Palaeontology
The Palaeontological Association
www.palass.org

[Palaeontology, Vol. 59, Part 3, 2016, pp. 447–462]

IMPLIED WEIGHTING AND ITS UTILITY IN PALAEONTOLOGICAL DATASETS: A STUDY USING MODELLED PHYLOGENETIC MATRICES

by CURTIS R. CONGREVE[1] *and* JAMES C. LAMSDELL[2,3]

[1]Department of Geosciences, Pennsylvania State University, State College, University Park, PA 16802, USA; e-mail: crcongreve@gmail.com
[2]Department of Geology and Geophysics, Yale University, PO Box 208109, New Haven, CT 06520, USA
[3]Division of Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA

Typescript received 20 October 2015; accepted in revised form 15 February 2016

Abstract: Implied weighting, a method for phylogenetic inference that actively seeks to downweight supposed homoplasy, has in recent years begun to be widely utilized in palaeontological datasets. Given the method's purported ability at handling widespread homoplasy/convergence, we investigate the effects of implied weighting on modelled phylogenetic data. We generated 100 character matrices consisting of 55 characters each using a Markov Chain morphology model of evolution based on a known phylogenetic tree. Rates of character evolution in these datasets were variable and generated by pulling from a gamma distribution for each character in the matrix. These matrices were then analysed under equal weighting and four settings of implied weights (k = 1, 3, 5, and 10). Our results show that implied weighting is inconsistent in its ability to retrieve a known phylogenetic tree. Equally weighted analyses are found to generally be more conservative, retrieving higher frequency of polytomies but being less likely to generate erroneous topologies. Implied weighting is found to generally resolve polytomies while also propagating errors, resulting in an increase in both correctly and incorrectly resolved nodes with a tendency towards higher rates of error compared to equal weighting. Our results suggest that equal weights may be a preferable method for parsimony analysis.

Key words: phylogenetic methods, cladistics, homoplasy, implied weights, parsimony, phylogeny.

THE advent of computationally based phylogenetic methods has had a dramatic effect on the manner in which we study evolution. Utilizing an optimization criterion such as parsimony, likelihood or posterior probability, biologists can take molecular and morphological characteristic data and generate phylogenetic hypotheses of relationship. Palaeontological studies have traditionally utilized maximum parsimony over parametric phylogenetics to reconstruct the evolutionary relationships of fossil taxa, potentially a legacy issue owing to the likelihood models for morphological character change being derived from modifications to models developed for nucleotide evolution (Lewis 2001; Wright & Hillis 2014). Parsimony itself, however, is not a model-less system (see Holder *et al.* 2010; Huelsenbeck *et al.* 2011; and references therein), and while the underlying basis of parsimony is to explain phylogenetic relationships while invoking the fewest number of character changes possible given the data, there has recently been a push towards a different form of optimization that seeks to maximize the total fit of characters on the tree through implied weighting (Goloboff 1993). The fit of a character is defined by the interplay of the number of changes the character undergoes, the minimum number of changes possible for the character, and a concavity constant that influences how severely homoplasious characters are downweighted. This results in a tree being selected based on the weights it assigns its characters given the topology, not the number of character changes itself. Although a number of palaeontological and neontological studies have experimented with implied weights (e.g. Mirande 2009; Cruz Mendes 2011; Cerdeño *et al.* 2012; Weiss *et al.* 2012), the method has largely failed to gain traction among biological workers since it has been considered to be largely non-parsimonious (Turner & Zandee 1995), only becoming actively used in other fields such as parsimony analyses of endemicity (Morrone 2014). The method is, however, getting a fair amount of exposure in palaeontological studies where it has been used to reduce the impact of homoplasy introduced by the inclusion of juvenile ontogenetic stages in analyses (e.g. Tschopp *et al.* 2015), to resolve the position of volatile taxa (e.g. Xu & Pol 2014), as a stress test to

doi: 10.1111/pala.12236

assess which parts of the tree are most impacted by homoplasy (e.g. Smith & Ortega-Hernandez 2014), or simply to increase resolution (e.g. Jones *et al.* 2014). Proponents of implied weighting suggest that the method is ideal for working on palaeontological data, especially data sets with high homoplasy or missing data. While the model has been demonstrated to increase internal consistency within real datasets (Goloboff *et al.* 2008*a*), this pattern is not in and of itself a true test of the efficacy of the methodology since the same could be said of any form of character weighting. Rather, a true test of the utility of this method is to compare how well implied weighting converges on a known tree. To this end, we have constructed a series of model datasets based on a single known tree, which serve to test the accuracy of implied weighting with a series of concavity constant (k) values. It is important that the veracity of this method be critically assessed, as the negative impact of phylogenetic inaccuracy not only influences systematics but cascades over into meta-analyses that rely on accurate phylogenetic frameworks to study broader patterns and processes.

HOMOLOGY AND THE PHILOSOPHICAL UNDERPINNINGS OF PARSIMONY

In order to evaluate the utility of implied weighting it is necessary to understand how it fits into parsimony theory and how parsimony has been applied to phylogenetic analyses. Core to the workings of parsimony analyses are the concepts of homology and homoplasy. The nature of homology is an oft-debated topic among phylogenetic systematists (see Wiley (2008) for a detailed overview). However, homology in its simplest form applies to characters that have shared evolutionary origins, therefore representing modifications to the same biological structure (the various permutations of the tetrapod forelimb form a classic example). We largely follow Wiley's (2008) reconciliation of Patterson's (1982) and Ghiselin's (2005) competing standpoints on the nature of homology; taxic homologues are represented in a dataset by the same state of a character, and act as synapomorphies for defining phylogenetic clades (homology *sensu* Patterson 1982), while transformational homologues are hypotheses that manifest as a nested set of phylogenetic statements, representing the shared evolutionary origins of the taxic homologues that share a transitive relationship (homology *sensu* Ghiselin 2005). These definitions are consistent with Van Valen's (1982) argumentation that homology is defined by information flow; homoplasy, therefore, is simply similarity achieved through independent evolution in different parts of the tree of life (Lankester 1870). The problem with this definition from a morphological standpoint, however, is that while homoplasy through convergence (similarity in form as mediated through interactions with the environment) is independent of information flow, homoplasy through parallelism (independent expression of a form mediated by the same genetic pathways) is not, and so might be considered homology under Van Valen's framework. However, Eldredge & Cracraft (1980) have shown that there is no distinction between convergence and parallelism in practical terms, and so we adopt the broader definition of homoplasy as any ahistorical characteristic that would act as a synapomorphy defining a polyphyletic group.

The sole criterion for maximum parsimony is to have the fewest number of character changes possible, resulting in phylogenetic trees that minimize homoplasy. While this basic principle underlies all parsimony analyses, there are a number of variations to this principle that can alter how the data are treated computationally, and these generally manifest as assumptions regarding character state ordering and character weighting. Character states in multistate (non-binary) characters can be treated as unordered (known as Fitch parsimony; Fitch 1970) in which any character state can change to any other character state at the cost of a single step. Alternatively multistate characters can be treated as ordered (known as Wagner parsimony; Farris 1970) whereby a character is constrained to pass through its states in the prescribed order, although the directionality of change is neutral. Therefore a change of state from 0 to 2 would cost two steps, as it has to pass through state 1 in order to do so. Wagner parsimony clearly adds an extra layer of assumption to the analysis, as we explicitly assume that we know the order of transformation of a homologue *a priori* and actively enforce that transformation series during the analysis. Character weighting, meanwhile, can be performed both *a priori* and *a posteriori* to the initial analysis, and there is some debate as to whether equal character weighting (often referred to as strict parsimony) assumes any less than differential character weighting (e.g. Chippindale & Weins 1994; Huelsenbeck *et al.* 1994; Vogt 2002).

The concept of character weighting has its origins among some of the earliest discussions of parsimony-based phylogenetics (Hennig 1966; Kluge & Farris 1969). *A priori* weighting includes character selection, whereby investigators select or exclude characters from the analysis based on perceived reliability (Hennig & Schlee 1978), often qualified in terms of apparent variation, structural complexity, and likelihood of homoplasticity. The more explicit forms of *a priori* weighting involve increasing the step-cost of transformations of certain characters relative to others based on their supposed importance or complexity; for example, the acquisition or loss of a limb may be assigned a higher cost than the acquisition or loss of

an individual digit, or the expression of a calcified endoskeleton may be assigned a higher cost than the expression of elongated phalanxes. As well as sharing the same underlying mechanisms as multistate character ordering (imposing a structure to the data through modifications to the step-cost of character transformations), *a priori* character weighting also requires a qualitative assessment of the importance of the characters and would seem to add an extra layer of assumptions to the analysis. While differentially weighting characters in this way may seem appealing, in practice this is far from simple. For example, it would seem logical that the loss of an entire set of limbs would be more important than variations in individual bone shape. However, limbless squamates are consistently grouped together in parsimony analyses of morphological data (Lee 2005; Gauthier *et al.* 2012), despite ample morphological evidence suggesting disparate origins of limb loss among a number of squamate groups (Greer 1991). In this case limb loss and general convergence on fossorial life habits results in parsimony retrieving a false grouping even under equal character weightings. Squamates therefore exhibit a predilection for limb reduction which has been shown to have a consistent underlying genetic cause linked to *Hox* gene expression (Sanger & Gibson-Brown 2004; Kohlsdorf *et al.* 2008). It is therefore possible for complex morphological structures to be drastically influenced by relatively simple changes in gene expression, and any *a priori* assumption of differential weighting appears to be at serious risk of imparting unwarranted bias to the analysis. While an assignment of equal weights is itself an assumption, it is at least neutral in terms of which specific characters are most 'important' to reconstructing the phylogeny, and in this way it can be considered more conservative than *a priori* differential weighting.

A posteriori methods provide alternative solutions to the issue of character weighting, and have their roots in the principal of reciprocal illumination (Hennig 1966), whereby each individual hypothesis is evaluated by the extent in which it agrees with the overall, favoured hypothesis given all available data. Whereas Hennig was of the opinion that homoplasy is a result of erroneous homology statements caused by convergence and advocated reformulating the actual homology statements of characters based on this process, modern *a posteriori* weighting methods instead seek to downweight the importance of homoplastic characters through either outright deletion (Wiley & Lieberman 2011) or reduction of their step-cost. This latter technique forms the basis of what might be the most widely used *a posteriori* weighting system, weighting through successive approximations (Farris 1969). This successive weighting procedure uses measures of character quality derived from an initial equally weighted analysis to evaluate character reliability. Each character is then assigned a weight based upon its consistency index (CI; a statistic that essentially quantifies how homoplasious the character is for a given tree topology) and the analysis is rerun. This procedure is then repeated a number of times until the weights become stable (the reweighting procedure results in no alteration to character weights). It is important to realise, however, that estimates of homoplasy can only be made in reference to an initial tree estimate, and because of this weighting through successive approximations has been argued to be circular (Cannatella & de Queiroz 1989; Swofford & Olsen 1990) or recursive (Carpenter *et al.* 1993; Carpenter 1994). Felsenstein (1981), however, likened it to a compatibility analysis, which attempts to circumvent problems of circularity by taking the average consistency index over all the most parsimonious trees rather than from a single topology. This approach is very similar methodologically to reciprocal illumination, and still results in somewhat arbitrary discrimination between equally parsimonious (and therefore equally optimal) topologies.

Implied weighting (Goloboff 1993) sets out to provide a non-iterative alternative to successive approximation. Both implied weighting and its extension, self-weighted optimization (Goloboff 1997) (which applies the same methods to assigning weights to character state transformations), estimate character weights concurrent with tree searching. The method utilizes a model (based on character fitness, f) to successively downweight characters that the model deems to be homoplasy against every possible generated topology, thereby honing in on the tree (or trees) that maximizes homology. The extent to which the model downweights characters deemed to be homoplasious depends upon the selected concavity constant, k. Low values of k ($k \sim 0$–2) strongly downweight homoplasy, whereas larger values (typically larger than 3) allow for some signal to come from homoplasy. The implied weighting method has been considered in conflict with the parsimony criterion (Turner & Zandee 1995; Kluge 1997*a*) and rejected on the grounds of the resulting weights being untestable and the resulting trees tautologous (Kluge 1997*b*). Goloboff (1995, 1997) countered these arguments by stating that parsimony applies to the shortest possible trees given appropriate weighting of the characters, and insists that the optimality criterion adopted by implied weighting is a refined way to measure parsimony in trees. This response sidesteps most of the philosophical objections to implied weighting, some of which are detailed below, while the performance of implied weighting compared to equal weighting has only been tested using actual datasets (Goloboff *et al.* 2008*a*), for which it is impossible to know the true phylogenetic topology. As such the method has been evaluated based on the resolution, support values and consistency of the

trees it produces, which are not in themselves correlates of the method's accuracy. Tests using simulated data, for which the tree topology is known, and discussion of the subsequent results, form the remainder of this paper.

PHILOSOPHICAL RAMIFICATIONS OF IMPLIED WEIGHTING

The basic justification of Goloboff's model is that we should expect data that has a high degree of homoplasy to vary more, and therefore have more character changes (or steps; Goloboff 1993). This is couched as being the most parsimonious interpretation (Goloboff 1995), but this model does not follow the core principles of parsimony since it is arguing that it is most parsimonious to assume a result will be less parsimonious under certain conditions. In essence, implied weighting is generating trees that are knowingly suboptimum (in regards to the parsimony optimization criteria), and instead maximizing 'homology' by minimizing 'homoplasy'. While the concept of minimizing homoplasy appears to be firmly in accord with the parsimony criterion, parsimony is tied to finding the simplest tree (the tree with the least number of steps), not the tree with the most internal agreement; in claiming that trees retrieved under implied weighting are the most parsimonious, the proponents of implied weighting essentially redefine parsimony based on the writings of Farris (1983), the relevant section of which is reproduced here in full:

> Now suppose that many independent characters support one placement of a taxon, while just one supports an alternative placement. Possible reinterpretations aside, if the characters are weighted equally, weight of evidence favors the first placement. If the data were different and the counts reversed, the second placement would then be favored. *If character weights were not all equal, either placement might be supported by the greater weight of evidence, depending on the character weights.* The process of selecting a placement would be the same whether weight of evidence were reflected by counts of equal weights or by sums of differing ones. In either case the decision is made by accepting the stronger body of evidence over the weaker, *and ad hoc hypotheses of homoplasy are required to the extent that evidence must be dismissed in order to defend the conclusion*. Farris (1983), p. 35 (emphasis ours).

While proponents of implied weights are correct that equal weighting is not more parsimonious than differential weighting, as under equal weights a weight is still assigned to each character (they just happen to all be equivalent), implied weighting appears to risk violating a key component of Farris' support of parsimony, namely avoiding uncorroborated suppositions (Farris 1983). This new optimization does not follow any standard model of evolution; it is not generating answers that are most likely in a probabilistic sense (as in likelihood) nor is it assuming evolution to be a rare event-based process (as in parsimony). Rather, it generates trees that are most internally consistent with their own data given successive weighting from the implied weighting model. Ultimately, arguments over whether implied weighting is parsimonious or not are moot; both equal and differential weighting seeks the shortest tree in terms of step-cost, and while we argue that differential weighting systems that produce trees that are logically inconsistent with trees produced under equal weighting systems are undesirable in that they are obviously less conservative in the way they treat the data, proponents of implied weighting simply occupy a different philosophical standpoint on the issue.

There is legitimate cause for concern, however, in whether or not implied weighting converges upon an accurate topology close to replicating the historical pattern of taxon relationships. Key to this issue is the question of how implied weighting identifies and handles homology, and whether homoplasy is always deleterious to retrieving an accurate topology. The results generated from the method strongly depend upon the value of k used in performing the analysis. Goloboff (1993, 1995) suggested that k values need to be tweaked for each dataset so that the model gives an appropriate response, implying that a value which is too high or too low will result in improper results. However, it should be argued that which characters are truly homologous and truly homoplasious are unknown to any researcher, so there is no way of actually determining how much of your data represent true homology statements. Therefore, it is effectively impossible to know which k value would be proper to use in any given dataset, especially since the true evolutionary tree is unknowable. At times, Goloboff (2014) seemed less concerned with whether the results are accurate, instead focusing on internal consistency; the resolution of an analysis and its internal consistency alone are not the same as the topology being the most accurate depiction of reality. Goloboff also claimed that simulations provide no test of accuracy, as simulations are incapable of incorporating all the hidden and unknown factors operating in the real world (Goloboff 2014). Such a statement clearly falls under the fallacy of *petitio principii*; simulations are not suitable for testing the implied weighting model because they do not represent the real world, as they are simulations. There is also the underlying assumption that implied weighting somehow better represents the hidden and unknown factors operating than does any simulation which could be constructed to

test it; these hidden and unknown factors are, however, exactly that: hidden and unknown. Once again the argument rests on implied weighting being preferred because it produces results consistent with itself. Other supporters of implied weighting simply state that implied weighting is the favoured weighting option from a philosophical standpoint, without any further support (Legg 2013; Legg & Vannier 2013; Jones *et al.* 2014), or again appeal to self-consistency (Legg *et al.* 2012; Legg & Caron 2014).

The veracity of arguments for implied weighting aside, it is agreed on all sides that results under implied weighting can be very dependent on the value of *k* used. This dependence is compounded by the way in which parsimony defines homology: as synapomorphy. While this makes sense from a computational perspective, the concept of homology as synapomorphy is intrinsically tied to Dollo's law, the supposition that characteristics cannot revert to previous states (Gould 1970). As demonstrated in the earlier squamate example, however, reversals do occur and the mechanisms for them are becoming increasingly understood (Teotónio & Rose 2001; Porter & Crandall 2003; Collin & Miglietta 2008). The problem with reversals is twofold. Monophyletic groups can be defined in part by reversals as demonstrated by desmognathine salamanders, which revert to an aquatic larval phase from direct-developer plethodontine salamanders (Chippindale *et al.* 2004), and in the non-filtratory morphology of the anterior trunk limbs of notostracan crustaceans (Oleson 2007), which are nested within the filtratory branchiopods. Furthermore, some groups for which monophyly is well established, such as branchiopods (Richter *et al.* 2007; Koenemann *et al.* 2010), are not defined morphologically by a single synapomorphy that does not undergo subsequent reversal in at least some species (Oleson 2007). In cases such as these, down-weighting homoplasy will strip clades of their support and risks destabilizing the entire tree. Källersjö *et al.* (1999) showed that homoplasy increases phylogenetic structure in analyses of molecular data; rapidly evolving nucleotide sites still provide support for clades even though they exhibit more overall homoplasy across the entire tree. Goloboff (2014) recently implemented a partitioning system that allows for different character sets to be analysed under different weighting schemes, yet still insists that homoplasy in morphological matrices behaves differently and only has an adverse influence when reconstructing tree topology (Goloboff *et al.* 2008*a*). However, we can see no logical reason why morphological homoplasy should be more deleterious to an analysis than molecular homoplasy.

One major flaw of implied weighting is that it assumes homoplastic characters have an equal likelihood of homoplasy across the entire tree. While preliminary studies have shown that homoplasy tends not to be clustered on phylogenetic trees (Sanderson 1991), there exist a number of known trends that would suggest homoplasy is not a truly random phenomenon and that localized homoplasy may occur. These trends include: parallelism (Simpson 1950), by which closely related taxa are more likely to exhibit homoplasy in characteristics with shared developmental origins; the fact that some lineages exhibit more of a predilection for paedomorphic change than others, paedomorphosis being one of the primary mechanisms by which character reversals are known to occur (Kluge 1988); and the fact that clades exhibit higher rates of evolution and reach maximal disparity early in their evolution (Ruta *et al.* 2006; Hughes *et al.* 2013), a higher rate of evolution passively increasing the chance of homoplasy. While uneven occurrences of homoplasy are a problem for any weighting system, localized homoplasy can result in characteristics that nonetheless otherwise define a clade being punitively downweighted under implied weights, as the method does not differentiate between localized homoplasy and general random occurrence. Furthermore, directed homoplasy (homoplasy due to adaptive convergence) can exhibit strong phylogenetic signal, and there is always the risk that, by maximizing character fit, implied weighting could converge on an erroneous topology through maximizing homoplasy.

This concern is, in our opinion, the most crucial. No study has been performed to see whether implied weighting is actually consistently minimizing homoplasy or maximizing fit in an *ad hoc* manner to the extent of retrieving a well-resolved but erroneous topology. In order to test this, however, it is impossible to use real datasets because the true tree is unknown and it is inappropriate to use branch support as a measure of accuracy as through maximizing character fit it could be possible to retrieve strong support for erroneous relationships. We must therefore simulate data for a given tree topology and compare the performance of equal weighting and implied weighting in retrieving the correct topology. Even through evaluating the efficacy of equal weighting versus implied weighting in this manner, the problem remains of how to determine which method produces the better tree, since a tree can be considered superior in two different ways; either the tree could have more resolved nodes (i.e. minimizing polytomies) or it could be more logically consistent with the true tree (i.e. fewer erroneous relationships). As an example, one method might resolve several polytomies, but also propagate a few errors. Is it better to resolve 10 more nodes correctly at the cost of one or two errors? To handle this issue, we turned to statistical hypothesis testing. When viewing phylogenetics from a hypothesis testing framework, a polytomy (the inability to differentiate the phylogenetic relationship of taxa) represents a type II error (a false negative; failing to assert a true sister-group relationship) whereas a node

that incorrectly describes the phylogenetic relationships constitutes a type I error (a false positive; in this case positing an incorrect sister-group relationship). Our criterion is simple, preferring type II over type I error. Such a result is more conservative, and subsequently less likely to result in spurious patterns under meta-analysis. Therefore, we consider the method that consistently produces consensus trees with the minimal number of incorrect nodes to be preferable.

METHOD

Key to evaluating the performance of implied weighting is ascertaining its ability to retrieve a known tree topology in controlled conditions. For this to be implemented it is necessary to populate a matrix for a set number of taxa in a given tree topology. A new tree file (Fig. 1) was constructed in Mesquite (Maddison & Maddison 2010) with 22 taxa and equal branch lengths with the topology based upon the strict consensus of a previous phylogenetic analysis of Ordovician trilobites by Congreve & Lieberman (2010). We simulated 100 matrices of categorical character evolution with 55 binary characters using a Markov *k*-state 1 parameter (Mk1) model (Lewis 2001). Typically a Mk1 model uses a single parameter that defines the rate of character change as equal for both gains and losses across the whole matrix (a generalization of the Jukes–Cantor model and the standard likelihood model for morphological evolution for both character simulation (e.g. Agnarsson *et al.* 2006) and ancestral character reconstruction (e.g. Nakatani *et al.* 2011; Martínez *et al.* 2014)). However, we modified this method, following Wright & Hillis (2014) to set the parameter that defines character change by randomly drawing from a gamma distribution (with a shape parameter of 0.5 and a rate parameter of 5) for each character within the matrix. This allows for every character within the matrix to have a variable rate of change, which results in more naturalistic datasets by accounting for the observed patterns of mosaic evolution (de Beer 1954; Stebbins 1983; Hopkins & Lidgard 2012; Hunt *et al.* 2015) and allows for the overall levels of homoplasy within each dataset to be highly variable. This gamma distribution was chosen to ensure that the majority of our rates of character change did not exceed 0.2 because our sample tree assumes constant branch length. As the rate of change for a character approaches 1.0 on our sample tree, said character would need to change at each branching event, which is not realistic for morphological data and would result in very few parsimony informative characters. Constant and autapomorphous characters were removed from the matrices and replaced by additional variable characters to ensure that each matrix contained 55 characters with the potential for some phylogenetic information.

FIG. 1. Initial phylogenetic tree used to generate the morphological datasets, modified from Congreve & Lieberman (2010).

These matrices were then subjected to heuristic searches employing random addition sequences followed by branch swapping (the *mult* command) with 100 000 repetitions in TNT (Goloboff *et al.* 2008*b*; made available with the sponsorship of the Willi Hennig Society). Each matrix was searched with characters treated as equally weighted as well as with implied weighting (the *piwe* command) under a variety of concavity values (k = 1, 3, 5, and 10). The resulting trees were summarized through strict consensus, as topologies of most parsimonious trees are considered equally likely in relation to each other, meaning that selecting one most parsimonious tree out of several is extremely *ad hoc* while preferring a node based on its frequency of occurrence among equally parsimonious trees (as majority rule consensuses do) is equally problematic. These strict consensuses were then analysed by determining which nodes within each consensus tree implied groupings that were logically consistent (labelled as correct nodes) or inconsistent (labelled as incorrect nodes) with the true tree. Then, each consensus tree was scored by counting the number of correct nodes, the number of incorrect nodes (type I errors), and the number of unknown relationships (polytomies). This is essentially a simplified, node- as opposed to edge- (branch) based, version of the Robinson–Foulds metric (Robinson & Foulds 1981) where trees are compared to a single fixed tree rather than one another. The maximum total score for a perfectly resolved tree was 19:0:0, while a maximally imperfectly resolved tree could score 0:19:0, and a complete polytomy would score 0:0:19. All matrices and trees as well as the correct, incorrect and unknown counts are available in the data archive (Congreve & Lamsdell 2016).

Another important avenue of investigation is understanding the model of evolution implied weights is proposing. Felsenstein (1981) suggested that character

weights should be proportional to the log normal of the character's rate of change (i.e. the Mk values input into our evolutionary model). Therefore, the ideal weights for each character in this model of evolution can be computed by taking the log normal of each character's rate of change and then standardizing those values by the slowest character (thereby obtaining a step-cost weight from 0 to 1). Following the study of Wagner (2012), we generated four characters with Mk values that were sampled from the midpoints of the four quartiles of our gamma distribution (Yang 1994). These characters (with rates of change: 0.00247, 0.02389, 0.07870, 0.23533; characters 56–59 respectively in the matrices) were added to the data matrices from our previous replicates (available in Congreve & Lamsdell 2016). These replicates were then run under the same implied weights settings as our previous analyses, but these replicates were generated through PAUP*4 (Swofford 2002) using 10 000 random stepwise replicates and TBR. PAUP*4 was used in these analyses because TNT does not allow the user to see how the method is weighting each character. In instances where a single run generated multiple weights for a single character, we took the best fit. The weights generated from the implied weights runs under all four values of k were then compared to the expected weights of the four characters from Felsenstein's model.

RESULTS

Character weighting

It is difficult to graphically show how implied weighting is generating its weights. The problem is that the fit is entirely contingent on tree topology; in this way implied weights just acts as a successive weighting system (not a concurrent weighting system as it claims) and is circular in its methodology. The best way to show this is to give the equation for how fits are calculated:

$$f = \frac{k}{e + k}$$

where f is fit, k is the arbitrarily defined constant, and e is the number of extra steps that the character must take on the topology. Therefore a character undergoing a single state change will have an e value of 0:

$$f = \frac{k}{0 + k} = 1$$

Any increase in steps will result in successively lower fits. The value of k influences the base value and so will result in higher or lower overall fits. For example, a character with three extra steps will have a weight of 0.25 under a k of 1:

$$f = \frac{1}{3 + 1} = 0.25$$

However, under a k of 10 its weight will be 0.769:

$$f = \frac{10}{3 + 10} = 0.769$$

Obviously the value of k selected has a huge impact on the eventual weight of the character. Furthermore, as the value for e can only be assigned after a tree topology is retrieved, it is impossible to work out *a priori* what weight implied weighting will assign to a character. Presumably characters are expected to have a step count of one less than the number of states (so binary characters are expected to exhibit one change, three state characters two changes, etc.) although this is never made explicit in the TNT documentation.

Summary statistics of the weights/fits of our four constant characters generated from implied weights under various values of k are provided in Table 1. Implied weights generally produced weights for the four characters that were appropriately ranked (i.e. the slowest character was given the highest fit and the fastest character the lowest fit) although roughly 5 of our 100 replicates inverted the rankings for some of these character weights. The method generally identified and properly weighted the slowest character (character 56) for all values of k, with only a few erroneous weights generated from each k value (these errors occurred randomly across different replicates). Meanwhile, $k1$ generated weights for the fastest character that roughly approximated the expected weights, while weights generated from all other values of k were too high for the fastest character. Where the method deviates the most from our expected results is in its ability to estimate characters with moderate rates of change (i.e. some homology and some homoplasy). All values of k generated weights that were, for the most part, substantially far from the expected results, with only a few outliers of $k1$ approaching the expected weights. Of particular note is character 57, which generated the highest variance out of all of the characters. Given the method's inability to consistently approximate the weights expected under Felsenstein's model for these moderately evolving characters, we propose that implied weights is not producing weights that follow current conventional models of morphological evolution.

Overall accuracy

By comparing the total number of correct and incorrect nodes for equal weighting and all k values (Fig. 2), we see

TABLE 1. Statistics for the character weights/fits generated under implied weights from 100 matrices including the same 4 characters with known rates of change.

	Character 56 expected weight 1.000				Character 57 expected weight 0.622			
	*k*1	*k*3	*k*5	*k*10	*k*1	*k*3	*k*5	*k*10
Mean	0.99334	0.996	0.99571	0.99668	0.83682	0.88534	0.90777	0.93965
Median	1	1	1	1	1	0.8	0.857	0.917
Mode	1	1	1	1	1	0.8	0.857	0.917
Variance	0.002195	0.000792	0.000601	0.000267	0.029678	0.010779	0.00527	0.001602
Minimum	0.667	0.8	0.857	0.917	0.5	0.667	0.75	0.846
Maximum	1	1	1	1	1	1	1	1
	Character 58 expected weight 0.423				Character 59 expected weight 0.241			
	*k*1	*k*3	*k*5	*k*10	*k*1	*k*3	*k*5	*k*10
Mean	0.64496	0.78707	0.8444	0.90586	0.288193	0.44619	0.55206	0.69556
Median	0.667	0.8	0.857	0.917	0.286	0.444	0.545	0.688
Mode	0.667	0.8	0.857	0.917	0.286	0.444	0.545	0.688
Variance	0.003767	0.00179	0.001361	0.000832	0.000815	0.001366	0.00135	0.001121
Minimum	0.4	0.571	0.667	0.786	0.222	0.3	0.462	0.611
Maximum	0.667	0.8	0.857	0.917	0.3333	0.571	0.667	0.786

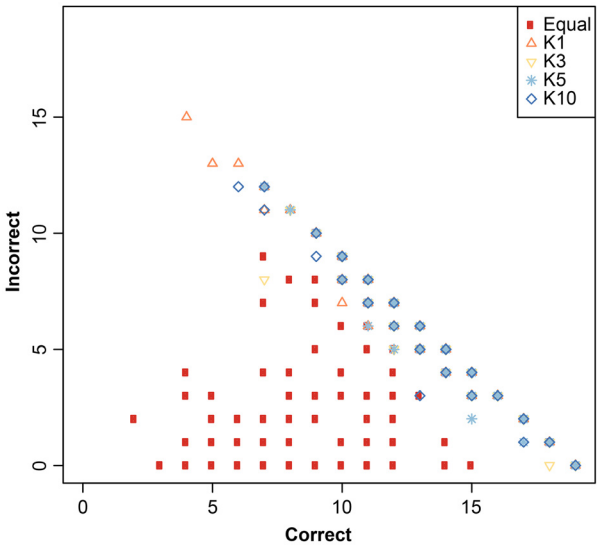

FIG. 2. Scatterplot of the total number of correct and incorrect nodes for all 100 replicates under equal weighting and implied weighting with *k* values of 1, 3, 5, and 10. Colour online.

that all values of *k* cluster closely together in a strong linear relationship that is defined by the total number of nodes possible in the tree (19). Meanwhile, values from equal weights are more randomly distributed throughout the scatterplot. This clustering of implied weights close to the limits of the data reflects the tendency for the method to consistently resolve ambiguous relationships when compared to equal weights, rarely if ever producing unknown relationships (polytomies). As we move from left to right on this line, implied weighting runs move from less accurate to more accurate as compared to the true evolutionary tree, however all values of *k* seem randomly distributed across this line.

In order to better parse out how well implied weighting performed relative to equal weights, we standardized each implied weighting run to equal weights by simply subtracting the correct/incorrect values of each equal weights tree from the correct/incorrect values in the implied weights consensus tree for the same run. In this manner we can observe whether implied weights produced trees that were more or less correct/incorrect than equal weights (Fig. 3). Data points above the x-axis represent implied weights trees that have more incorrect nodes than their equal weights counterparts while below the x-axis represent trees with fewer incorrect nodes. Similarly, data points to the left of the y-axis represent implied weights trees with fewer correct nodes than their equal weights counterparts, while to the right of the y-axis represent trees with more correct nodes. Different *k* values plot seemingly randomly within a dispersed distribution that is bounded within two lines, the bottom line resulting from implied weights consistently pushing to completely resolve its trees without polytomies and the top line resulting from the limits of the total number of nodes in the analysis. While no clear clusters exist, if we divide up the datasets into three groups related to how well the data performs relative to equal weights, 12% of the analyses increase incorrect nodes and decrease correct nodes relative to equal weights, 56% of the analyses increase both incorrect and correct nodes relative to equal weights, and

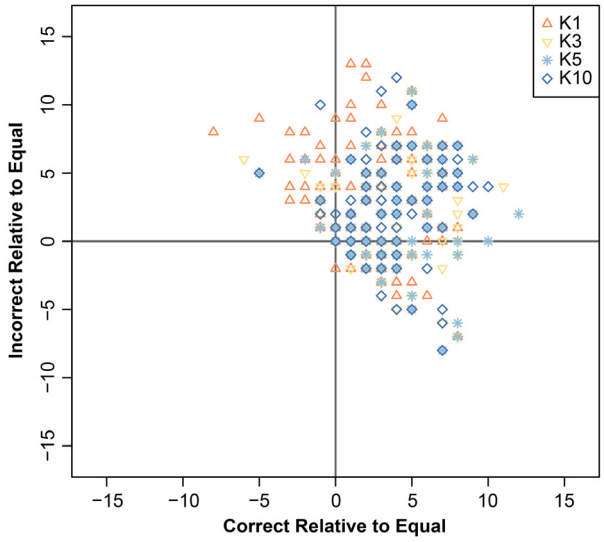

FIG. 3. Scatterplot of the total number of correct and incorrect nodes relative to equal weighting for all 100 replicates under implied weighting with *k* values of 1, 3, 5, and 10. Data points to the left and right of centre represent trees generated under implied weights with fewer correct or more correct nodes relative to the equal weights tree respectively. Data points below and above centre represent trees generated under implied weights with fewer incorrect or more incorrect nodes relative to the equal weights tree respectively. Colour online.

30% of the analyses increase correct nodes and decrease incorrect nodes relative to equal weights. This suggests that a total of 68% of our implied weights runs increase the total number of incorrect nodes relative to equal and therefore increase the amount of type I errors.

Resolving polytomies

To investigate the effects of ambiguous relationships (polytomies) on our data, we first plotted both the raw correct and incorrect node values against unknowns for equal weights and all values of *k* (Fig. 4A). As expected, implied weighting replicates consistently group together on the far left of the graph since the method tends to resolve nearly all of the nodes on the tree, rarely resulting in more than one most parsimonious tree. Equal weight values are randomly distributed throughout, although there is an interesting distinction between the correct vs unknown plots and the incorrect vs unknown plots. Equal weight values cluster along a linear relationship defined by the overall limits of the data in the correct plots but stay below the line in the incorrect plots. This suggests that equal weights is tending to push the limits of the data when resolving nodes correctly, but is perhaps more conservative when it comes to resolving nodes incorrectly.

To determine how effectively implied weighting resolves polytomies occurring in equal weighted analyses, we conducted a similar standardization on our data as when looking at overall accuracy. In these standardized scatterplots (Fig. 4B) points moving further left represent instances where implied weighting is resolving polytomies

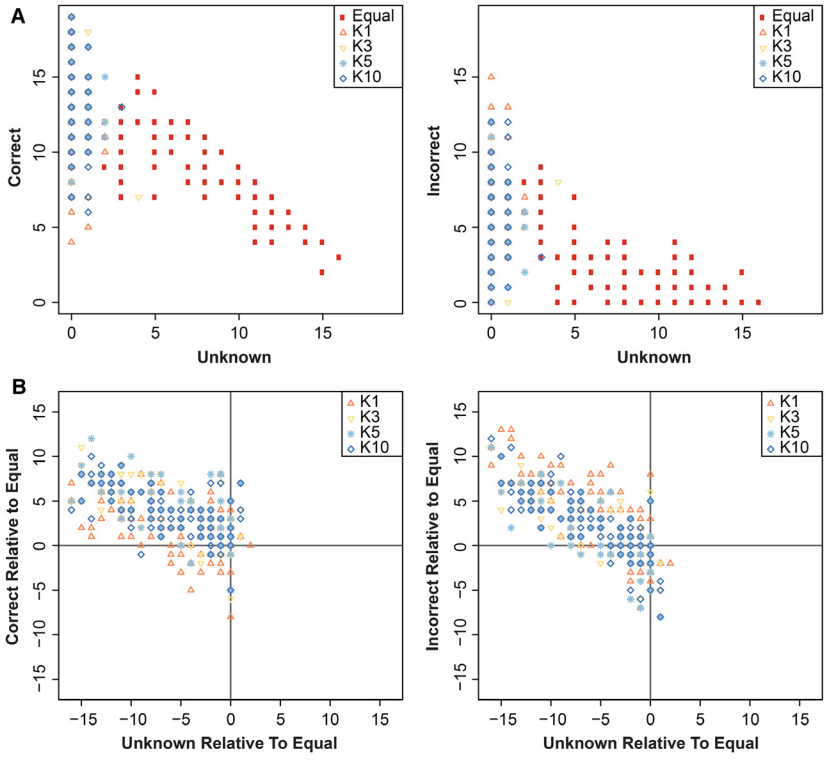

FIG. 4. A, 'Raw' scatterplots represent plots of correct nodes vs unknowns (left) and incorrect nodes vs unknowns (right). Equal weights on the top left scatterplot follows closely the line which is defined by the limits of the total number of nodes in the dataset, while equal weights on the top right scatterplot does not. B, 'Standardized' scatterplots represent plots of correct relative to equal vs unknown relative to equal (left) and incorrect relative to equal vs unknown (right) relative to equal. As data points move left on the x-axis, they represent more polytomies in equal weights that implied weights is resolving. Colour online.

present in the equal weighted analyses. Both standardized scatterplots show some linear relationship; however, the slope of the incorrect relative to equal vs unknown relative to equal is much steeper with tighter clustering of the overall data, suggesting that there is a strong correlation towards a greater increase in error as the number of polytomies resolved increases. To examine this in further detail, regression analyses were conducted on the standardized plots of both the correct vs unknown and incorrect vs unknown for all values of *k* (Fig. 5). In all instances the slope of the regression line for the incorrect vs unknown plot was substantially steeper than the correct vs unknown plots. While all regression lines were shown to be statistically significant, it is important to note that the r-squared values of the incorrect vs unknown plots were higher than those of the correct vs

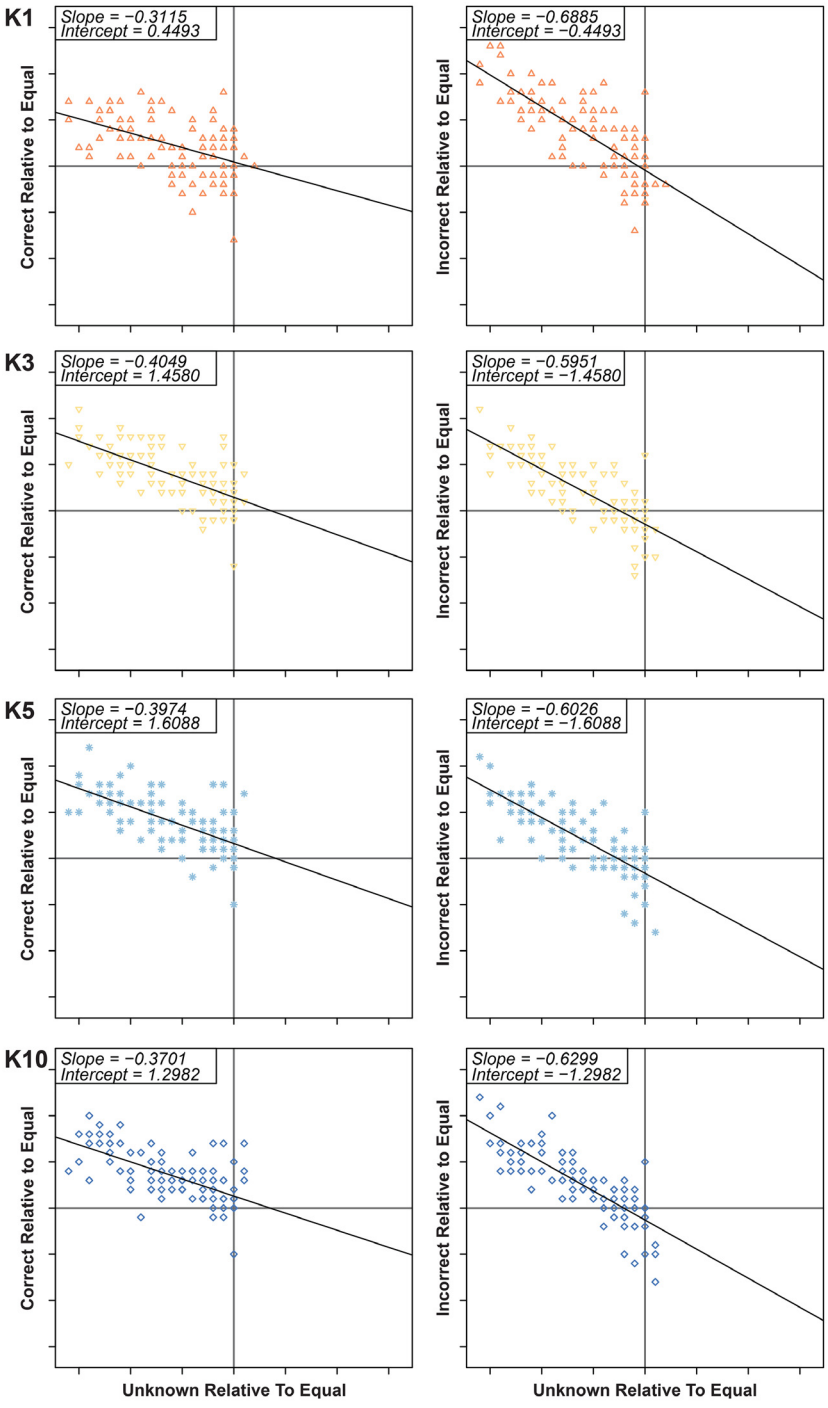

FIG. 5. Scatterplots with regression lines of correct relative to equal vs unknown relative to equal and incorrect relative to equal vs unknown relative to equal for each *k* value. Slopes on the incorrect relative to equal vs unknown relative to equal are consistently steeper than slopes on the correct relative to equal vs unknown relative to equal for all values of *k*. Colour online.

unknown plots (incorrect: $k1 = 0.601$, $k3 = 0.617$, $k5 = 0.618$, $k10 = 0.665$; correct: $k1 = 0.232$, $k3 = 0.421$, $k5 = 0.412$, $k10 = 0.413$) which implies a closer linear relationship between incorrect nodes vs unknown than correct nodes vs unknown. This suggests a stronger tendency for implied weighting to resolve polytomies in equal weighted analyses incorrectly rather than correctly.

Consistency of errors in equal weights across multiple k *values*

To determine how often errors in equal weighted analyses propagated across multiple values of k within a single dataset (i.e. implied weights trees retrieved an identical erroneous sister-group topology as the equal weight tree), we conducted a simple count of the number of times at least one error present in the equal weighted analysis appeared in at least three k values and all four k values (Congreve & Lamsdell 2016). A total of 81 of our analyses contained at least one incorrect node under equal weighting. Of these analyses, 43 propagated errors into all four k values, and 57 propagated errors into at least three k values. Our results suggest that when equal weighting incorrectly resolves a node, this error is repeated in four different k values roughly 50% of the time, and in at least three different k values roughly 70% of the time.

DISCUSSION

Efficacy of implied weighting

Simulation studies have shown that maximum parsimony performs worse when rates of change (and in turn homoplasy) are heterogeneous (Huelsenbeck 1994; Kuhner & Felsentein 1994; Wagner 1998) and implied weighting implicitly seeks to improve the performance of parsimony in these situations (Goloboff *et al*. 2008*a*). As such, in incorporating rate heterogeneity our simulations were designed to favour implied weighting over equal weights parsimony. However, the results of our analyses suggest that while implied weighting tends to be far less conservative than equal weighting, this is often to the method's detriment. Like other weighting methods, implied weighting consistently works to resolve ambiguity, but our results suggest that in doing so it tends to increase the number of erroneous relationships relative to equally weighted analyses. While implied weights can improve an analysis relative to equal weights, it does so seemingly randomly, with no clearly superior k value and no strong trend within the data. Again, because the level of homoplasy within an analysis is unknowable *a priori*, it is impossible to choose the single best k value that will improve the overall accuracy of our phylogenetic hypothesis. Furthermore, if we treat the process of picking the correct k value for an analysis as being effectively random, we run into a major problem. When compared directly to equal weights, 68% of the implied weights trees showed a clear increase in the number of incorrectly resolved nodes. If all we want to do is to minimize the number of incorrect nodes (type I errors), then this presents a problem for implied weights because the general pattern is an overall increase in error relative to equal weights.

However, it is important to consider that this perspective might be skewed due to how we treated polytomies in our analysis. As a hypothetical example, given how we have scored this data, an analysis in which the equal weights consensus was entirely a polytomy would by definition be considered superior to an implied weights tree with only one error. Such an example would not be a fair treatment of the accuracy of implied weights. This is why we also compared the number of correct and incorrect nodes with the number of unknowns, and the strong correlation we recovered between resolving polytomies and increasing error in implied weights is troubling. Our data suggests that using implied weighting to resolve datasets that are highly conflicted in equal weights results in a high number of erroneous relationships. Generally speaking equal weighting tends to favour collapsing nodes into polytomies in the consensus tree rather than erroneously reconstructing parts of the tree. Implied weights simply 'picks' a solution to these conflicts seemingly at random, and more often than not it tends to be incorrect. We strongly advise against using implied weights to resolve polytomies found in equal weighted analyses since it shows a far stronger trend towards incorrectly resolving data rather than correctly resolving data. This adds a further note of caution to the practice of resolving polytomies, particularly via purely methodological means, as spurious resolutions can have serious impact on the results of downstream meta-analyses (Rabosky 2015). It should be remembered that the occurrence of polytomies in and of themselves is not detrimental to a result, and researchers should refrain attempting to force a methodological solution when only further data can elucidate the issue with any accuracy.

Our results also corroborate the accusations of tautology that have plagued implied weights, with a clear trend towards propagating erroneous relationships present in equal weighted analyses. Furthermore, the method will sometimes randomly break down correct relationships recovered in equal weights. As such, we find little utility in using implied weights to gauge the accuracy of equal weighted analyses since congruence or disagreement between the two methods does not necessarily convey any real information about the accuracy of the data. Furthermore, since implied weights is functioning like other

weighting methods, some degree of congruence between equal weighted analyses and implied weights should always be expected because they have a circular relationship.

Handling missing data

Goloboff (2014) recently proposed a solution to dealing with the issue of missing data under implied weighting, based on the assumption that characters with a large number of missing entries will be unfairly upweighted in the analysis as homoplastic state changes are not sampled. Under this scheme, missing data are assumed to add some level of homoplasy to the character, although precisely how much depends on the number of missing entries per character and on a model defined by the investigator. There are a number of undesirable side effects to this solution, however, that make its implementation troublesome. The first is that it has the potential to be overly punitive to characters with a large amount of missing data. This is problematic as downweighting these characters based on the potential for homoplasy has the potential to drown out any signal they impart to the analysis. Much work has been done on the influence of missing data on phylogenetic inference, often within the realm of incomplete taxa, but with increasing focus on characters with a high proportion of missing data (Kearney 2002). It has been comprehensively shown that incompletely coded characters can provide a strong signal (Wiens 2003*b*), and that characters with high levels of missing data can still improve phylogenetic accuracy under many conditions (Wiens 1998, 2003*a*, 2006; Edgecombe 2010; Wiens & Morrill 2011), while ignoring characters because of missing data can lead to loss of resolution (Kearney & Clark 2003). Assuming homoplasy in missing data under implied weighting effectively results in characters that exhibit missing data being unfairly discounted during the analysis and, by extension, dismisses the signal which such characters have been shown to impart.

The second issue is that implied weighting negatively impacts characters that are inapplicable for certain taxa. Inapplicable codings arise when a characteristic is contingent on the presence of another character that is absent in the target taxon. Rather than assign a specific state, the character is considered to be inapplicable and is treated as missing data during the analysis (Maddison 1993; Strong & Lipscomb 1999). Alternative strategies of treating inapplicables as separate character states (often a repetition of absence) or formulating every character state as a separate binary character (Pleijel 1995) result in extra emphasis being placed on the absence of a single structure and act as a form of weighting (Seitz *et al.* 2000), skewing the analysis so that taxa lacking the structure are more likely to resolve together. In contrast, treating inapplicables as missing data does not bias the analysis as missing data cannot produce groupings for which there is no evidence (Kearney 2002), so this is the preferred protocol for handling such situations. This does, however, result in an increase of missing data; even among living taxa almost a quarter of coded characters in analyses can be treated as missing due to inapplicable characters (Edgecombe 2010). Goloboff's (2014) new method for handling missing data is therefore wholly inappropriate to use with inapplicable characters as it will result in assuming homoplasy where no homoplasy can exist, in turn penalizing characters through downweighting simply because they are inapplicable for some taxa.

CONCLUSIONS

Our data suggest that implied weighting often results in unpredictable fixation on sometimes spurious tree topologies. When the method is consistent it often mirrors the results of equal weight analyses, including replicating errors. The method is particularly poor at resolving datasets with large amounts of conflict/polytomies. Under equal weighting, the retrieved consensus trees are shown to be generally more conservative (i.e. higher frequency of polytomies), resulting in fewer erroneous nodes. While lack of resolution is certainly problematic, resolving trees based on a potentially random topology is more problematic in terms of our understanding of taxonomic relationships and evolutionary patterns and processes.

Recent work (Legg *et al.* 2013; Garwood & Dunlop 2014; Smith & Ortega-Hernandez 2014) has treated implied weighting as a stress test for the robustness of phylogenetic results. While such a method has its merits, and is certainly preferable to placing favour on trees retrieved under a single value of k, our analysis indicates two major problems with this approach. The first is that errors which appear in the equal weight analysis are propagated under implied weights, therefore suggesting that a stress test of combining equal weights with implied weights under a variety of k values could result in no conflict despite the relationships being wrong. Second, the method occasionally generates more erroneous nodes than the equal weight analyses, thereby implying conflicts that do not exist. Since it is inconsistent with regard to whether or not it is conservative, this lack of predictable behaviour limits its utility as a stress test. Furthermore, the inconsistent behaviour of implied weighting means that the method should not be used to conclusively determine sister taxon relationships (*contra* Xu & Pol 2014).

Given that application of statistical methods (such as likelihood and Bayesian inference) to phylogenetic analysis of morphological data is at present still far from ubiquitous (Kolaczkowski & Thornton 2004; Lee &

Worthy 2011; Wright & Hillis 2014), parsimony is currently the most widespread option for palaeontological research on morphology. As appropriate models for statistical inference of morphological change continue to be developed and refined (Lee & Palci 2015), it remains important to understand the limits and strengths of various parsimony methods. It should be stated that the topologies retrieved from previous implied weight analyses should not be dismissed outright, given the random behaviour of implied weighting in our analyses. Rather, implied weight trees should not be preferred over topologies retrieved under equal weights, and it would be advisable for authors to report both in future studies. While we would be cautious to over generalize, perhaps Patterson's (1982) statement still rings true today: good characters weight themselves.

Acknowledgements. We would like to thank Martin Stein (Natural History Museum of Denmark), Mark Patzkowsky (Pennsylvania State University), and Jacques Gautier (Yale University) for discussions and observations that were vital to formulating the methods used in our analysis. The paper benefited greatly from detailed and insightful reviews from David Bapst (South Dakota School of Mines and Technology) and Peter Wagner (Smithsonian Institution). Both authors contributed equally to the manuscript. Authorship priority was decided by a six game competition consisting of best of three rounds of Divekick, Towerfall Ascension, STARWHAL, Nidhogg, Ultra Street Fighter IV, and Super Puzzle Platformer. CRC won 4 to 2.

DATA ARCHIVING STATEMENT

The following appendices containing data for this study are available in the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.7dq0j):
Appendix S1: TNT files
Appendix S2: Trees
Appendix S3: Spreadsheet of node counts
Appendix S4: Run counts of repeated errors
Appendix S5: Character weights test NEXUS files
Appendix S6: Character weights.

Editor. Marcello Ruta

REFERENCES

AGNARSSON, I., AVILÉS, L., CODDINGTON, J. A. and MADDISON, W. P. 2006. Sociality in theridiid spiders: repeated origins of an evolutionary dead end. *Evolution*, **60**, 2342–2351.

BEER G. de 1954. *Archaeopteryx* and evolution. *Advancement of Science*, **11**, 160–170.

CANNATELLA, D. C. and DE QUEIROZ, K. 1989. Phylogenetic systematics of the anoles: is a new taxonomy warranted? *Systematic Zoology*, **38**, 57–69.

CARPENTER, J. M. 1994. Successive weighting, reliability and evidence. *Cladistics*, **10**, 215–220.

—— STRASSMAN, J. E., TURILLAZZI, S., HUGHES, C. R., SOLIS, C. R. and CERVO, R. 1993. Phylogenetic relationships among paper wasp social parasites and their hosts (Hymenoptera: Vespidae: Polistinae). *Cladistics*, **9**, 129–146.

CERDEÑO, E., VERA, B. and SCHMIDT, G. I. 2012. An almost complete skeleton of a new Mesotheriidae (Notoungulata) from the Late Miocene of Casira, Bolivia. *Journal of Systematic Palaeontology*, **10**, 341–360.

CHIPPINDALE, P. T. and WEINS, J. J. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. *Systematic Biology*, **43**, 278–287.

—— BONETT, R. M., BALDWIN, A. S. and WIENS, J. J. 2004. Phylogenetic evidence for a major reversal of life-history evolution in plethodontid salamanders. *Evolution*, **58**, 2809–2822.

COLLIN, R. and MIGLIETTA, M. P. 2008. Reversing opinions on Dollo's Law. *Trends in Ecology & Evolution*, **23**, 602–609.

CONGREVE, C. R. and LAMSDELL, J. C. 2016. Data from: Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Dryad Digital Repository*. doi: 10.5061/dryad.7dq0j

—— and LIEBERMAN, B. S. 2010. Phylogenetic and biogeographic analysis of deiphonine trilobites. *Journal of Paleontology*, **84**, 128–136.

CRUZ MENDES, A. 2011. Phylogeny and taxonomic revision of Heteropachylinae (Opiliones: Laniatores: Gonyleptidae). *Zoological Journal of the Linnean Society*, **163**, 437–483.

EDGECOMBE, G. D. 2010. Palaeomorphology: fossils and the inference of cladistic relationships. *Acta Zoologica*, **91**, 72–80.

ELDREDGE, N. I. and CRACRAFT, J. 1980. *Phylogenetic patterns and the evolutionary process*. Columbia University Press, New York.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. *Systematic Zoology*, **18**, 374–385.

—— 1970. Methods for computing Wagner trees. *Systematic Zoology*, **19**, 83–92.

—— 1983. The logical basis of phylogenetic analysis. 7–36. *In* PLATNICK, N. I. and FUNK, V. A. (eds). *Advances in cladistics, vol. 2*. Proceedings of the Second Meeting of the Willi Hennig Society, Columbia University Press, New York, 593 pp.

FELSENSTEIN, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society*, **16**, 183–196.

FITCH, W. M. 1970. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, **19**, 99–113.

GARWOOD, R. J. and DUNLOP, J. A. 2014. Three-dimensional reconstruction and the phylogeny of extinct chelicerate orders. *PeerJ*, **2**, 1–33.

GAUTHIER, J. A., KEARNEY, M., ANDERSON MAISANO, J., RIEPPEL, O. and BEHLKE, A. D. B. 2012. Assembling the squamate tree of life: perspectives from the phenotype and the fossil record. *Bulletin of the Peabody Museum of Natural History*, **53**, 3–308.

GHISELIN, M. T. 2005. Homology as a relation of correspondence between parts of individuals. *Theory in Biosciences*, **124**, 91–103.

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

—— 1995. Parsimony and weighting: a reply to Turner and Zandee. *Cladistics*, **11**, 91–104.

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

—— 2014. Extended implied weighting. *Cladistics*, **30**, 260–272.

—— CARPENTER, J. M., ARIAS, J. S. and ESQUIVEL, D. F. M. 2008*a*. Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics*, **24**, 1–16.

—— FARRIS, J. A. and NIXON, K. C. 2008*b*. TNT, a free program for phylogenetic analysis. *Cladistics*, **24**, 774–786.

GOULD, S. J. 1970. Dollo on Dollo's Law: irreversibility and the status of evolutionary laws. *Journal of the History of Biology*, **3**, 189–212.

GREER, A. E. 1991. Limb reduction in squamates: identification of the lineages and discussion of the trends. *Journal of Herpetology*, **52**, 166–173.

HENNIG, W. 1966. *Phylogenetic systematics*. University of Illinois Press, Urbana, 280 pp.

—— and SCHLEE, D. 1978. Abriss der phylogenetischen systematik. *Stuttgarter Beiträge zur Naturkunde Serie A (Biologie)*, **319**, 1–11.

HOLDER, M. T., LEWIS, P. O. and SWOFFORD, D. L. 2010. The Akaike Information Criterion will not chose the No Common Mechanism model. *Systematic Biology*, **59**, 477–485.

HOPKINS, M. J. and LIDGARD, S. 2012. Evolutionary mode routinely varies among morphological traits within fossil species lineages. *Proceedings of the National Academy of Sciences*, **109**, 20520–20525.

HUELSENBECK, J. P. 1994. Performance of phylogenetic methods in simulation. *Systematic Biology*, **44**, 17–48.

—— SWOFFORD, D. L., CUNNINGHAM, C. W., BULL, J. J. and WADDELL, P. J. 1994. Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? *Systematic Biology*, **43**, 288–291.

—— ALFARO, M. E. and SUCHARD, M. A. 2011. Biologically inspired phylogenetic models strongly outperform the No Common Mechanism model. *Systematic Biology*, **60**, 225–232.

HUGHES, M., GERBER, S. and WILLS, M. A. 2013. Clades reach highest morphological disparity early in their evolution. *Proceedings of the National Academy of Sciences*, **110**, 13875–13879.

HUNT, G., HOPKINS, M. J. and LIDGARD, S. 2015. Simple versus complex models of trait evolution and stasis as a response to environmental change. *Proceedings of the National Academy of Sciences*, **112**, 4885–4890.

JONES, F. M., DUNLOP, J. A., FRIEDMAN, M. and GARWOOD, R. J. 2014. *Trigonotarbus johnsoni* Pocock, 1911, revealed by X-ray computed tomography, with a cladistic analysis of the extinct trigonotarbid arachnids. *Zoological Journal of the Linnean Society*, **172**, 49–70.

KÄLLERSJÖ, M., ALBERT, V. A. and FARRIS, J. S. 1999. Homoplasy *increases* phylogenetic structure. *Cladistics*, **15**, 91–93.

KEARNEY, M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. *Systematic Biology*, **51**, 369–381.

—— and CLARK, J. M. 2003. Problems due to missing data in phylogenetic analyses including fossils: a critical review. *Journal of Vertebrate Paleontology*, **23**, 263–274.

KLUGE, A. G. 1988. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Biology*, **38**, 7–25.

—— 1997*a*. Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic systematics. *Zoologica Scripta*, **26**, 349–360.

—— 1997*b*. Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics*, **13**, 81–96.

—— and FARRIS, J. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, **18**, 1–32.

KOENEMANN, S., JENNER, R. A., HOENEMANN, M., STEMME, T. and VON REUMONT, B. M. 2010. Arthropod phylogeny revisited, with a focus on crustacean relationships. *Arthropod Structure & Development*, **39**, 88–110.

KOHLSDORF, T., CUMMINGS, M. P., LYNCH, V. J., STOPPER, G. F., TAKAHASHI, K. and WAGNER, G. P. 2008. A molecular footprint for limb loss: sequence variation of the autopodial identity gene *Hoxa-13*. *Journal of Molecular Evolution*, **67**, 581–593.

KOLACZKOWSKI, B. and THORNTON, J. W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, **431**, 980–984.

KUHNER, M. K. and FELSENTEIN, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology & Evolution*, **11**, 459–468.

LANKESTER, E. R. 1870. On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. *Annals & Magazine of Natural History*, **6**, 34–43.

LEE, M. S. Y. 2005. Squamate phylogeny, taxon sampling, and data congruence. *Organisms Diversity & Evolution*, **5**, 25–45.

—— and PALCI, A. 2015. Morphological phylogenetics in the genomic age. *Current Biology*, **25**, R922–R929.

—— and WORTHY, T. H. 2011. Likelihood reinstates *Archaeopteryx* as a primitive bird. *Biology Letters*, **8**, 299–303.

LEGG, D. A. 2013. Multi-segmented arthropods from the middle Cambrian of British Columbia (Canada). *Journal of Paleontology*, **87**, 493–501.

—— and CARON, J.-B. 2014. New middle Cambrian bivalved arthropods from the Burgess Shale (British Columbia, Canada). *Palaeontology*, **57**, 691–711.

—— and VANNIER, J. 2013. The affinities of the cosmopolitan arthropod *Isoxys* and its implications for the origin of arthropods. *Lethaia*, **46**, 540–550.

—— SUTTON, M. D., EDGECOMBE, G. D. and CARON, J.-B. 2012. Cambrian bivalved arthropod reveals origin of arthrodization. *Proceedings of the Royal Society B*, **279**, 4699–4704.

—— —— —— 2013. Arthropod fossil data increase congruence of morphological and molecular phylogenies. *Nature Communications* **4**, 1–7.

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

MADDISON, W. P. 1993. Missing data versus missing characters in phylogenetic analysis. *Systematic Biology*, **42**, 576–581.

—— and MADDISON, D. R. 2010. Mesquite: a modular system for evolutionary analysis. Version 2.73. http://mesquiteproject.org

MARTÍNEZ, A., DI DOMENICO, M. and WORSAAE, K. 2014. Gain of palps within a lineage of ancestrally burrowing annelids (Scalibregmatidae). *Acta Zoologica*, **95**, 421–429.

MIRANDE, J. M. 2009. Weighted parsimony phylogeny of the family Characidae (Teleostei: Characiformes). *Cladistics*, **25**, 574–613.

MORRONE, J. J. 2014. Parsimony analysis of endemicity (PAE) revisited. *Journal of Biogeography*, **41**, 842–854.

NAKATANI, M. M., MIYA, M., MABUCHI, K., SAITOH, K. and NISHIDA, M. 2011. Evolutionary history of the Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. *BMC Evolutionary Biology*, **11**, 1–25.

OLESON, J. 2007. Monophyly and phylogeny of Branchiopoda, with focus on morphology and homologies of branchiopod phyllopodous limbs. *Journal of Crustacean Biology*, **27**, 165–183.

PATTERSON, C. 1982. Morphological characters and homology. 21–74. *In* JOYSEY, K. A. and FRIDAY, A. E. (eds). *Problems in phylogenetic reconstruction*. Academic Press, London, 442 pp.

PLEIJEL, F. 1995. On character coding for phylogeny reconstruction. *Cladistics*, **11**, 309–315.

PORTER, M. L. and CRANDALL, K. A. 2003. Lost along the way: the significance of evolution in reverse. *Trends in Ecology & Evolution*, **18**, 541–547.

RABOSKY, D. L. 2015. No substitute for real data: a cautionary note on the use of phylogenies from birth–death polytomy resolvers for downstream comparative analyses. *Evolution*, **69**, 3207–3216.

RICHTER, S., OLESEN, J. and WHEELER, W. C. 2007. Phylogeny of Branchiopoda (Crustacea) based on a combined analysis of morphological data and six molecular loci. *Cladistics*, **23**, 301–336.

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

RUTA, M., WAGNER, P. J. and COATES, M. I. 2006. Evolutionary patterns in early tetrapods. I. Rapid initial diversification followed by decrease in rates of character change. *Proceedings of the Royal Society B*, **273**, 2107–2111.

SANDERSON, M. J. 1991. In search of homoplastic tendencies: statistical inference of topological patterns in homoplasy. *Evolution*, **45**, 351–358.

SANGER, T. J. and GIBSON-BROWN, J. J. 2004. The developmental bases of limb reduction and body elongation in squamates. *Evolution*, **58**, 2103–2106.

SEITZ, V., ORTIZ GARCIA, S. and LISTON, A. 2000. Alternative coding strategies and the inapplicable data coding problem. *Taxon*, **49**, 47–54.

SIMPSON, G. G. 1950. Some principles of historical biology bearing on human origins. *Cold Spring Harbor Symposia on Quantitative Biology*, **15**, 55–66.

SMITH, M. R. and ORTEGA-HERNANDEZ, J. 2014. *Hallucigenia*'s onychoporan-like claws and the case for Tactopoda. *Nature*, **514**, 363–366.

STEBBINS, G. L. 1983. Mosaic evolution: an integrating principle for the modern synthesis. *Experientia*, **39**, 823–834.

STRONG, E. E. and LIPSCOMB, D. 1999. Character coding and inapplicable data. *Cladistics*, **15**, 363–371.

SWOFFORD, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, MA.

—— and OLSEN, G. J. 1990. Phylogeny reconstruction. 411–501. *In* HILLIS, M. and MORITZ, C. (eds). *Molecular systematics*. Sinauer Associates, Sunderland, MA, 655 pp.

TEOTÓNIO, H. and ROSE, M. R. 2001. Perspective: reverse evolution. *Evolution*, **55**, 653–660.

TSCHOPP, E., MATEUS, O. and BENSON, R. B. J. 2015. A specimen-level phylogenetic analysis and taxonomic revision of Diplodocidae (Dinosauria, Sauropoda). *PeerJ*, **3**, 1–298.

TURNER, H. and ZANDEE, R. 1995. The behaviour of Goloboff's tree fitness measure *f*. *Cladistics*, **11**, 57–72.

VAN VALEN, L. M. 1982. Homology and causes. *Journal of Morphology*, **173**, 305–312.

VOGT, L. 2002. Testing and weighting characters. *Organisms Diversity & Evolution*, **2**, 319–333.

WAGNER, P. J. 1998. A likelihood approach for evaluating estimates of phylogenetic relationships among fossil taxa. *Paleobiology*, **24**, 430–449.

—— 2012. Modelling rate distributions using character compatability: implications for morphological evolution among fossil invertebrates. *Biology Letters*, **8**, 143–146.

WEISS, F. E., MALABARBA, L. R. and CLAUDIA, M. 2012. Phylogenetic relationships of *Paleotetra*, a new characiform fish (Ostariophysi) with two new species from the Eocene–Oligocene of south-eastern Brazil. *Journal of Systematic Palaeontology*, **10**, 73–86.

WIENS, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? *Systematic Biology*, **47**, 625–640.

—— 2003*a*. Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? *Journal of Vertebrate Paleontology*, **23**, 297–310.

—— 2003*b*. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology*, **52**, 528–538.

—— 2006. Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics*, **39**, 34–42.

—— and MORRILL, M. C. 2011. Missing data in phylogenetic analyses: reconciling results from simulations and empirical data. *Systematic Biology*, **60**, 719–731.

WILEY, E. O. 2008. Homology, identity and transformation. 9–21. *In* ARRATIA, G., SCHULTZE, H.-P. and WILSON, V. H. (eds). *Mesozoic fishes 4. Homology and phy-*

logeny. Verlag Dr. Friedrich Pfeil, München, Germany, 502 pp.

—— and LIEBERMAN, B. S. 2011. *Phylogenetics: theory and practice of phylogenetic systematics*. Wiley-Blackwell, New Jersey, 406 pp.

WRIGHT, A. and HILLIS, D. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of a phylogeny from discrete morphological data. *PLoS One*, **9**, 1–6.

XU, X. and POL, D. 2014. *Archaeopteryx*, paravian phylogenetic analyses, and the use of probability-based methods for palaeontological datasets. *Journal of Systematic Palaeontology*, **12**, 323–334.

YANG, Z. 1994. Maximum phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, **39**, 306–314.