver, J., Fleming, J., Pisani, D., Donoghue, P., 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. Biol. Lett. 12, 1–5.

O'Reilly, J., Puttick, M., Pisani, D., Donoghue, P., 2018. Empirical realism of simulated data is more important than the model used to generate it: a reply to Goloboff et al. Palaeontology 61, 631–635.

Posada, D., Buckley, T., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793–808.

Puttick, M., O'Reilly, J., Tanner, A., Fleming, J., Clark, J., Holloway, L., Lozano Fernandez, J., Parry, L., Tarver, J., Pisani, D., Donoghue, P., 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. Proc. R. Soc. B 284, 20162290.

Puttick, M., O'Reilly, J., Pisani, D., Donoghue, P., 2018. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. Palaeontology 62, 1–17.

Ronquist, F., Rasnitsyn, A., Roy, A., Eriksson, K., Lindgren, M., 1999. Phylogeny of the Hymenoptera: a cladistic reanalysis of Rasnitsyn's (1988) data. Zool. Scr. 28, 13–50.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.

Samson, R., Choate, P., Keating, J., Randle, E., 2018. Parsimony, not Bayesian analysis, recovers more stratigraphically congruent phylogenetic trees. Biol. Lett. 14, 20180263.

Sankoff, D., Rousseau, P., 1975. Locating the vertices of a Steiner tree in an arbitrary space. Math. Program. 9, 240–246.

Simmons, M., 2012. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Mol. Phylogenet. Evol. 62, 472–484.

Simmons, M., 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. Mol. Phylogenet. Evol. 80, 267–280.

Simmons, M., Goloboff, P., 2013. An artifact caused by undersampling optimal trees in supermatrix analyses of locally sampled characters. Mol. Phylogenet. Evol. 69, 265–275.
Smith, M., 2019. Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. Biol. Lett. 15, 20180632.
Steel, M., 1989. *Distributions on bicoloured evolutionary trees*. PhD Thesis. Massey University.
Steel, M., 2002. Some statistical aspects of the maximum parsimony method. In: De Salle, R., Giribet, G., Wheeler, W. (Eds.), Molecular Systematics and Evolution: Theory and Practice. Birkhäuser Verlag, Basel, Switzerland, pp. 125–139.
Steel, M., 2005. Should phylogenetic models be trying to "fit an elephant"? Trends Genet. 21, 307–309.
Steel, M., Penny, D., 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17, 839–850.
Steel, M., Penny, D., 2004. Two further links between MP and ML under the Poisson model. Appl. Math. Lett. 17, 785–790.
Sterli, J., Pol, D., Laurin, M., 2013. Incorporating phylogenetic uncertainty on phylogeny-based palaeontological dating and the timing of turtle diversification. Cladistics 29, 233–246.
Swofford, D. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.
Swofford, D., Olsen, G., Waddell, P., Hillis, D., 1996. Phylogenetic inference. In: Hillis, D., Moritz, C., Mable, B. (Eds.), Molecular Systematics, 2nd ed. Sinauer, Sunderland, MA, pp. 407–514.
Tuffley, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59, 581–607.
Wheeler, W., 2016. Systematics: A Course of Lectures. John Wiley and Sons, Oxford.
Wheeler, W., Lucaroni, N., Hong, L., Crowley, L., Varón, A., 2014. POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics 31, 189–196.
Wright, A., 2017. Editor's note on 'Putting fossils in trees' special issue. Biol. Lett. 13, 20170103.
Wright, A., Hillis, D., 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS ONE 9, e109210.
Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.
Yang, Z., 1996. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42, 294–307.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1 Supplementary materials.