WILEY-BLACKWELL

Cladistics 24 (2008) 1–16

Cladistics

10.1111/j.1096-0031.2008.00209.x

Weighting against homoplasy improves phylogenetic analysis of morphological data sets

Pablo A. Goloboff[a]*, James M. Carpenter[b], J. Salvador Arias[c] and Daniel Rafael Miranda Esquivel[c]

[a]*CONICET, Instituto Miguel Lillo, Miguel Lillo 205, 4000 SM de Tucumán, Argentina;* [b]*Division of Invertebrate Zoology, The American Museum of Natural History, Central Park at 79th St, New York, NY 10024, USA;* [c]*Laboratorio de Sistemática y Biogeografía, Escuela de Biología, Universidad Industrial de Santander, A.A. 678 Bucaramanga, Colombia*

Accepted 10 December 2007

Abstract

The problem of character weighting in cladistic analysis is revisited. The finding that, in large molecular data sets, removal of third positions (with more homoplasy) decreases the number of well supported groups has been interpreted by some authors as indicating that weighting methods are unjustified. Two arguments against that interpretation are advanced. Characters that collectively determine few well-supported groups may be highly reliable when taken individually (as shown by specific examples), so that inferring greater reliability for sets of characters that lead to an increase in jackknife frequencies may not always be warranted. But even if changes in jackknife frequencies can be used to infer reliability, we demonstrate that jackknife frequencies in large molecular data sets are actually improved when downweighting characters according to their homoplasy but using properly rescaled functions (instead of the very strong standard functions, or the extreme of inclusion/exclusion); this further weakens the argument that downweighting homoplastic characters is undesirable. Last, we show that downweighting characters according to their homoplasy (using standard homoplasy-weighting methods) on 70 morphological data sets (with 50–170 taxa), produces clear increases in jackknife frequencies. The results obtained under homoplasy weighting also appear more stable than results under equal weights: adding either taxa or characters, when weighting against homoplasy, produced results more similar to original analyses (i.e., with larger numbers of groups that continue being supported after addition of taxa or characters), with similar or lower error rates (i.e., proportion of groups recovered that subsequently turn out to be incorrect). Therefore, the same argument that had been advanced against homoplasy weighting in the case of large molecular data sets is an argument in favor of such weighting in the case of morphological data sets.

Character weighting in cladistics has traditionally been a controversial issue. Authors in favor of weighting had usually considered that characters with more homoplasy are less reliable. This was the basis of Farris's (1969) successive weighting method and its non-iterative descendants, implied weighting and auto-weighted optimization (Goloboff, 1993, 1997; see general discussion in De Laet, 1997). However, Källersjö et al. (1999), in analyzing a large rbcL matrix (with about 2500 taxa, hereinafter rbcL-2500), found that the number of well supported groups and average jackknife resampling frequency, were strongly decreased when the most homoplastic characters (the third positions) were eliminated from the analysis. Källersjö et al. (1999) observed that "contrary to earlier expectations, increasing saturation and frequency of change … actually improve the ability to recognize well-supported phylogenetic groups." In addition to concluding that the common practice of eliminating third positions from phylogenetic analysis is probably pernicious, Källersjö et al. (1999, p. 93) also suggested that weighting methods that "rest on the idea that more homoplasy implies less reliability and less structure … may not be well advised."

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

As Källersjö et al. (1999) considered that their findings provided possible arguments against down-weighting characters on the basis of homoplasy, Farris (2001) proposed an alternative method, support weighting. Support weighting relates reliability to the number of well-supported groups set off by the character (i.e., the number of well-supported groups for which changes in the character appear as synapomorphies), and is explicitly intended to estimate weights regardless of homoplasy. Farris (2001) tested the method on the jackknife tree for rbcL-2500, and it gave third positions (which discriminate more groups) higher weights than first and second (which discriminate very few groups).

Although Källersjö et al. (1999) did not consider their results as providing evidence against weighting in general, other authors did. Even authors who had otherwise used only philosophical (or merely rhetorical) arguments to champion the exclusive and mandatory use of equal weights have referred to the empirical findings of Källersjö et al. with great approval (e.g., Grant and Kluge, 2005, p. 602; Kluge, 2005, p. 27).

A reanalysis of rbcL-2500, presented below, shows that the groups supported by first and second positions, even if few, are compatible with those groups supported by third positions. In other words, even if first and second positions distinguish few groups, they do so reliably. The extrapolation from jackknife frequencies to reliability of individual characters is not justified. Furthermore, using implied weighting to analyze large molecular data sets improves jackknife frequencies (and associated measures), as long as the weighting strength is properly rescaled.

In the case of morphological data sets, a trend opposite to that of Källersjö et al. (1999) had been documented before. Goloboff (1997; using 14 morphological data sets, with 14–47 taxa) showed that average jackknife frequencies were increased, relative to equal weights, when using either implied weighting, successive weighting, or self-weighted optimization. Ramírez (2003) also documented a similar trend. The present paper reports the most extensive comparison carried out to date between the results under equal and differential character weighting in morphological data sets. Our results show that (for morphological data), jackknife frequencies and other resampling measures are clearly improved when weighting against homoplasy. This is the same criterion that critics (e.g., Grant and Kluge, 2005, p. 602; Kluge, 2005, p. 27) had used to argue against weighting in the case of large molecular data sets. Therefore, defending equal weights on the basis of jackknife frequencies in the case of large molecular data sets requires—by the same logic—that weighting against homoplasy be defended in the case of morphological data sets.

Kluge's criticisms of weighting

Kluge (1997a,b, 2005) has published the most prominent and vocal criticisms against weighting, pretending that he has rejected weighting on the basis of Popper's ideas on falsification, and the very concepts of the nature of evidence and objectivity in science. Thus (the reader is led to conclude), those in favor of weighting oppose Popper, evidence, objectivity and scientific methods; but we do not.

Kluge (1997b) repeats Turner and Zandee's (1995) characterization of weighting as producing "unparsimonious" trees (simply by defining "most parsimonious" as "having fewest steps under equal weights"), despite the fact that Goloboff (1995) had replied to exactly that same argument. Just like Turner and Zandee before, Kluge (1997b) appeals to Farris's (1983) demonstration that parsimony maximizes explanatory power, and pretends that Farris (1983) showed that weighted hypotheses provide defective explanation. But if the support for arguments against weighting is supposed to come from Farris (1983), then there is no support at all: Farris (1983, p. 1011) had been (as noted by Goloboff, 1993, 1995) explicit that parsimony is not equivalent to equal weights, and that step counts must be weighted step counts, when some characters represent stronger evidence than others. Kluge (2005, p. 27) seems later to have realized this much, because he no longer cites Farris (1983) in support of the idea that weighting leads to unparsimonious hypotheses: "weighting leads to suboptimal, less-parsimonious, not more parsimonious, phylogenetic hypotheses when it comes to the data of observation (Kluge, 1997b; see, however, Farris, 1983)." It is especially ironic that Kluge (2005) cites Kluge (1997b) as providing justification for his statement regarding weighting and unparsimonious hypotheses, because there is no justification in Kluge (1997b) other than an appeal to Farris (1983).

Kluge (1997b) also refers to Popper's formula of corroboration,

$$C_{h,e,b} = \frac{p_{(e,hb)} - p_{(e,b)}}{p_{(e,hb)} + p_{(e,b)} - p_{(he,b)}}$$

(where C = corroboration, e = evidence, h = hypothesis, and b = background knowledge; see Farris, 1995 for discussion of this formula, and note here that Popper never meant his formula to be applied in real cases, as the actual values of the terms cannot be objectively measured; Popper simply used it to illustrate some general relationships between probability and corroboration). According to Kluge (1997b, p. 352):

> all of the justifications for differential character weighting … follow a verificationist agenda—the application of weights supposedly improves one's chances of discovering objective truth … Weighting under any such guise negatively impacts C

> and S [severity of test, measured with a similar formula], either by adding to b, or by reducing the empirical content of h.

Contrary to Kluge's unsubstantiated assertions, justifications for weighting have nothing to do with verificationism (this accusation simply delves into scientist's attitudes, amounting to nothing more than *ad hominem* psychologism). Kluge has not shown that weighting decreases *C* or *S* (adding to *b* changes the values of the terms in the equation, possibly increasing *C*, *S* or content). Kluge has not even shown that weighting adds to *b*; in fact, one might well say the opposite: equal weighting presupposes that all the characters are equally correlated with phylogeny and that no character can be more reliable than others, while successive or implied weighting presuppose nothing regarding character reliability (e.g., with perfectly congruent data all characters receive equal weights). And, finally, even if weighting truly reduced the content of the hypothesis, this does not mean that the hypothesis will be less corroborated or less severely tested: "corroboration depends on presently available evidence, while logical improbability (content) does not" (Farris, 1995, p. 114).

Somewhat less philosophically, Kluge (1997b, p. 355) also argues that one problem with homoplasy weighting is that it does not specify the mechanism that might make more homoplasious characters less reliable: "Fallibility does not specify why something is fallible … self-consistency weighting presupposes a character that behaves badly in one part of the cladogram … must do so as well elsewhere in the cladogram. But why that must be so?" The lack of specification is actually one of the strengths of the method, as the mechanisms themselves are bound to remain nebulous prior to a phylogenetic analysis. Sometimes it is argued that strongly adaptive characters are less reliable, which is reasonable, as characters that can change rapidly in response to varying environments are unlikely to be well correlated with phylogeny (e.g., salt glands are surely not a strong reason to believe that all birds that came to live near sea shores are a monophyletic group). Sometimes the very opposite is argued: characters involved in important adaptations are very reliable—this is also reasonable, because some characters may be so critical to survival that changing them is very unlikely (e.g., a female mammal without mammary glands is not very likely to ever produce viable offspring). Similar arguments can and have been construed for non-adaptive characters, both for reliability and lack thereof. In practice, the only way to judge whether some evolutionary mechanism determines a good correlation with phylogenetic groupings is by testing its effects on the character in question, that is, through homoplasy. Once it is determined that a given character is poorly correlated with phylogeny, then the researcher may proceed with inquiries as to the possible causes of the lack of correlation, but the causes themselves need not be known before the phylogenetic analysis. It is true that, as pointed out by Kluge, homoplasy weighting presupposes that whatever mechanism is responsible for a good or bad correlation with phylogeny, it remains relatively constant through time and across the tree. This may well not be the case, but it is wholly irrelevant to the comparison between standard homoplasy weighting and equal weights—equal weights has even stronger assumptions of uniformity, assuming that the reliability is the same over all characters, in addition to being constant through time and across the tree.

Aside from mere rhetoric, Kluge (1997b, p. 355) also resorts to arguments that superficially appear as empirically grounded:

> A further complication is the fact that the cladogram over which the homoplasy is counted is limited to the terminal taxa in the matrix, which does not take account of the character evolution which has occurred more globally or which has taken place within terminal taxa. Thus, at best, a suite of weights can constitute only a crude hypothesis on frequencies of incongruence due to supposed homoplasy and/or investigator error.

This actually has empirical consequences: if a given matrix incorrectly reflects amounts of homoplasy, the results obtained from weighted analyses of a limited sample of taxa could be biased, and those results would greatly change when adding more taxa. However, mere change is not enough: adding taxa or characters may well change the results of analyses under equal weights; what matters here is whether uniform or differential weights changes more when adding taxa or characters. Despite his "concern for evidence" (Kluge, 1989), he never bothered to examine whether any evidence supports his assertion that weighted results will be more unstable to addition of taxa. Evidence actually refutes his assertion, as shown in this paper.

The other empirical (or quasi-empirical) argument advanced by Kluge (2005, p. 27) is that, under differential weights "there is also a potential loss of information because an incongruent character state can in fact increase phylogenetic structure, e.g., a reversed state can be diagnostic of a monophyletic group (Källersjö et al., 1999)." That is, incongruent characters can in fact increase jackknife frequencies. It is surprising that Kluge (2005) cites Källersjö et al.'s results, as those results were based on resampling methods, which he rejects as unscientific (Grant and Kluge, 2003). In any event, as we show below, weighting against homoplasy increases jackknife frequencies relative to equal weights, invalidating the last of Kluge's (2005) arguments.

Methods

The molecular data sets analyzed here have 439–921 taxa (average size 587.3 taxa), and 463–5520 characters. They comprise rbcL, cytochrome *b*, protein and ribosomal sequences. The 70 morphological data sets have 50–170 taxa (average size 82.3 taxa), and 31–381 characters. The data sets used (obtained mostly from their authors) are indicated in Appendix 1.

All the analyses were done using TNT (Goloboff et al., 2003b). The scripts used (available from the first author) made the runs in parallel, using a cluster of 10 slaves plus master, all 3.0 GHz Pentium IV machines, running under Suse Linux 9.1. Justification for the different experiments and comparisons is provided in the sections describing the results; specific details on the search techniques used for each type of comparison are in Appendix 2.

Several values of the concavity constant k (which determines how strongly homoplasious characters are downweighted; see Goloboff, 1993, 1995) were used for implied weighting. For clarity (and given that no weighting strength seemed to have a significantly better performance over all data sets and for all measures), the plots for implied weighting under different weighting strengths do not distinguish between the results for different concavities.

In the comparisons of results for reduced data sets with results for complete data sets, the groups supported by the complete data set are referred to as "correct" groups. This is not intended to suggest that the groups for the complete data set are truly monophyletic (in nature), but rather to the fact that results based on an increased amount of evidence are, necessarily, to be preferred over those based on only part of the evidence.

Abbreviations

The following abbreviations are used to define stability measures: r, number of groups supported by the reduced data set; $\sim r$, number of groups supported by the entire data set and not by the reduced data set; e, number of groups supported by the entire data set; $\sim e$, number of groups supported by the reduced data set and not by the complete data set; f, number of groups in a fully resolved tree (number of taxa for the reduced data set, minus 2); d, number of groups recovered by both the reduced and the complete data set.

Caveats to interpreting changes in jackknife frequencies

Källersjö et al.'s (1999) results constitute the only empirical finding that seems to be against downweighting homoplastic characters. The argument is based on the idea that since adding third positions (highly homoplastic) to a data set comprising only first plus second positions increases the jackknife frequencies, then it follows that positions with more homoplasy may nonetheless define groups more reliably. While increased jackknife frequencies may often indicate that the added characters more reliably define groups, this needs not be so in general. It is obvious that a very conservative character will define only few groups, but every group which the character defines is highly believable. If the majority of the characters in a matrix are extremely conservative, resulting cladograms will be poorly resolved. But every one of the groups is likely to be correct, that is, it is unlikely to be falsified should we subsequently obtain a larger set of characters. If less conservative characters are added to the matrix, then a most parsimonious tree for the enlarged data set will display more groups. However, the association of one of the added (and considerably homoplastic) characters with a given group does not, in itself, reliably allow us to conclude the existence of the group. This effect seems to be precisely that seen in rbcL-2500. In other words, first plus second positions, in the case of rbcL-2500, determine few (360)[1] well-supported groups, just because they are too conservative to identify more groups. Third positions alone determine 1349 groups, and all positions together determine 1394. Of the 360 groups determined by first plus second positions, 272 groups are supported, and an additional 53 groups are uncontradicted, by third positions alone. Of those 360 groups, 317 are also well supported when all positions are considered. In other words, of the 360 groups supported by first and second position, 88% are supported by the complete data set, and 90% are compatible with the groups for third positions alone. The groups supported by first and second position therefore are few, but apparently not mistaken. It is as if we had a one-character data set, in which all vertebrates are scored for the presence or absence of mammary glands. The character would determine a single group, but that one group is (for all we know) a perfectly correct group; the fact that a single group can be determined from that character hardly means that the character is unreliable. That characters with more homoplasy are less reliable had traditionally been considered to follow from the definition of homoplasy (see Goloboff, 1993, for citations and discussion), as the tree implied by a (single) homoplastic character will be more and more incorrect to the extent that the character has more and more homoplasy. The present examples suggest that reliability can indeed be

[1]Källersjö et al. (1999) had recovered 431 groups. The present analysis used TBR-collapsing, which eliminates more ambiguous groups. The values obtained here for third positions alone, and for all positions together, are essentially identical to those obtained by Källersjö et al.

considered to follow from absence (or rarity) of homoplasy.

As an additional example of the problems that may arise when adding characters and concluding, from increased jackknife frequencies, that the added characters are more reliable, consider Fig. 1. The graphs plot the effect of adding random characters to well-structured matrices of 60 taxa (100 such matrices were simulated). Each matrix was formed by generating a random polytomous tree (with polytomies of maximum degree 7), and then creating 12 perfect characters for each of the groups in this polytomous random tree. The resulting matrix is free of homoplasy, and any of the characters unequivocally determines a "correct" group. Then, half as many character as the homoplasy-free characters were added independently to each cell of the matrix at random, so that every cell had $P_{(1)} = 0.3$ and $P_{(0)} = 0.7$. Average jackknife frequencies (standardized by reference to a perfectly resolved, perfectly supported tree) and number of nodes with jackknife frequency above 50% (divided by the number of taxa minus 2) were calculated before and after the addition of the random characters. In Fig. 1, the x-axis corresponds to values prior to the addition of the random characters, and the y-axis to values after the addition. The diagonal, that is, represents a tie: points on that diagonal had values unchanged after adding the random characters. As can be appreciated in the plots, both average jackknife frequency and number of well-supported groups are consistently increased when random characters are added.

The data sets of Fig. 1 are quite extraordinary, to be sure; the effect is produced by having numerous perfectly congruent characters determining each group while leaving large parts of the tree unresolved; the addition of half as many random characters under those circumstances will rarely suffice to make a previously supported group unsupported, simply adding some resolution to the polytomous parts of the tree. Extraordinary or not, however, the example shows that it is possible for randomly generated characters to increase jackknife frequencies and number of well-supported groups, thus indicating that caution is needed to interpret changes in those jackknife measures.

It is obvious that rbcL-2500 is not an exact equivalent of the data sets in Fig. 1. For example, the random characters of Fig. 1 determine no well-supported groups by themselves (they do so only in conjunction with the well-structured characters), while third positions alone determine many groups in rbcL-2500. The examples are alike, however, in that the well-structured data determine few groups simply because of low variability, not because of the presence of conflicting information.

The results of Fig. 1 show that changes in jackknife frequencies when deleting (or downweighting) characters may not have an unequivocal interpretation. They may or may not give an indication of the intrinsic reliability of the removed characters.

Morphological data sets, implied versus equal weights

Resolution and jackknifing

Under homoplasy weighting, the (estimated) consensus of optimal trees was much more resolved than under equal weights (Fig. 2A). This is uncontroversial; the extra resolution under implied weighting is often interpreted as the result of the increased precision of the fitting function, which makes exact ties less likely. However, assigning weights as a random function of the number of steps (see Appendix 2 for details on the weighting function) produces trees which are, on average, less well-resolved than trees under equal weights (see Fig. 2A). This suggests that the extra resolution under homoplasy weighting may be the result of more than just making exact ties less likely.

The number of strongly supported groups (i.e., with jackknife frequency above 50%) when weighting against homoplasy was clearly higher than under equal weights; when using random weights (Fig. 2B) it was not significantly different from equal weights. The average

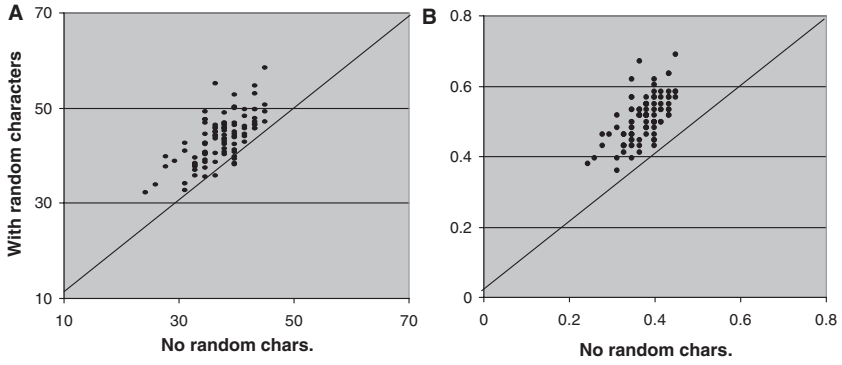

Fig. 1. Addition of random data. (A) Effect on jackknife frequencies; (B) number of nodes above 50%.

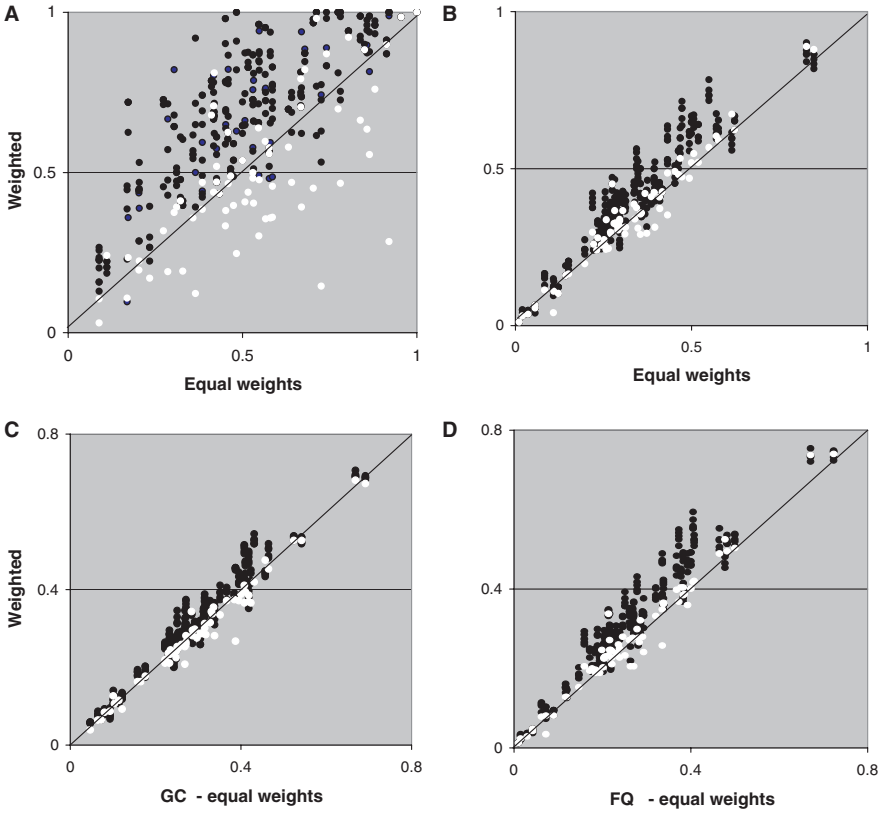

Fig. 2. (A) Resolution; (B) number of groups with jack frequency above 50%; (C) average group support, measured with GC; (D) average group support, measured with raw frequency. For $k = 5$–16 (black) and random weights (white), on morphological data sets.

group support was increased when weighting against homoplasy, regardless of whether the support was measured as frequency differences (GC statistic, of Goloboff et al., 2003a; see Fig. 2C) or raw frequencies (Fig. 2D). For GC (which Goloboff et al., 2003a; proposed as a better measure of group support), the average group support for random weights tended to be lower than for equal weights, while for raw frequencies random and equal weights produced similar results.

Stability under addition of characters/taxa

If perhaps similar in some regards (and possibly correlated in most cases), jackknifing and stability measures are not identical. Although jackknifing is sometimes justified by reference to "stability" (e.g., Siddall, 2002, p. 85; Hovenkamp, 2004), other authors (e.g., Farris et al., 1996; Farris, 2002, p. 352; Goloboff et al., 2003a, p. 326) have justified jackknifing as intended primarily as a measure of degree of support; that is, a measure of whether a hypothesis of monophyly for the group in question involves character conflict. Note that evaluations of stability to variation in the parameters of an analysis (often called "sensitivity analysis"; Wheeler, 1995; Giribet, 2003), have a different goal and will not be considered in this section.

Goloboff (1993) argued that, if weighting methods function as expected, they should produce more stable results. Studying whether or not the results produced by analyzing limited amounts of evidence remain stable to the addition of new evidence (characters, taxa) in actual taxonomic practice, would be very difficult[2]. But it is possible to have a rough estimate of that stability by eliminating part of the evidence, and comparing the results produced from analysis of the reduced data set with those produced when analyzing the complete data set. In this case, the reduced data set is the "current" data set, and the actual (complete) data set stands for the data set that the researcher might obtain in the future. Note that these statistics are relative to the results of the reduced data set, not the complete data set; in this regard, our comparisons are more appropriate than those in Goloboff (1997) or Ramírez (2003), which divided by numbers of groups for the complete data set. Average group supports (in the previous section) indicate the proportion of groups supported for the com-

[2]Such a study could be done by comparing the results produced by successive stages of a morphological data set, as a taxonomic study progresses by adding taxa and/or characters; this could be done if taxonomists routinely preserved copies of all versions of their data sets. We know of only a few colleagues who actually follow such practice.

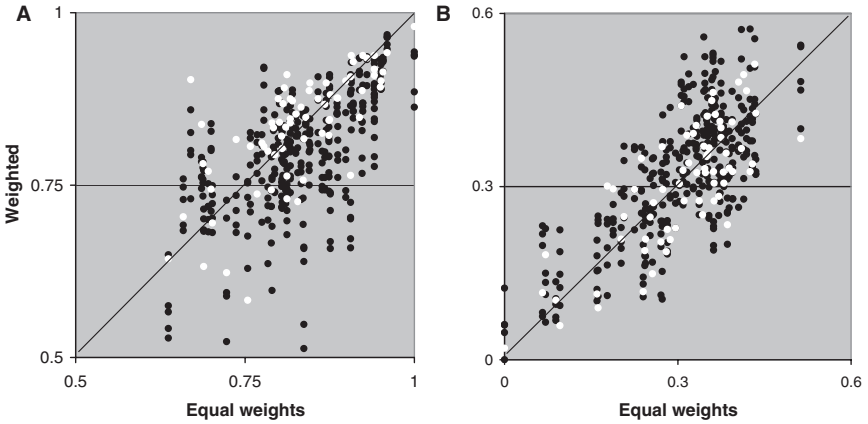

Fig. 3. Proportion of nodes recovered, when adding characters (A) or taxa (B), for $k = 5$–16 (black) and random weights (white), on morphological data sets.

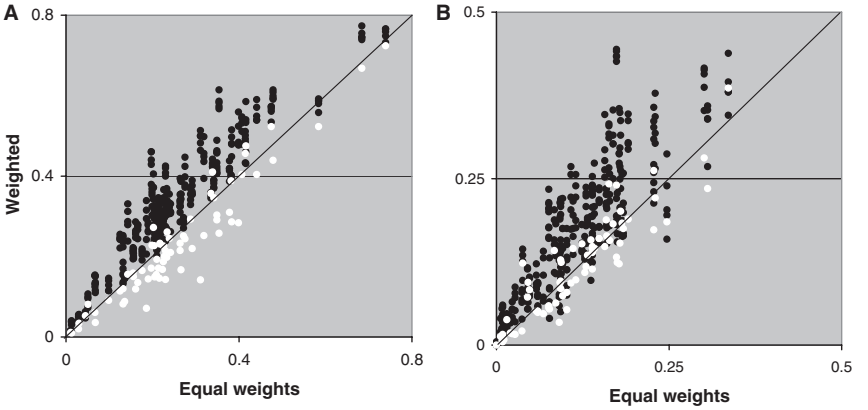

Fig. 4. Nodes recovered (relative to a fully resolved tree), when adding characters (A) or taxa (B), for $k = 5$–16 (black) and random weights (white), on morphological data sets.

plete data set that are supported by the reduced one, while stability measures indicate proportion of groups supported by the reduced data set (i.e., "now"), which will be supported when we add the remaining evidence and complete the data set (i.e., "in the future").

Of all the groups found in the (estimated) strict consensus, if weighting against homoplasy, the proportion (d/r) of groups likely to remain supported (relative to equal weights) when adding characters tends to be lower (Fig. 3A), and when adding taxa it is about the same (Fig. 3B). This measure tends to produce better results for less resolved trees; if the (reduced) data set, for example, supports only three groups, just recovering one of those three when adding taxa/characters will produce a high (0.66) proportion of groups recovered, and will be perceived as "better" than recovering, say, 10 of 20 groups. Random weights produces (according to d/r) results that (when adding characters) tend to be better than equal weights (with 24 cases below, and 33 cases above the diagonal that indicates a tie; see Fig. 3A). A measure under which random weights seems to produce better results than equal weights seems problematic[3]. Thus, a measure that does not favor under-resolved results seems desirable, and that is obtained by calculating the number of groups likely to remain supported when adding characters or taxa, relative to a fully resolved tree (d/f). Under such a measure, when adding characters or taxa, the results remain much more stable if weighting against homoplasy (Fig. 4A,B). In other words, if weighting against homoplasy, a larger number of supported groups are likely to remain supported when new characters or taxa are found in the future.

When weighting against homoplasy the trees after adding characters or taxa have, clearly, more similar topologies, than when using equal or random weights,

[3]Unless one is willing to consider that random weights truly improve results—we are not.

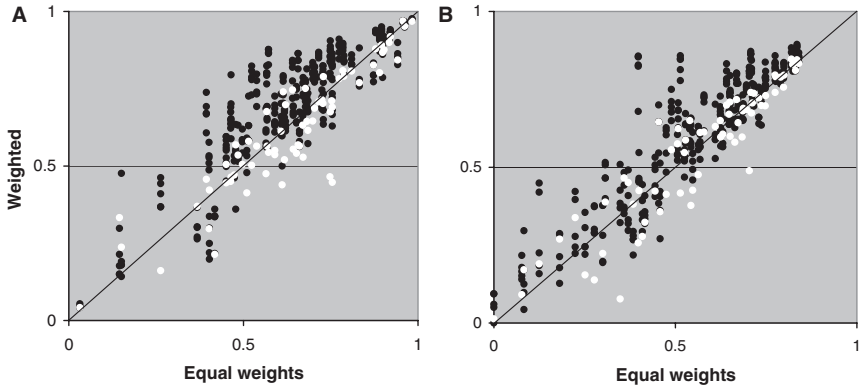

Fig. 5. Retention index of MRP of tree after addition of characters (A) or taxa (B), mapped on to tree before the addition, for $k = 5$–16 (black) and random weights (white), on morphological data sets.

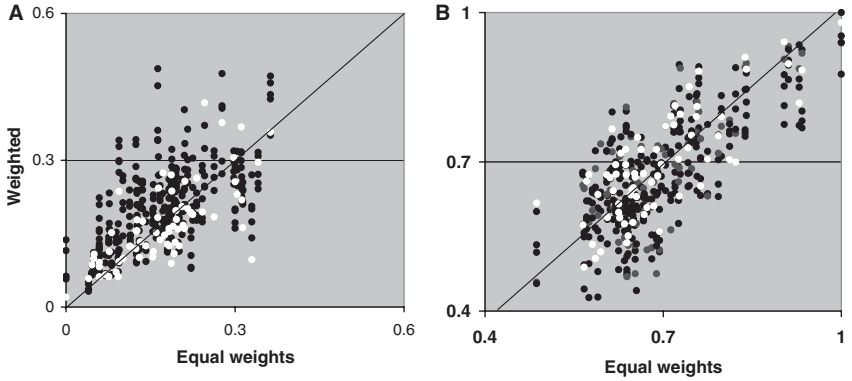

Fig. 6. Proportion of groups supported after addition of characters (A) or taxa (B) that were not supported before addition, for $k = 5$–16 (black) and random weights (white), on morphological data sets.

as measured by the retention index (Farris, 1973, 1989) of the MRP of the tree before the addition mapped on to the tree after the addition (Fig. 5A,B).

While the previous measures indicate that weighting against homoplasy increases precision (in the sense of finding more of the groups that will eventually be found as supported when adding evidence), they do not say much on the proportion of mistaken groups, that is, errors. In our examples, weighting against homoplasy is somewhat more likely to miss some of the groups that will be supported when more characters are added ($\sim r/e$; Fig. 6A), although that is in part because of the fact that so many groups will be supported (i.e., as indicated in Fig. 2A). As with the proportion of groups recovered (Fig. 3A), the proportion of groups missed when using random weights tends to be lower than for equal weights (Fig. 6A). When adding taxa, the proportion of groups supported by the complete data set that had been missed when analyzing the reduced one is (on average) similar for homoplasy weighting, equal weights and random weights. It is clear that $\sim r/e$ is influenced by resolution and will tend to produce lower rates when the tree for the complete data set is poorly resolved (i.e., not many groups can be missed then). If perhaps relevant, we do not view this measure as critical to the choice of whether and how to weight. More meaningful is the proportion of groups supported by the reduced data set that are likely to become not supported when adding taxa or characters ($\sim e/r$; Fig. 7A,B), which is similar for homoplasy weighting and equal weights. This measure is more meaningful because concluding the existence of a group that will turn out to be incorrect will often be more disturbing than not being able to conclude the existence of a group that will turn out to be correct.

Molecular data sets and homoplasy weighting

For their analysis of rbcL-2500 Källersjö et al. (1999) compared structure with third positions included or excluded, but this is an extreme form of weighting. As discussed above, the groups determined by first and second positions are, if few, mostly congruent with those

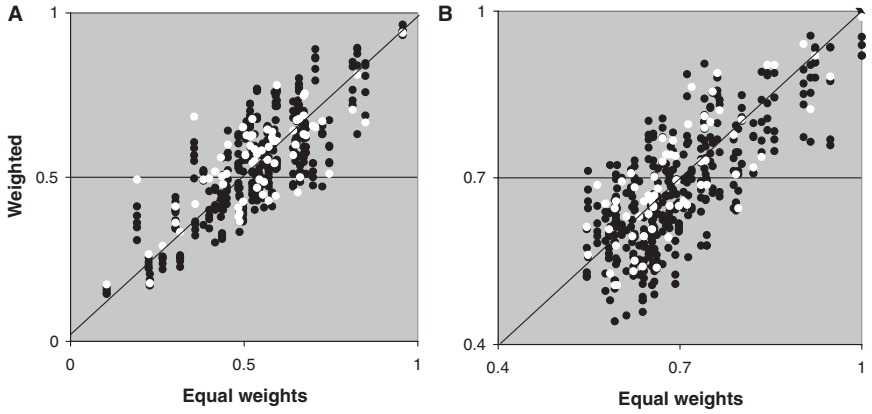

Fig. 7. Proportion of groups supported before addition of characters (A) or taxa (B) that were not supported after addition, for $k = 5$–16 (black) and random weights (white), on morphological data sets.

determined by third positions. The complementarity and lack of conflict between the results for first plus second positions on the one hand, and third positions on the other, suggest that downweighting third positions, instead of completely eliminating them, may possibly produce well resolved trees.

To test whether standard homoplasy weighting improves the results in the case of large molecular data sets, 10 molecular data sets were tested (using implied weighting under $k = 20$). The results are shown in figs 8–10 (black dots). Except for one of the measures (average jackknife frequency, Fig. 8B), in almost every case the results indicate that implied weighting produces worse results than equal weights. The proportion of groups supported by the complete data set but not the reduced ($\sim r/e$), and the proportion of groups supported by the reduced data set but not the complete ($\sim e/r$) were also higher (i.e., worse) in the case of implied weighting (data not shown).

Those results—contrary to what the complementarity in the results for third and first plus second positions suggested—show that downweighting, instead of eliminating, does not improve jackknife frequencies or measures. A more careful consideration of implications of weighting, however, shows that even a concavity of 20 produces a weighting that is, in the case of these large data sets, very far from mild. In the case of Chase et al.'s (1993) classic 500-taxon rbcL matrix, for example, there are characters that can have up to 325 extra steps. Under $k = 20$, the cost of adding the last extra step to such a character is about 280 times lower than the cost of adding an extra step to a character with no homoplasy. It is clear therefore that such a weighting strength actually comes close to complete elimination of many characters.

An alternative is to set k to a value such that, for the data set at hand, the maximum possible ratio for the implied weights cannot exceed a certain value, thus determining the admitted range of weights. For concavity k, adding the first extra step to a character without homoplasy has a cost, v_1, equal to

$$v = k/(k+1)$$

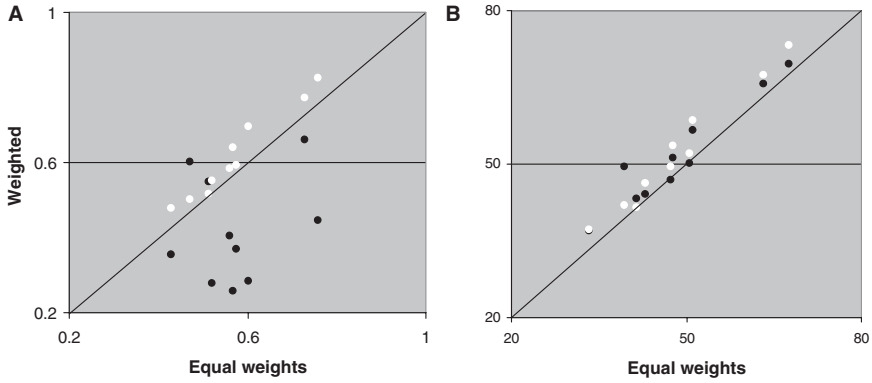

Fig. 8. (A) Nodes with jackknife frequency above 50%; (B) average jackknife frequency, for implied weights ($k = 20$, black) and range weighting (range = 15, white), versus equal weights, on molecular data sets.

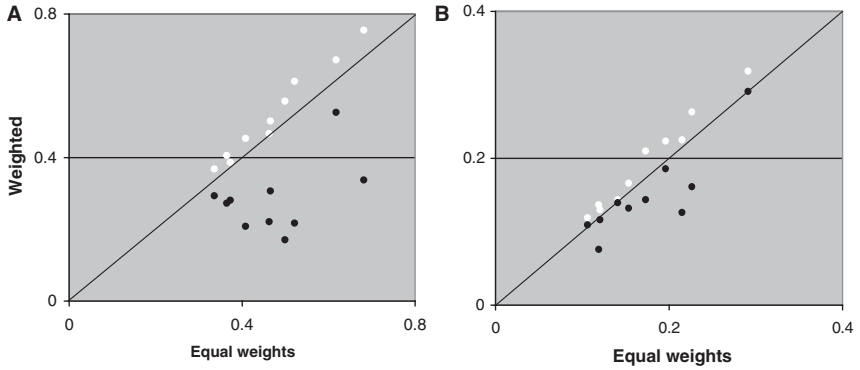

Fig. 9. Nodes recovered (relative to a fully resolved tree), when adding characters (A) or taxa (B), for $k = 20$ (black) and range = 15 (white), on molecular data sets.

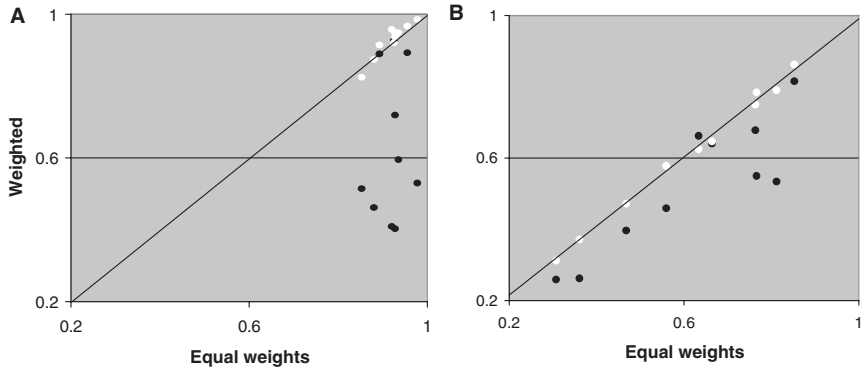

Fig. 10. Retention index of MRP of tree after addition of characters (A) or taxa (B), mapped on to tree before the addition, for implied weighting ($k = 20$, black) and range weighting (range = 15, white), on molecular data sets.

For a maximum possible number of extra steps equal to g, adding the last step of homoplasy has cost v_g equal to

$$v_g = (k/k + g) - (k/k + g - 1)$$

The value of k which will produce the desired range of weights, N, is the one[4] for which $v = N \times v_g$. In our examples, we used $N = 15$. Evidently, the actual value of k varied from data set to data set, with the constant parameter being the permissible range of weights.

When downweighting homoplastic characters in this way, the results are clearly improved over those for equal weights (Figs 8–10, white dots). The proportion of groups supported by the complete data set but not the reduced ($\sim r/e$), and the proportion of groups supported by the reduced data set but not the complete ($\sim e/r$) were also lower (i.e., better) in the case of range weighting.

These results suggest that the poor results of implied weighting in large molecular data sets are caused by weighting too strongly against homoplasy, rather than by weighting against homoplasy per se. Goloboff (1993, p. 89) had noted that "whether the same concavity should be used for different numbers of taxa, remains to be investigated"; the present results and discussion suggest a way to determine at least reasonable ranges of concavity to use for a given data set.

Discussion and conclusions

Our results show that changes in jackknife frequencies when removing or adding groups of characters may not tell much about the reliability of those characters. Källersjö et al.'s (1999) results do not provide unequivocal evidence against downweighting characters with homoplasy. If, despite this, Källersjö et al.'s results continue being cited as evidence against homoplasy weighting (as we suspect they will), then the results for morphological data sets become relevant: downweighting characters according to their homoplasy produces more strongly supported groups, and more stable results. Therefore, invoking Källersjö et al.'s (1999)

[4]The actual formula for the equivalence is

$$k = \frac{(N - 2g - 1)^2 + 4(N - 1)(g^2 + g) - (N - 2g - 1)}{2(g - 1)}$$

results as a reason to not weight characters in the case of large molecular data sets necessarily implies that character weighting is highly desirable in the case of morphology.

Part of the reason why Källersjö et al.'s (1999) results may have been interpreted so widely as an indication that weighting methods are flawed (e.g., Miller and Hormiga, 2004; Kluge, 2005; with a few notable exceptions such as Fontal-Cazalla et al., 2002) is perhaps that those results agree, superficially at least, with a general preconception against character weighting. The degree to which preconception can influence perception is exemplified by Scott, 2005) otherwise excellent morphological analysis of ranid phylogeny. She states that she prefers the results based on equal weights, because "the position that all characters should be weighted equally in phylogenetic analysis … is the least assumption-laden approach" (Scott, 2005, p. 515). But then she finds that

> The topology obtained from the simultaneous analysis of molecular and morphological data with equal weights is … more similar to the result of the separate analysis of morphology under weak implied weighting than it is to the analysis of morphology under equal weights, and does not display the questionable relationships mentioned above for the equal weights morphology analysis (Scott, 2005, p. 516).

Surprisingly, Scott (2005, p. 516) considers that "[t]his may be interpreted as demonstrating the positive synergistic effect of analyzing both datatypes simultaneously". Scott, disapproving of weighting because it is "assumption-laden", misses the obviously complementary interpretation: that this indicates the advantages of weighting against homoplasy in morphological data sets[5].

Critics of weighting often argue that weighting is based on strong assumptions and it produces trees with decreased explanatory power. This is rarely justified explicitly; one of the exceptions is Miller and Hormiga (2004), who stated that

> [s]ince the minimization of ad hoc hypotheses is critical to the relationship between evidence and science (Farris, 1983), defense of a hypothesis requiring more than the minimum number of character state changes to explain observation must be accompanied by a compelling argument for why the parsimony criterion should be relaxed (Miller and Hormiga, 2004, p. 401).

but if all that counts in phylogenetic inference is the raw number of character state changes, then this automatically rules out any type of character except non-additive characters. Considering any character as additive is equivalent to admitting that minimization of some changes is more important. As Lipscomb (1992) has cogently argued, information on relative degrees of similarity between states, when available, can and should be used to score the corresponding character as additive (e.g., when similarity is observed to be nested). It is especially ironic that Miller and Hormiga (2004), like so many others (Turner and Zandee, 1995; Kluge, 1997b) pretend that Farris (1983) showed that character weighting leads to reduced explanatory power, when Farris has always been in favor of character weighting, both before (Farris, 1969) and after (Farris, 2001) the discovery that third positions define more groups in the rbcL-2500 matrix. Farris (1983, pp. 17–18) explicitly stated that weighting characters did not decrease explanatory power:

> In portraying weighting as an alternative to parsimony, Watrous and Wheeler apparently intended to equate the parsimony criterion with simple counting of equally weighted homoplasies. That usage reflects both a lack of familiarity with the way in which parsimony has long been used by other phylogeneticists and a misunderstanding of the nature of character weighting. … In the absence of any convincing reason for doing otherwise, the characters of a study are often treated in practice as if they all provided equally cogent evidence on phylogenetic relationship. No one supposes, however, that characters in general all deserve the same weight—that they all yield equally strong evidence. Drawing conclusions despite conflicting evidence requires that some evidence be dismissed as homoplasy. It is surely preferable to dismiss weaker evidence in deference to stronger. A decision reached by weighting characters, at any rate, can hardly rest on a basis different from parsimony. … 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' position aside, no philosopher of science would quarrel with the idea that some ad hoc hypotheses are more perturbing than others. For example, consider the case where two alternative biological theories can be rescued from falsification by just one ad hoc hypothesis each: to defend theory A it must be postulated that the watch used by the observer who produced the falsifying experiment was running 5 min late; to defend theory B, it must be postulated that the observer's watch was running 12 h late—that he mistook day for night. How could anyone claim that these two theories are equally falsified?

Equally unconvincing is the idea that differential weighting relies on stronger assumptions than analyses under equal weights (Kluge, 1997b; Miller and Hormiga, 2004; Scott, 2005; and many others). The argument somehow suggests (without being explicit) that methods such as successive or implied weighting force specific sets of differential weights on to the

[5]Giannini and Simmons (2005, p. 426) also found that "analyses under implied weights were slightly superior to equal weights with respect to the number of nodes shared with combined analyses." Unlike Scott (2005), Giannini and Simmons (2005) do interpret this situation as indicating that properly weighted analyses are preferable.

characters, while the truth is that only equal weights forces any specific set of weights: all equal. None of these authors has explained why assuming, prior to the analysis, that all the characters are indeed equally reliable, is a stronger assumption that just not assuming so. In the best possible case, one might say that both equal weighting and implied or successive weighting rely on different notions as to how homoplasy may affect reliability: either not at all, or to some degree. But the statement that homoplasy does not affect reliability at all is, indeed, a strong statement.

Our defense of homoplasy weighting must not be interpreted as a statement that it is the ideal form of weighting. Goloboff's (1997) method of self-weighted optimization [not tested here; in Goloboff's (1997) comparisons it performed similarly to implied and successive weighting] is an obvious possibility of improvement, but there are other possible lines, such as trying to use information on how distantly homoplastic changes occur to influence the cost of those changes. In the absence of better formalizations, the only thing that can be compared is the methods that have been actually formalized by now—and these are clearly superior to equal weights. The notion that equal weighting should be preferred because current weighting methods are less than perfect cannot be seriously defended[6], for the only comparison that is relevant to the choice between current weighting methods and equal weights is their relative performance, not the performance of as yet undiscovered methods.

Likewise, the demonstration that a properly rescaled weighting function also produces improvements for large molecular data sets need not be interpreted as an indication that implied weighting should be used routinely in the analysis of molecular data sets. Obviously, many problems that necessarily affect the analysis of sequence data make it difficult to recommend a universal application of implied weighting. For prealigned sequence data, establishing categories of characters or transformations may be preferable to evaluating every character separately (as implied weighting normally does). This does not necessarily mean the often used ts/tv ratios, but other less explored—and probably simpler—possibilities, such as collectively weighting all the sites in a region of the sequence according to the average homoplasy of all the sites in that region. Furthermore, analysis of fragments of unequal length needs to consider simultaneously indels and possibly other types of rearrangements (e.g., Sankoff, 1975; Wheeler, 1996, 2005). As a consequence, combined analyses of morphology and molecules may remain problematic, at least until a reasonable way to apply implied weighting to some parts of the data set but not others is developed[7]. This may make a universal and acritical application of combined or "total evidence" analyses (*sensu* Kluge, 1989) problematic; analyzing different types of data separately (so that the best method of analysis can be used in each case) and then combining or comparing results may be more meaningful in many cases. Of course, analyzing some data subsets separately does not mean that one is not considering all of the available evidence; it simply means that there is no metric that can be meaningfully used to produce just one combined analysis.

Our experiments used a wide range of concavities (values of k between 5 and 16 correspond to implied weights, for a character with 10 extra steps, between 0.125 and 0.387), all of which produced similar degrees of supports and estimations of stability/congruence. In other words, the results outperformed equal weights regardless of the concavity constant chosen. This, of course, does not eliminate the need to examine results under different concavities, as in a "sensitivity analysis" (Wheeler, 1995; Giribet, 2003). This also points to another notion, which seems mistaken: that the need to explore different parameters (whether ts/tv ratios or concavity) stems from being unable to determine, a priori, the exact value that the parameter should take. There may not be one unique, "true" value of k or ts/tv ratios (e.g., they may vary over the tree, and these parameters certainly are not intended to reflect exact probabilities of events in a stochastic model). The need to explore different concavities, in the case of implied weighting, is simply because concluding the existence of a group because it occurs under a given k is obviously unwarranted if the same group is absent from the results for another value of k (within a reasonable range of concavities). This reasoning also casts doubt on using a sensitivity analysis as a "metacriterion" to determine the "optimal" value of a parameter, as proposed by Ramírez (2003) for k. Regardless of the fact that some value of k may produce the highest value of jackknife frequencies (or any other measure), the monophyly of a taxonomic group unsupported in some very close value of k cannot be considered as firmly established. Thus, only those groups present in the results for all concavities explored should be used (possibly with their minimum values of Bremer or jackknife support), so that sensitivity analysis is used to produce more

[6]Especially when equal weighting shares the same imperfection: Kluge (1997b) implies that homoplasy weighting is problematic because it assumes that a character is equally reliable over all parts of the tree, but ignores the fact that equal weights assumes exactly the same.

[7]In older versions of Poy (see Wheeler et al., 2006), implied weighting for sequences was achieved by calculating the fit as a function of the number of transformations in entire fragments; this has the problem that fragments of different length may be weighted using functions of different strengths. In the current version (Poy 4.0, Varón et al., 2007), implied weighting is not implemented.

conservative conclusions (and not as a metacriterion to maximize resolution).

Acknowledgments

The authors wish to thank V. Albert, J. Farris, M. Källersjö, M. Mirande, D. Pol, M. Ramírez and C. Szumik, for comments and discussion. PAG thanks L. Taher for the derivation of the formula in footnote 4. All the authors who provided data sets (listed in Appendix 1) are gratefully thanked. The cluster on which most of the computations for the present study were carried out was funded by NSF (AToL grant 0228699 to Ward Wheeler). PAG was supported by the ANPCyT (PICT 12605 FONCYT-BID), CONICET (PIP 02567) and Consejo de Investigaciones de la Universidad Nacional de Tucumán, Argentina. JSA and DRME were supported by the grant 11020513563 Colciencias (Consejo Nacional de Ciencia y Técnica), Colombia.

References

Chase, M.W., Soltis, D.E., Olmstead, R.G., Morgan, D., Les, D.H., Mishler, B.D., Duvall, M.R., Price, R.A., Hills, H.G., Qiu, Y.L., Kron, K.A., Rettig, J.H., Conti, E., Palmer, J.D., Manhart, J.R., Sytsma, K.J., Michaels, H.J., Kress, W.J., Karol, K.G., Clark, W.D., Hedren, M., Gaut, B.S., Jansen, R.K., Kim, K.J., Wimpee, C.F., Smith, J.F., Furnier, G.R., Strauss, S.H., Xiang, Q.Y., Plunkett, G.M., Soltis, P.S., Swensen, S.M., Willimas, S.E., Gadek, P.A., Quinn, C.J., Eguiarte, L.E., Golenberg, E., Learn, G.H. Jr, Graham, S.W., Barret, S.C.H., Dayanandan, S., Albert, V.A., 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Ann. Mol. Bot. Gard. 80, 528–580.

De Laet, J., 1997. A reconsideration of three-item analysis, the use of implied weighting in cladistics, and a practical application in Gentianaceae. PhD Thesis, presented to the Katholieke Universiteit Leuven, Faculteit der Wetenschappen.

Farris, J.S., 1969. A successive approximations approach to character weighting. Syst. Zool. 18, 374–385.

Farris, J.S., 1973. On comparing the shapes of taxonomic trees. Syst. Zool. 22, 50–54.

Farris, J.S., 1983. The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (Eds.), Advances in Cladistics 2: Proceedings of the Second Meeting of the Willi Hennig Society. Columbia University Press, New York, pp. 7–36.

Farris, J.S., 1989. The retention index and the rescaled consistency index. Cladistics 5, 417–419.

Farris, J., 1995. Conjectures and refutations. Cladistics 11, 105–118.

Farris, J.S., 2001. Support weighting. Cladistics 17, 389–394.

Farris, J., 2002. RASA attributes highly significant structure to randomized data. Cladistics 18, 334–353.

Farris, J., Albert, V., Källersjö, M., Lipscomb, D., Kluge, A., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.

Fontal-Cazalla, F.M., Buffington, M.L., Nordlander, G., Liljeblad, J., RosFarré, P., NievesAldrey, J.L., PujadeVillar, J., Ronquist, F., 2002. Phylogeny of the Eucoilinae (Hymenoptera: Cynipoidea: Figitidae). Cladistics 18, 154–199.

Giannini, N., Simmons, N., 2005. Conflict and congruence in a combined DNA morphology analysis of megachiropteran bat relationships (Mammalia: Chiroptera: Pteropodidae). Cladistics 21, 411–437.

Giribet, G., 2003. Stability in phylogenetic formulations and its relationship to nodal support. Syst. Biol. 52, 554–564.

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

Goloboff, P.A., 1995. Parsimony and weighting: a reply to Turner and Zandee. Cladistics 11, 91–104.

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

Goloboff, P., Farris, J.S., 2001. Methods for quickconsensus estimation. Cladistics 17, S26–S34.

Goloboff, P., Pol, D., 2005. Parsimony and bayesian phylogenetics. In: Albert, V. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 148–159.

Goloboff, P., Farris, J., Källersjö, M., Oxelmann, B., Ramírez, M., Szumik, C., 2003a. Improvements to resampling measures of group support. Cladistics 19, 324–332.

Goloboff, P., Farris, J., Nixon, K., 2003b. T.N.T. Tree Analysis Using New Technology. Program and documentation, available at http://www.zmuc.dk/public/phylogeny/tnt.

Grant, T., Kluge, A., 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics 19, 379–418.

Grant, T., Kluge, A., 2005. Stability, sensitivity, science and heurism. Cladistics 21, 597–604.

Hovenkamp, P., 2004. In support of support. In: Stevenson, D. (Ed.), Abstracts of the 22nd Annual Meeting of the Willi Hennig Society. Cladistics 20, 76–100.

Källersjö, M., Albert, V.A., Farris, J.S., 1999. Homoplasy increases phylogenetic structure. Cladistics 15, 91–93.

Kluge, A., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38, 7–25.

Kluge, A., 1997a. Testability and the refutation and corroboration of cladistic hypotheses. Cladistics 13, 81–96.

Kluge, A., 1997b. Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic systematics. Zool. Scr. 26, 349–360.

Kluge, A., 2005. What is the rationale for "Ockham's Razor" (a.k.a. parsimony) in phylogenetic inference? In: Albert, V. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 15–42.

Lipscomb, D., 1992. Parsimony, homology, and the analysis of multistate characters. Cladistics 8, 45–65.

Miller, J., Hormiga, G., 2004. Clade stability and the addition of taxa: a case study from erigonine spiders (Araneae: Linyphiidae, Erigoninae). Cladistics 20, 385–442.

Nixon, K.C., 1999. The parsimony ratchet a new method for rapid parsimony analysis. Cladistics 15, 407–414.

Ramírez, M., 2003. The spider subfamily Amaurobioidinae (Araneae, Anyphaenidae): a phylogenetic revision at the generic level. Bull. Am. Mus. Nat. Hist. 277, 1–262.

Sankoff, D.M., 1975. Minimal mutation trees of sequences. SIAM. J. Appl. Math, 28, 35–42.

Scott, E., 2005. A phylogeny of ranid frogs (Anura: Ranoidea: Ranidae), based on a simultaneous analysis of morphological and molecular data. Cladistics 21, 507–574.

Siddall, M., 2002. Measures of support. In: DeSalle, R., Giribet, G., Wheeler, W. (Eds.), Techniques in Molecular Systematics and Evolution. Birkhäuser-Verlag, Basel, pp. 81–101.

Turner, H., Zandee, R., 1995. The behaviour of Goloboff's tree fitness measure F. Cladistics 11, 57–72.

Varón, A., Vinh, L.S., Bomash, I., Wheeler, W.C., 2007. POY 4.0 Beta 1983. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php.

Wheeler, W.C., 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44, 321–331.
Wheeler, W.C., 1996. Optimization alignment: The end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.
Wheeler, W.C., 2005. Alignment, dynamic homology, and optimization. In: Albert, V. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Cambridge, pp. 73–80.
Wheeler, W., Aagesen, L., Arango, C., Faivovich, J., Grant, T., D'Haese, C., Janies, D., Smith, W., Varón, A., Giribet, G., 2006. Dynamic Homology and Phylogenetic Systematics: a Unified Approach Using POY. The American Museum of Natural History, New York, 365 pp.

Appendix 1: List of morphological data used

These can be found at http://www.cladistics.org/journal/data/. Because some matrices contain unpublished data, all the taxon names in the matrices have been randomized. For each data set, the number of characters and taxa are indicated first, followed by the source.

1 agonum, 138 × 150: Carabid beetles, genus *Agonum*. Liebherr & Schmidt. 2004. Dtsch. Entomol. Z. 51, 151–206.
2 amerem, 64 × 56: American Eumeninae (Vespidae). Unpublished (J. Carpenter).
3 amphi, 156 × 85: Anuran amphibians. Unpublished (a version of the matrix in A. Haas. 2003. Cladistics 19, 2389; J. Faivovich, pers. comm.).
4 anyph, 200 × 93: Anyhpaenid spiders. Ramírez, M. et al. 2004. Zootaxa 668, 18.
5 apoidea, 139 × 54: Bees and sphecid wasps (Apoidea). Melo, G. 1999. Sci. Pap. Nat. Hist. Mus. Univ. Kansas 14, 155.
6 araneo, 302 × 83: Araneoid spiders. Agnarsson, I. Inv. Syst. 2003, 17, 719–734.
7 astr, 36 × 103: *Astragalus* legumes. Camp, P., in Platnick, N. 1989. Cladistics 5, 145–161.
8 ausmat, 96 × 72: Thynnine wasps (Tiphiidae). L. Kimsey. 2000. J. Hymen. Res. 9, 18–28.
9 bats, 250 × 75: Bats. Unpublished (modified from Giannini & Simmons. 2005. Cladistics 21, 411–437; N. Giannini, pers. comm.).
10 bemby, 163 × 53: Carabids (genus *Bembidion* et al.). Maddison, D. R. 1993. Bull. Mus. Comp. Zool. 153, 143–299.
11 bertetal, 59 × 54: Tinamou species. Bertelli et al. 2002. Syst. Biol. 51, 959–979.
12 bivalvia, 183 × 76: Bivalves (Mollusca). Giribet & Wheeler. 2002. Invert. Biol. 121, 271–324.
13 bomb, 44 × 50: Bees (genus *Bombus*). Williams, P. 1994. Syst. Entomol. 19, 327–344.
14 bracon, 89 × 126: Braconid wasps. (D. Quicke, pers. comm.)
15 brochu, 164 × 62: Gavialids (Crocodilia). C. Brochu. 1997. Syst. Biol. 46, 479–522.
16 bryo, 43 × 56: Polytrichales (Bryophita). J. Hyvonen et al. 2004. Mol. Phyl. Evol. 31, 915–928.
17 camilo, 110 × 50: Scorpions (genus *Bothriurus*). Unpublished (Camilo Mattoni, Ph.D. Thesis).
18 caronieto, 110 × 64: Mayflies. Unpublished (C. Nieto, Ph.D. Thesis).
19 centip, 222 × 80: Centipedes. Edgecombe & Giribet. 2004. J. Zool. Syst. Evol. Res. 42, 89–134.
20 cephalo, 101 × 78: Cephalopods. Lindgren et al. 2004. Cladistics 20, 454–486.
21 cocos, 268 × 53: Crocodyles. A version from Pol & Apesteguia. 2005. Am. Mus. Novit. 3490, 1–38.
22 corydo, 83 × 68: Corydoradine fishes. Britto, M. 2003. Proc. Acad. Nat. Sc. Philadelphia 153, 119154.
23 cristian2, 128 × 64: *Liolaemus* lizards. Unpublished (C. Abdala, Ph.D. Thesis).
24 crust, 352 × 68: Crustaceans and other arthropods. Giribet et al. 2005. Crustacean Issues 16, 307–352.
25 das, 60 × 85: *Dasybasis* (Tabanidae). Unpublished (González et al.).
26 dinos, 276 × 50: Prosauropods. Unpublished (D. Pol, Ph.D. Thesis).
27 diony, 381 × 145: Dionychan spiders. Unpublished (M. Ramírez).
28 dro, 217 × 159: Drosophilid flies. Grimaldi, 1990. Bull. Am. Mus. Nat. Hist. 197, 1–139.
29 embia, 186 × 157: Embiopterans. Unpublished (C. Szumik).
30 entelo, 247 × 55: Entelogyne spiders. Griswold et al. 2005. Proc. Calif. Acad. Sci. 4th Ser. 56, Suppl. II:1–324.
31 erigo, 176 × 82: Erigonid spiders. Miller, J. & G. Hormiga. 2004. Cladistics 20, 385–442.
32 ethe, 51 × 58: Iguanid lizards. Etheridge & de Queiroz. 1988. Stanford University Press.
33 fannia, 157 × 83: Muscoid flies (genus *Fannia*). Unpublished (C. Domínguez, Ph.D. Thesis).
34 firefly, 100 × 96: Branham, M. A. & J. W. Wenzel. 2003. Cladistics 19, 1–22.
35 gig_nw, 71 × 66: Genus *Gidantodax* (Simuliidae, Diptera). Pinto Sanchez et al. 2005. Insect Syst. Evol. 36, 219–240.
36 gui_m, 76 × 55: Tingid heteropterans. Guilbert, E. 2001. Zool. Scr. 30, 313–324.
37 holmorph, 176 × 85: Holometabolous insects. Whiting, M. et al. 1997. Syst. Biol. 46, 1–68.
38 hymen, 169 × 77: Hymenopteran families. Ronquist, F. et al. 1999. Zool. Scr. 28, 13–50.
39 kearney, 162 × 80: Amphisbaenians. Kearney, M. 2003. Herpet. Monogr. 17, 1–74.
40 liebherr, 206 × 170: Platynine carabids. Liebherr & Zimmerman. 1998. Syst. Entomol. 23, 137–172.

41 lobo3, 45 × 76: *Liolaemus* lizards. Unpublished (F. Lobo).
42 lorica, 215 × 128: Loricariid fishes. Armbruster, J. 2004. Zool. J. Linn. Soc. 141, 1–80.
43 ltbees, 131 × 83: Longtongued bees. RoigAlsina, A. & Michener, C.D. 1993. Univ. Kansas Sci. Bull. 55, 123–162.
44 lucena, 119 × 66 Characid fishes, unpublished (Carlos A.S. Lucena, Ph.D. Thesis).
45 lucho3, 139 × 83: *Trichomycterus* fishes and related genera. Unpublished (Luis Fernandez, pers. comm.).
46 lycos, 147 × 98: Ctenid spiders and relatives. Unpublished (an earlier version of Silva, D. 2003. Bull. Am. Mus. Nat. Hist. 274, 1–86).
47 mammals, 319 × 90: Tetrapods. Ruta et al. 2003. Biol. Rev. 78, 251–345.
48 marcos2, 370 × 91: Characid fishes. Unpublished (M. Mirande).
49 mischo, 60 × 73: *Mischocyttarus* wasps. Unpublished (O. Silveira, Ph.D. Thesis).
50 mitt, 159 × 78: Chrysomelid beetles. From Platnick, N. 1989, Cladistics 5, 145–161.
51 molina, 123 × 73: Leptohyphid mayflies. Unpublished (Molineri, Ph.D. Thesis, an earlier version of Molineri, C. 2006. Syst. Entomol. 31, 711).
52 morph, 252 × 117: Hexapod orders. Wheeler, W. et al. 2001. Cladistics 17, 113–169.
53 nixseed, 103 × 49: Seed plants. Nixon et al. 1994. Ann. Mo. Bot. Gard. 81, 484–533.
54 norell, 222 × 56: Troodontid dinosaurs. Xu & Norell. 2004. Nature 431, 838–841.
55 nsfmorph, 31 × 51: *Polistes* wasps. Unpublished (Pickett et al.).
56 odonata, 132 × 121: Dragonflies. An enlarged version of Rehn, A. 2003. Syst. Entomol.
57 parambly, 132 × 92: *Paramblynotus* wasps. Liu, Z. et al. in press, Bull. Am. Mus. Nat. Hist.
58 pilo, 149 × 113: Pilophorine hemipterans: Schuh, R. 1991. Cladistics 7, 157–189.
59 po, 95 × 68: Polistine wasps: Arevalo, E. et al. 2004. BioMed Central Evol. Biol. 4, 8.
60 prendi, 115 × 71: Scorpion genera: Unpublished (L. Prendini, with duplicates removed).
61 pulawski, 74 × 135: Species of *Tachysphex* (Sphecidae). Unpublished (W. Pulawski, with duplicates removed).
62 realdata, 124 × 90: Vespid wasps. Unpublished (Carpenter et al.).
63 ropa, 95 × 106: *Ropalidia* wasps. Unpublished (Kojima and Carpenter).
64 sch, 75 × 76: Phylinae bugs (Hemiptera). Schuh, R. 1984. Bull. Am. Mus. Nat. Hist. 177, 1–476.
65 tab_m, 96 × 65: Tabanids (Diptera). Unpublished (Coscarón and Miranda-Esquivel.).
66 tenu, 262 × 56: Tenuipalpid mites. QuirozGonzales (Ph.D. Thesis), in Platnick, N. 1989. Cladistics 5, 145–161.
67 tetrao, 219 × 58: Tetraodontiform fishes. Santini & Tyler. 1999. Am. Zool. 39, 10.
68 total, 104 × 84: Nemesiid spiders. Goloboff, P. 1985. Bull. Am. Mus. Nat. Hist. 224, 1–189.
69 virg7, 93 × 75: Lizards (muscles). Unpublished (V. Abdala).
70 west, 73 × 66: Legumes. Crisp & Weston. 1987. Adv. Legume Syst., Part 3, R. Bot. Gard. Kew, table 4.

Appendix 2: Methods of analysis for the comparisons carried out

Morphological data sets

Equal weights and all the types of weighting (implied weighting with different concavities, weights increasing with homoplasy, and weights following a random trajectory as homoplasy increases) were analyzed identically for all the comparisons.

Weighting functions. Implied weighting used different concavities of the standard function implemented in TNT (see Goloboff, 1993; documentation of TNT). The random weighting used a fitting function such that the cost of adding a step to a character with a given number of extra steps was a random number in the range 0.02–2. The functions that increased with homoplasy were similar, except that the cost c_x of adding a step to a character with x extra steps was defined as $c_{x-1} + (1/2.x)$, for each possible number of extra steps (i.e., a function that monotonically increases the weight with homoplasy, but less and less as homoplasy is larger).

Jackknifing. The jackknifing used symmetric resampling; Goloboff et al. (2003a) demonstrated that deletion-only jackknife with a probability of elimination different from 0.5 produces distortion when characters have different weights or transformation costs. Each resampled data set was therefore analyzed by means of symmetric resampling, with p(del) = p(up) = 0.33. A total of 100 replications of resampling was done for each data set; every resampled data set was analyzed by means of two random addition sequence Wagner trees (RAS) plus TBR and six cycles of ratchet (Nixon, 1999, as modified in TNT), collapsing the trees on equally optimal TBR swappings (see Goloboff and Farris, 2001).

Stability comparisons. To estimate the stability under addition of characters, the results obtained when eliminating characters with probability 0.33 were compared with the results obtained for the entire data set. Average values for 25 replications are reported in all

cases. Every reduced data set was analyzed by building 12 RAS trees, subjecting each to TBR, sectorial search (default parameters), and 15 iterations of tree-drifting; the trees were collapsed on equally optimal TBR swappings. To avoid bias due to unequal frequency of groups in optimal and quasi-optimal trees (see discussion in Goloboff and Farris, 2001, and examples in Goloboff and Pol, 2005), the best trees for each addition sequence were retained even if worse than the trees for other addition sequences. These results were compared with the results for the complete data set, which were calculated with a quick-consensus estimation obtained by consensing the endpoints of 15 independent replications (each of which used three RAS, TBR, sectorial search, and 15 iterations of tree-drifting; branches were collapsed when minimum possible length was zero).

To estimate the stability under addition of taxa, 10% of the taxa were eliminated at random, 25 times. The consensus for each reduced data set was estimated, and compared with the consensus including all the taxa (pruned to have the same taxon composition as the reduced data set). This required that the consensus be re-estimated for each set of taxa to be eliminated, and thus the consensus estimation was done (for both the reduced and complete data sets) with less exhaustive algorithms than in the case of character stability, using simply 15 RAS plus TBR (saving a single tree, but collapsing on equally optimal TBR rearrangements).

Molecular data sets

Jackknifing. Resampling was identical to that for equal weights and non-iterative weights, except that each resampled data set was analyzed by means of a single RAS plus TBR (collapsing the trees on equally optimal TBR rearrangements).

Stability comparisons. The stability under character addition (estimated by deleting characters with $P = 0.33$) analyzed each reduced data set with the same search algorithms as for the morphological data sets (but, due to time constraints, using only five cycles of tree-drifting instead of 15, and collapsing branches when minimum length was zero). The results for the complete data set were calculated with a quick-consensus estimation obtained by consensing the endpoints of 15 independent replications (each of which used a single RAS, TBR, sectorial search, and five iterations of tree-drifting; branches were collapsed when minimum possible length was zero).

The stability to taxon addition was estimated identically to that for morphological data sets.