WILEY-BLACKWELL

Cladistics 31 (2015) 550–567

Cladistics

10.1111/cla.12098

Parsimony analysis of unaligned sequence data: maximization of homology and minimization of homoplasy, not minimization of operationally defined total cost or minimization of equally weighted transformations

Jan De Laet*

Göteborgs Botaniska Trädgård, Carl Skottsbergs Gata 22A, SE-413 19, Göteborg, Sweden

Accepted 21 August 2014

Abstract

Wheeler (2012) stated that minimization of *ad hoc* hypotheses as emphasized by Farris (1983) always leads to a preference for trivial optimizations when analysing unaligned sequence data, leaving no basis for tree choice. That is not correct. Farris's framework can be expressed as maximization of homology, a formulation that has been used to overcome the problems with inapplicables (it leads to the notion of subcharacters as a quantity to be co-minimized in parsimony analysis) and that is known not to lead to a preference for trivial optimizations when analysing unaligned sequence data. Maximization of homology, in turn, can be formulated as a minimization of *ad hoc* hypotheses of homoplasy in the sense of Farris, as shown here. These issues are not just theoretical but have empirical relevance. It is therefore also discussed how maximization of homology can be approximated under various weighting schemes in heuristic tree alignment programs, such as POY, that do not take into account subcharacters. Empirical analyses that use the so-called 3221 cost set (gap opening cost three, transversion and transition costs two, and gap extension cost one), the cost set that is known to be an optimal approximation under equally weighted homology in POY, are briefly reviewed. From a theoretical point of view, maximization of homology provides the general framework to understand such cost sets in terms that are biologically relevant and meaningful. Whether or not embedded in a sensitivity analysis, this is not the case for minimization of a cost that is defined in operational terms only. Neither is it the case for minimization of equally weighted transformations, a known problem that is not addressed by Kluge and Grant's (2006) proposal to invoke the anti-superfluity principle as a rationale for this minimization.

Introduction

Wheeler (2012) recently stated that minimization of *ad hoc* hypotheses as emphasized by Farris (1983) always leads to a preference for trivial optimizations when analysing unaligned sequence data. In this, trivial optimizations are optimizations with trivial implied alignments. Trivial alignments, in turn, are alignments that are obtained by simply juxtaposing all observed sequences, as for example in Fig. 5c. Pointing out that this leaves no basis for tree choice, Wheeler jumped to the conclusion that parsimony must signify minimization of total cost or steps instead.

Wheeler (2012, his Fig. 1; see Fig. 7 here) discussed an example that he presented as a further elaboration of an earlier example of Kluge and Grant (2006, their Fig. 1; see also Fig. 1 of Grant and Kluge, 2009, and Fig. 6 here). Similarly to Wheeler, they took their example to mean that trivial alignments are optimal according to Farris' (1983) justification of parsimony in terms of minimizing *ad hoc* hypotheses (Kluge and Grant, 2006, p. 277). They argued that an alternative parsimony rationale, in terms of the anti-superfluity principle, provides sufficient basis to conclude that explanatory power in unweighted parsimony is maximized when equally

*Corresponding author:
E-mail address: jan.de.laet@anagallis.be

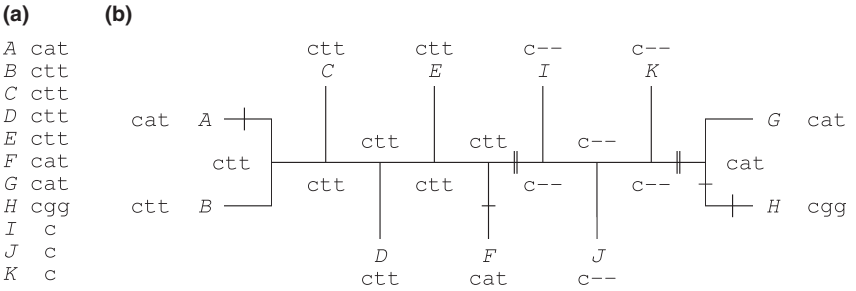

Fig. 1. A set of observed sequences (a) with a tree alignment (b), illustrating the difference between subsequence and compositional homology. Single bars across branches indicate substitutions, double bars indel events. See text for further explanation.

weighted transformations are minimized (Kluge and Grant, 2006, p. 284). Frost et al. (2001, p. 354) had earlier argued for that same minimization directly on grounds of descriptive efficiency and explanatory power. This minimization, however, suffers from a problem that seems to be insurmountable in the context of analyses of unaligned sequence data (De Laet, 2005, pp. 111–114). As discussed here, invoking the anti-superfluity principle does not address let alone solve that problem. Throughout, my discussion of Kluge and Grant (2006) applies equally well to Grant and Kluge (2009). For brevity, hereafter I refer only to Kluge and Grant (2006).

The examples of Kluge and Grant (2006) and Wheeler (2012) are remarkably similar to one of the examples that I discussed in an exploration of Farris' (1983) conceptual framework beyond the realm of independent single-column characters (De Laet, 2005, p. 111, Fig. 6.13; see Fig. 5 here). There, I argued that Farris' framework can be formulated as a maximization of homology, a point of view that provides the keys to solve the problems with inapplicables (Maddison, 1993) and to extend parsimony to the analysis of unaligned sequence data. That example was included specifically to illustrate that maximization of homology does not lead to a preference for trivial alignments when analysing unaligned sequences. Kluge and Grant (2006) and Wheeler (2012) do not seem to have been aware that this example and its discussion are at odds with their analyses of trivial alignments. The problems with their points of view are best exposed by explicitly formulating maximization of homology as a minimization of *ad hoc* hypotheses in the sense of Farris (1983), as I will do here.

Kluge and Grant (2006, p. 277) correctly pointed out that differences in justifications of parsimony—effectively leading to different numerical criteria—have empirical significance, making this an issue of general interest and practical relevance. I will therefore also provide some general background on maximization of homology, recap and further discuss how it can be best approximated in POY (Wheeler et al., 2003; Varón et al., 2010, 2013), and briefly review how the approximation under equally weighted homology has been performing with empirical data.

The discussion in this paper relies heavily on the notion of a tree alignment (Sankoff, 1975; Sankoff and Cedergren, 1983). A tree alignment for a set of observed sequences on a given tree—a frame sequence in the terminology of Sankoff (1975)—can be thought of as that tree in which the nodes are labelled with the rows of a multiple alignment that includes the observed sequences (at the leaves) as well as reconstructed sequences (at the inner nodes). The main focus of this paper is on different criteria than can be used to evaluate given tree alignments and on how such criteria make sense from a biological point of view, not on search algorithms given a criterion. Algorithms such as optimization alignment or direct optimization (Wheeler, 1996) and iterative pass optimization (Wheeler, 2003b), both available in POY, are best seen as heuristic approximations of optimal tree alignments (De Laet, 2005, p. 105; see also, e.g., Varón et al., 2008; Varón and Wheeler, 2012). They can be used with different criteria to evaluate tree alignments and so do not require special consideration from this point of view.

When discussing homology in sequence data, it is useful to distinguish between subsequence homology on the one hand and base-to-base or compositional homology on the other (De Laet, 2005, p. 106), two components of sequence homology that cannot be reduced to one another. Subsequence homology refers to statements of homology about presence or absence of entire subsequences, compositional homology to homology of bases in homologous positions of homologous subsequences. In a given tree alignment, whether or not a shared absence[1] or presence of a subsequence in two nodes is homologous, or whether or not a shared identical base at an identical position in a homologous

[1]This does not lead to the almost infinite regress that Platnick (2013) imagined: absence of a particular subsequence needs only to be ascertained and can only be considered evidence down (up) to the level where the context for its insertion arose (got lost). Below (above) that level, its absence or presence are inapplicable. A discussion of the status of its shared absence in, say, Platnick's desk and my desk is therefore as devoid of meaning as Platnick's (2013, p. 10) discussion of the status of the shared absence of spinnerets in his desk and in scorpions.

subsequence is homologous, is determined by checking all nodes on the path through the tree that connects the two nodes that are involved: only when all these nodes have the same condition as the end nodes of that path can a hypothesis of homology be made.

As an example, the tree alignment of Fig. 1b postulates two indel events, both involving the subsequence of length two that is found at positions two and three of that given tree alignment. As a result, that subsequence is homologous among terminals A–F on the one hand and between G and H on the other. It is not homologous between those two groups of terminals. Absence of that subsequence, in turn, is homologous among terminals I–K. In general, indel events in different parts of a tree can affect partially overlapping subsequences, and subsequence homology is best thought of in the negative and across single branches with indels: every indel on a branch implies a case of subsequence non-homology across that branch compared with the situation where the indel would not have occurred. Either a subsequence that was present before got lost or a subsequence that was not present before was gained.

When looking at a given position of a tree alignment, the indels that involve that position set off distinct regions of applicability for that position in the tree. In the example of Fig. 1b, positions 2 and 3 of the tree alignment are applicable in the region of the tree that connects terminals A–F and in the region of the tree that connects terminals G and H. They are not applicable outside these regions. Compositional or base-to-base homology is restricted to within such regions of applicability. Given the tree alignment, the *a*'s in the second position of A and in the second position of G, for example, cannot be homologous because they are in different regions of applicability for that position: the path through the tree that connects leaf nodes A and G contains inner nodes that do not have that position. This makes these two bases not comparable in the sense of De Laet (2005, p. 107).

Within a region of applicability, observed bases at a given position are comparable. Such comparable bases are homologous when they are identical and when all internal nodes that connect the leaf nodes that are involved also have that same base at that position. As an example, the *t* in the third position of A and the *t* in the third position of F are homologous because all intervening nodes have a *t* at that position as well. The comparable *a*'s in the second position of these terminals, on the other hand, are not homologous because there are intervening nodes that do not have an *a* there. In an extreme case, subsequences can be homologous as subsequences and yet harbour no compositional homology at all. In the example, this is the case for the subsequence at positions 2 and 3 in terminals G and H.

Examples in the figures are presented as unrooted. They can be rooted by selecting one or more appropriate outgroups. Gap and indel terminology is as in De Laet (2005, p. 98). As an example, a sequence as *g a - - - t t g c* has one gap or indel, of length 3, that extends over three unit gaps. The cost of a gap of length n is the gap opening cost + $(n - 1)$ times the gap extension cost. When gap opening cost and gap extension cost are equal, the term unit gap cost refers to either. This differs from usage as for example in Varón et al. (2010, 2013). Following existing usage (e.g. Schwikowski and Vingron, 1997; Wheeler, 2003a), I refer to a multiple alignment for the observed sequences that is obtained by removing the reconstructed sequences from a tree alignment as an implied alignment.

Maximization of homology and inapplicables

In parsimony analysis, the problems with missing characters or inapplicables (Maddison, 1993) can be overcome by maximizing the amount of similarity that can be interpreted as homology (De Laet, 2005). In the case of inapplicables that arise in the analysis of unaligned sequence data that are thought to be hierarchically related through indel events and base substitutions, a computationally harder and more general problem than what Maddison had in mind, equally weighed homology on a given tree is maximized with tree alignments that simultaneously minimize the total number of indels, substitutions, and subcharacters (De Laet, 2005, pp. 105–108). Trees that are minimal according to this criterion are trees that maximize the amount of similarity in observed sequences that can be interpreted as homology (De Laet, 2005, p. 108). Such trees, as hypotheses of genealogy, can be seen as hypotheses of maximum explanatory power.

The number of subcharacters that appears in this optimization, a new parameter both in the context of parsimony analysis and in the context of tree alignments, provides the link with inapplicables: for each position in a tree alignment, it is the number of distinct groups of comparable bases (De Laet, 2005, p. 107; see Fig. 3 for some examples). This amounts to the number of regions in the tree where that position is applicable. When a given position of a given indel in a given tree alignment has more than one subcharacter, a tree alignment that is equivalent in terms of homology can be obtained by replacing each position of that indel with multiple positions that each have one or more subcharacters, and such that the overall number of indels does not increase (see, for example, Figs. 5c and 5d). The resulting tree alignment has more positions, but the number and identity of the subcharacters as well as the numbers of indels and substitutions remain the same. The order of the groups of positions that replace the original positions has no meaning.

With sequence data, an indel by necessity comes with one or more positions that are inapplicable in some terminals.[2] So in a sequence data set with no inapplicable characters there are by definition no indel events, and each position of the tree alignment has exactly one subcharacter or region of applicability on every possible tree. In addition, each such character can in principle be scored for every terminal, and, to the degree that there is no missing information, has so been scored. In terms of tree alignments, this would be a tree alignment in which all observed and reconstructed sequences have the same length. This, in turn, amounts to a multiple alignment without indels. Typical examples would be most *rbc*L datasets of the first 1428 positions of that gene (see Chase et al., 1993, p. 531), each position specifying a positional character.

With such a dataset, minimization of the sum of indels, subcharacters, and substitutions reduces to minimization of substitutions because there are no indels and because the single subcharacter of each position on any tree has no effect on the optimization. This is numerically equivalent to a standard parsimony analysis of independent single-column characters, which directly demonstrates that minimization of the sum of indels, subcharacters, and substitutions, to maximize homology, is indeed a proper generalization of parsimony beyond independent single-column characters, and not some other criterion. As discussed below, the equivalence extends in general to a rationale in terms of minimization of *ad hoc* hypotheses of homoplasy in the sense of Farris (1983).

Maddison (1993) examined the problems with inapplicables as they classically arise with morphological data and with aligned nucleotide and protein sequence data. It is useful to distinguish between morphological data on the one hand and sequence data on the other, however. The reason is that, in general, alignment and tree evaluation cannot be properly separated with sequence data (De Laet, 2005, pp. 99–105[3]), which is precisely why this a harder problem than inapplicables with (most) morphological data[4] and why the general solution for inapplicables with such data involves tree alignments rather than static prior alignments. Maddison (1993, p. 580) suggested development of algorithms that would keep track of interactions among characters and that would restrict counting of steps in some characters to within regions of the tree where they are applicable. This is different from maximization of homology, which in the case of morphological data reduces to keeping track of interactions among characters and simultaneously minimizing the number of subcharacters, the number of gains and losses of morphological structures (the morphological counterpart of indel events), and the number of transformations among states of coded aspects of these morphological structures (the morphological counterpart of substitutions). This is computationally less complex than the case of sequence data (see De Laet, 2005, pp. 110–111) because there are more prior constraints on transformation series.

The difference between the approach that Maddison suggested and maximization of homology is illustrated in Fig. 2, using a subset of three characters and four terminals from Friedemann et al.'s (2014) dataset on Acercaria (Hexapoda). The first retained character, character 19, describes absence or presence of a tegula at the base of the forewings. The two other characters code two aspects of tegulae when present: size and shape (character 20) and attachment to the body wall (character 21). Terminals in which no tegula is present at the base of the forewings are scored as inapplicable for these two characters. That these comparative data are coded in this way conveys the idea that tegulae should be considered homologous even when they differ in further described aspects. If, for example, there would be prior grounds to reject homology of small tegulae and tegulae that are enlarged, then the coding that Friedemann et al. (2014) used is inadequate. Under this alternative prior assumption, small tegulae and enlarged tegulae should be coded in two different absence/presence characters, each with their own subordinate character that describes attachment to the body wall.

All terminals in this reduced data set have forewings, so character 19 is applicable throughout and as a result it functions as a single subcharacter on any possible tree. Steps in that character are gains and/or losses of a tegula at the base of the forewings. Steps in characters 20 and 21 are transformations between the states of the coded aspects of these tegulae. Keeping track of interactions among characters

[2]This is different from morphological data because losses and/or gains of morphological features do not necessarily imply presence of inapplicable characters: even if some feature is absent in some terminals and so coded in an absence/presence character, as long as that feature has no further characters there are no inapplicables.

[3]Simmons et al. (2011, p. 413) correctly pointed out that the conclusion that I drew from the example of my Fig. 6.8 (De Laet, 2005, p. 103) was worded too strongly: even if similarity alignment does not properly deal with the local symmetries in the sequences of that example (it assigns different scores to pairs of multiple alignments that are identical up to that symmetry), it does allow to find both optimal trees because of additional (pairs of) optimizations that I did not consider. But it suffices to change the second sequence for the outgroup in that example from *tcca* into *ccca* to rectify that. Doing so, only the first alignment of Simmons et al.'s Fig. 3b is optimal under similarity alignment, and the example so modified is sufficient to demonstrate that this alignment approach can indeed lead to rejection of a tree that the data at hand cannot logically distinguish from a tree that it accepts.

[4]An exception would be morphological data that can be conceptualized as a sequence. An example can be found in Agolin and D'Haese (2009).

(a)

```
c19: tegulae of the forewing
     (0) present
     (1) absent
c20: size and shape of tegulae
     (0) small
     (1) enlarged, with broad extension encircling entire margin
c21: attachment of tegulae to body wall
     (0) narrow
     (1) broad
```

	c19	c20	c21
F (Frankliniella, Thysanoptera)	0	0	1
H (Hackeriella, Hemiptera, Coleorrhyncha)	1	–	–
C (Ceratocombus, Hemiptera, Heteroptera)	1	–	–
D (Dictyophora, Hemiptera, Auchenorrhyncha)	0	1	0

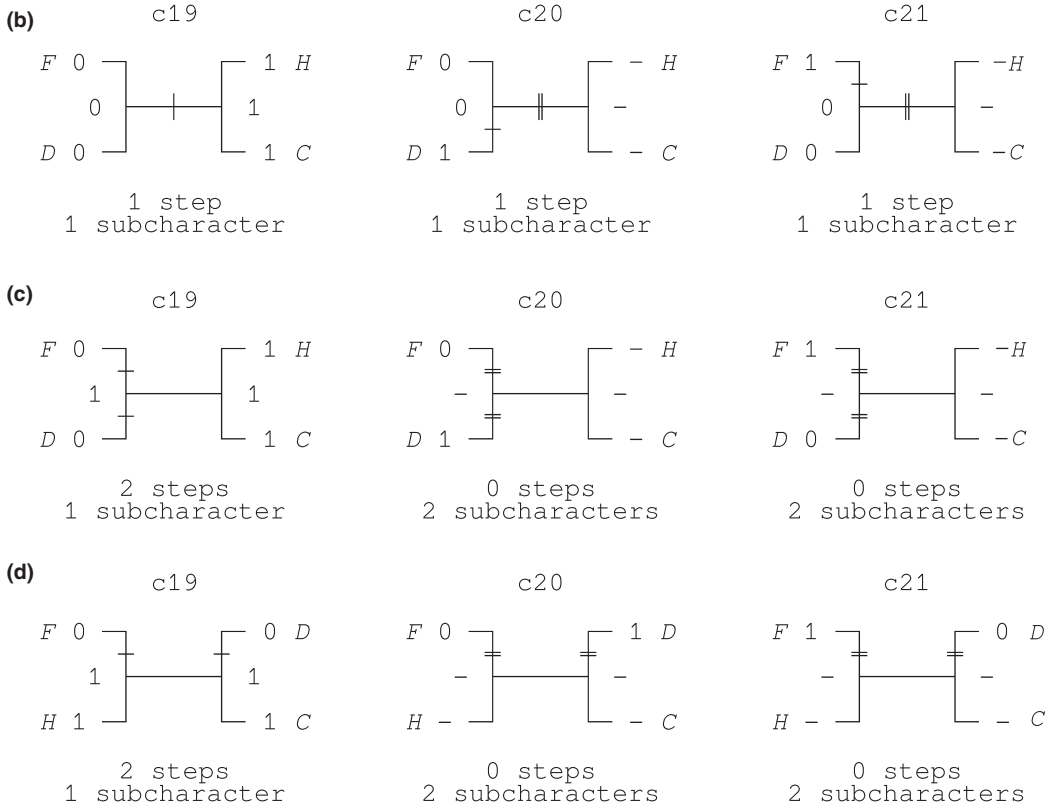

Fig. 2. Friedemann et al.'s (2014) tegula characters for a subset of four terminals (a), two explanations of those characters on the tree that has a split between the terminals with a tegula at the base of the forewing and those without (b, c), and one explanation on an alternative tree (d). A dash indicates inapplicability of a character in a terminal or at a tree node. Single bars across branches indicate steps within subcharacters (regions of applicability for a character), double bars boundaries of subcharacters. See text for further explanation.

boils down to the requirement that the explanations of the observed states on a tree should be internally consistent.[5] If, for example, an explanation of those characters on a tree requires that an internal node of that tree is optimized with absence of tegulae (character 19), then that same node on that same tree in that same explanation cannot at the same time be optimized as having an enlarged tegula with a broad extension encircling the entire margin (character 20).

When simultaneously minimizing gains and losses (steps in character 20), transformations (steps in characters 20 and 21), and subcharacters, the explanation of Fig. 2b is optimal by one (six versus seven in Figs. 2c and 2d). That explanation on that tree allows both

[5]When analysing unaligned sequence data this requirement of internal consistency is implicitly dealt with by using tree alignments (De Laet, 2005, p. 106).

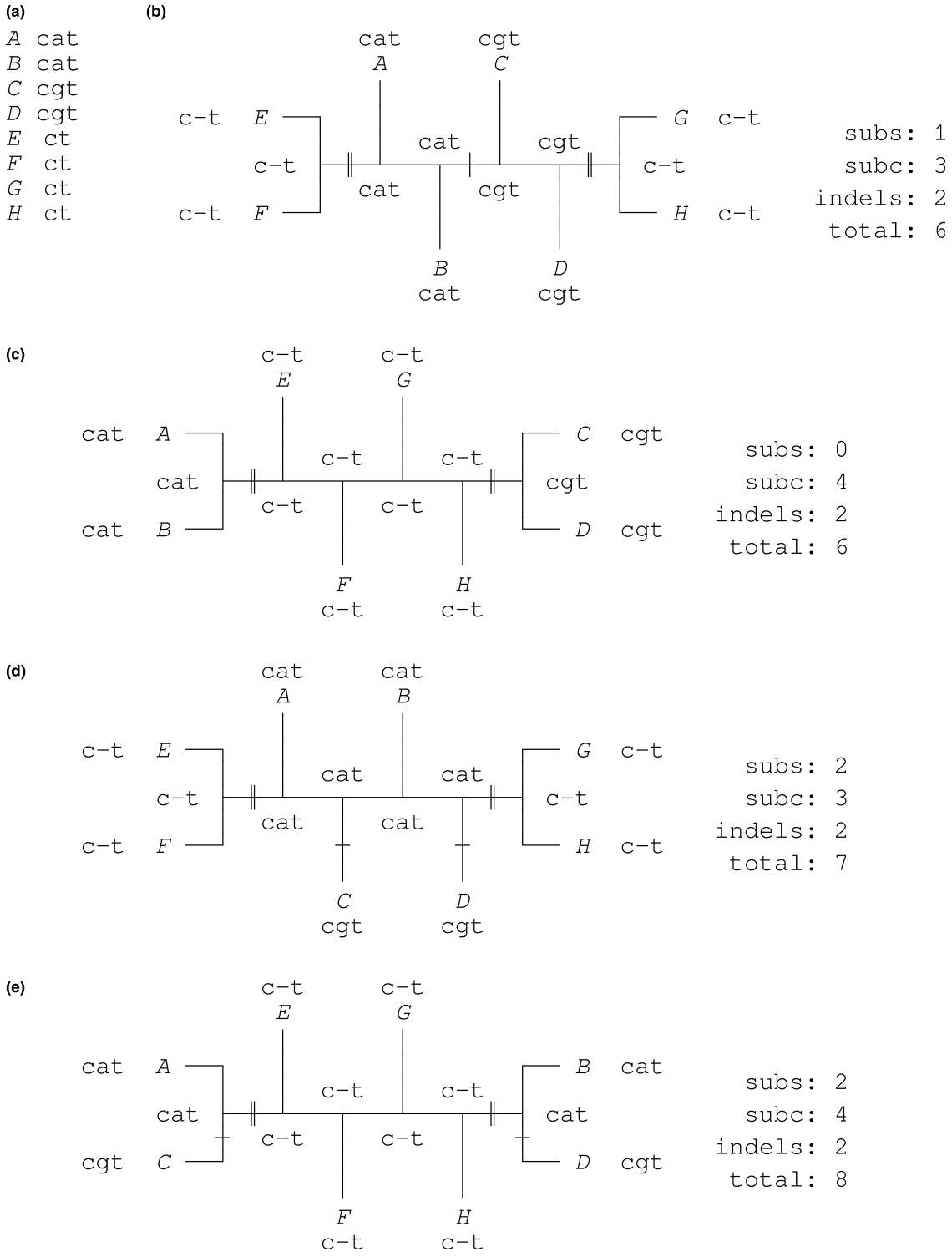

Fig. 3. A sequence character (a) and four tree alignments with two indels (b–e). The tree alignments differ only in the number of subcharacters (regions of applicability for observed residues) and steps within subcharacters for the second position of the alignment. Subs, subc, and indels are total numbers of substitutions, subcharacters, and indel events, respectively. Single bars across branches indicate substitutions, double bars indel events. See text for further explanation.

the shared absence and the shared presence of a tegula in the terminals to be explained by inheritance and common descent. The alternative explanation on that same tree that is presented in Fig. 2c only allows the shared absence to be so explained. In this alternative explanation, all steps and all boundaries of regions of applicability are on terminal branches, leaving the internal branch unsupported. As a result, the two other unrooted trees for four terminals allow for an explanation with the same number of steps and subcharacters and with the same amount of explained similarity, as illustrated for one of those trees in Fig. 2d. The numerical difference between the optimal score of Fig. 2b and the suboptimal scores of Figs. 2c and 2d—one—exactly measures their difference in independent statements of observed similarity that can be explained by common descent and inheritance.

When ignoring subcharacters and minimizing just steps in characters in regions of the tree where they are applicable, the explanations of Figs. 2c and 2d are optimal by one (two versus three in Fig. 2b). How this shift in preference comes about can be understood by comparing Figs. 2b and 2c, two alternative explanations on the same tree. With one step in character 19, the observed shared presence of tegulae in Fig. 2b is homologous, and characters 20 and 21 each have one region of applicability. In that region of applicability, both require one transformation. If, on the other hand, it is assumed that presence of tegulae on that tree is not homologous (Fig. 2c), character 19 requires two steps, resulting in two regions of applicability for each of characters 20 and 21. Within those regions, no steps are required, resulting in a decrease of total steps by one compared with the explanation of Fig. 2b. The net result is that, even if the tree allows for a homoplasy-free interpretation that tegulae are homologous, their presence is considered non-homologous under this criterion, and just minimizing steps within regions of applicability leads to a conclusion that contradicts the intended meaning of the coded comparative data. This is not the case when counting total homology in a logically correct way, which is achieved by simultaneous minimization of steps (gains/losses as well as transformations) and subcharacters. As will be clear, a similar distinction lies at the heart of Kluge and Grant's (2006) and Wheeler's (2012) mistaken view on the optimality of trivial alignments.

Such morphological character suites with inapplicables are not uncommon. Recently published examples in this journal include the characters of the embolic membrane where applicable in the spider subfamily Mynogleninae and relatives (Araneae: Linyphiidae; Frick and Scharff, 2014; their characters 64–74); of the ventromedian carina on segment V where applicable in Hormuridae and relatives (Scorpiones: Scorpionoidea, Monod and Prendini, 2014; their characters 109–111); of the manidibular exopod and of the maxillular exopod where applicable in Cambrian pancrustacean larval fossils (Wolfe and Hegna, 2014; their characters 36–40 and 48–52); and of the lateral organs where applicable in Protodrilidae and relatives (Annelida; Martínez et al., 2014; their characters 25–29). In the current absence of a program that maximizes homology for such morphological data,[6] it remains an open question if or to what degree this approach may affect phylogenetic inference.

Ad hoc hypotheses of homoplasy

A feature that is observed to be shared by organisms is so either because it can be explained by inheritance from a common ancestor—homology—or because it is a homoplasy (Farris, 1983, p. 18). So, given a tree with optimized characters, a shared feature that cannot be explained as a homology on that tree constitutes a hypothesis of homoplasy relative to that tree. If no external evidence can be found that supports this hypothesis of homoplasy, the only indication for homoplasy is in the structure of the tree itself, and the hypothesis of homoplasy is required to defend that tree as a genealogical hypothesis. Being required but having no supporting evidence of its own, it is *ad hoc* (Farris, 1983, p. 10). As Farris (2008, p. 826) pointed out,

> *Ad hoc* hypotheses of homoplasy, then, correspond to observed similarities that are explained neither by inheritance from a common ancestor, nor, so far as is known, by anything else. They could simply be called unexplained similarities, and indeed this would often be clearer, although understanding "*ad hoc* hypotheses of homoplasy" is still necessary when discussing earlier literature.

Consider an unordered character i with ns_i states and $nobs_i$ scored observations, and call nt_{ij} the number of terminals that have been assigned state j of character i (ordered characters can be dealt with by decomposition into binary additive characters). In this way, the sum of nt_{ij} over all states j equals $nobs_i$. For each state j of character i, there are $nt_{ij} - 1$ independent shared observed similarities. Summed over all states j, this amounts to $nobs_i - ns_i$ independent shared similarities in character i, a tree-independent value. When the character has no homoplasy on a given tree, that tree allows to explain all $nobs_i - ns_i$ similarities as homology, and it requires the minimal amount of change or steps for that character. With every

[6] A program for tree searches under this criterion that can handle nested levels of absences and presences of structures and substructures is under development (see De Laet, 2013). This includes the restricted case of sequences with multiple alignments that do not have partially overlapping gaps.

additional step beyond this number comes a loss of an independent shared feature that can be explained as homology, an *ad hoc* hypothesis of homoplasy on the tree involved. The amount of homoplasy in such a character on that tree is then measured directly by the number of such extra steps.

But a dataset of unaligned observed sequences has no predefined positional characters and no predefined coded similarities at that level. The coded similarity is at a higher level: the hypothesis that the sequences at hand are orthologues that are hierarchically related through two kinds of transformations, indels and substitutions. Such a dataset can be viewed as a single character, a sequence character, that can be optimized on a tree as a tree alignment. In general, two tree alignments for a sequence character have different implied alignments. In terms of parsimony analysis of independent single-column characters, this amounts to comparing two different data sets. Datasets, moreover, that may differ in the prior amounts of shared similarities in base matches. But at the level of the sequence character as a whole the underlying data are the same observations that are coded and conceptualized in the same way: sequences that can transform into each other through indels and substitutions.[7] That two tree alignments—optimizations of a sequence character on a tree—may come with different implied alignments is then no more surprising than that different optimizations of single-column characters may come with different implied transformations.

Take, for example, an unordered multistate character such as a single position of a regular multiple alignment in a position that has no gaps. Prior to the analysis it is not specified if, say, observed residues *a* are transformations of, say, residues *g* or of residues *t*. Such hypotheses are left open to optimization, and once most parsimonious trees have been obtained, the implied transformations are read from those trees. These implied transformations, moreover, may differ among different most parsimonious trees. The case of sequence characters is not qualitatively different, it just operates at a more general level where indel events are taken into account as well. As a result, positional correspondences within the putatitive orthologues are left open to optimization. As above, implied positional correspondences may then differ among different optimal trees. There are, however, consequences for what can be considered *ad hoc* hypotheses of homoplasy in the sense of Farris (1983).

Strictly speaking, only optimal tree alignments are optimizations of a sequence character, but here I use the term optimization more loosely to apply to suboptimal tree alignments as well. A suboptimal tree alignment can be considered an optimization of a sequence character on a tree in the sense that it provides an explanation of the observed sequences that is free from internal contradictions. This ensures that expressions for the two components of sequence homology in tree alignments—subsequence and compositional homology—can be derived independently and then added to get an expression that can be used to compare levels of total homology between different tree alignments (De Laet, 2005, p. 106). As discussed below, these two components of sequence homology or explained sequence similarity have a corresponding component of sequence homoplasy or unexplained sequence similarity. As with homology, relative expressions for the two components of sequence homoplasy in tree alignments can be calculated independently and then added to get a relative measure of total homoplasy that can be used to compare different tree alignments.

First consider subsequence homology. Each indel event beyond the minimum number required by the data at hand (one less than the number of different lengths of observed sequences) can be considered an *ad hoc* hypothesis in the above sense: there is no external evidence supporting it, but it is required to defend the tree alignment as a genealogical hypothesis. As with parsimony of single column characters (Farris, 1983, p. 13), this need not imply that indel events are rare.

The simple fact that putative orthologous sequences of different lengths have been observed provides such external evidence for the minimum number of indels required. Similar external evidence may be present for additional indel events, but this does not affect minimization. As an example, consider a data set of a coding region that is fully conserved except for a missing triplet near the start in half of the observed sequences, and a missing triplet near the end in the other half. As there are no length differences, the minimum number of indels is zero. But there are two observable shifts in codon correspondences that can be taken as evidence for the hypothesis that two indels have occurred nevertheless—it is precisely these shifts that warrant the expression that triplets are missing. This is evidence that is external to any particular tree, so these two hypothesized indels are no longer *ad hoc* in the sense discussed here, even if

[7] I have previously and wrongly suggested (De Laet, 2005, p. 96) that maximization of homology might also provide sufficient basis to select among competing conceptualizations of morphological data when those different conceptualizations would involve different kinds of transformations between different kinds of conceptualized structures, as for example when interpreting the vegetative region in some species of the angiosperm genus *Utricularia* as a shoot-like leaf or as a branched stem system without leaves. This is qualitatively different from the case of sequence data: with sequence data, however (sub) sequences are aligned, they remain conceptualized as (sub)sequences that are related through indels and substitutions. Therefore, different tree alignments are indeed just different optimizations of the same character. That would not be the case with the *Utricularia* example.

beyond the absolute minimum number required. But an assessment up to a constant is sufficient for the purpose of minimization. So when comparing two tree alignments, it is the one with fewer indels that performs best for this component of homoplasy, whether or not there is external evidence for indels beyond the minimum required.

The amount of compositional homology (independent base matches) in a tree alignment is equal to the sum of the lengths of the observed sequences minus the number of substitutions within subcharacters minus the number of subcharacters (De Laet, 2005, p. 107). When comparing two tree alignments, every match of two observed bases in the first tree alignment that is not present in the second means that there are two observed residues in the raw sequences that the first tree alignment can explain as an identity through common descent and inheritance but that cannot be so explained in the second, where an independent origin of these identical bases must be postulated, either through an insertion or through a substitution. Every such case amounts to an *ad hoc* hypothesis in the sense of Farris (1983) that is required on the second tree alignment but not on the first. This goes the other way around as well. With this relative formulation of *ad hoc* hypotheses in pairwise comparisons, minimization can be achieved without needing an absolute count: to decide which of both tree alignments comes with the fewest, it is sufficient to know which one has fewer. This can be done by looking at their difference in substitutions and subcharacters (the observed lengths cancel out), and maximizing homology amounts to minimizing homoplasy in this component as well.

Some examples are presented in Fig. 3. The four tree alignments shown all have two indel events, so they have the same level of subsequence homoplasy. With two different observed sequence lengths, the minimum number of indels is one, so that level of homoplasy is one. In simple examples like this, without partially overlapping indel events, instances of subsequence homoplasy are easily tracked to absence/presence of specific subsequences over all observed sequences. In this case, the single instance of homoplasy in Figs. 3b and 3d is in the absence of a short subsequence of length one in sequences E–H. In Figs. 3c and 3e it is in the presence of that same subsequence in sequences A–D.

When it comes to compositional homology, the four tree alignments differ only in the second of the three positons. The optimal tree alignments of Figs. 3b and 3c allow to explain the shared presence of residue *a* in in the middle of the sequences of A and B and the shared presence of residue *g* in the middle of C and D as homology, albeit in different ways. In Fig. 3b these four residues are in a single subcharacter, requiring one substitution. In Fig. 3c they are in two subcharacters, with all observed residues *a* in the first and all observed residues *g* in the other. In both cases, the shared presences of those residues can be explained as homology, and no homoplasy is present.

The tree alignment of Fig. 3d is suboptimal by one: it cannot explain the shared occurrence of residue *g* in the middle of the sequences of C and D. This constitues an *ad hoc* hypothesis of homoplasy that is not required on the optimal tree alignments of Figs. 3b and 3c. The tree alignment of Fig. 3e, finally, is the worse explanation. Compared with Fig. 3d, it has still one more unexplained simlarity: the shared presence of residue *a* in the middle of the sequences of A and B can no longer be explained as homology. The instance of homoplasy in Fig. 3d shows up as an extra step within a subcharacter, very much as homoplasy in independent single-column characters without inapplicables. But when inapplicables are present, homoplasy can also show up as extra subcharacters when comparing two tree alignments. This is the case, for example, for one of the instances of homoplasy in Fig. 3e relative to Fig. 3b, which illustrates that counting extra steps in subcharacters indeed no longer suffices to determine *ad hoc* hypotheses of homoplasy in the sense of Farris (1983) when inapplicables are present.

This discussion also suggests an extension of the parsimony criterion to accommodate inversions and translocations. In general, all hypotheses of translocations and inversions can be considered *ad hoc* in the sense discussed here. And even if, as was the case for indels, some hypotheses of translocations and inversions could have evidence that is external to particular trees, it would not make a difference for relative optimization of homology or homoplasy. Given two tree alignments that may include inversions and/or translocations, it is straighforward to check which one performs better (an example under equal weighting is presented in Fig. 4). Allowing inversions and translocations, however, dramatically increases computational complexity of finding optimal tree alignments.

Equally weighted transformations

Before moving to a detailed analysis of trivial alignments in the next section, it is useful to have a closer look at Kluge and Grant's (2006) view of parsimony analysis first. On theoretical grounds, they preferred minimization of equally weighted evolutionary transformations, including indels, a minimization that Frost et al. (2001) had argued for before. According to Frost et al. (2001, p. 354) it "renders the highest degree of descriptive efficiency and maximizes the explanatory power of all lines of evidence (i.e. characters)". Kluge and Grant (2006, p. 282) invoked the anti-superfluity

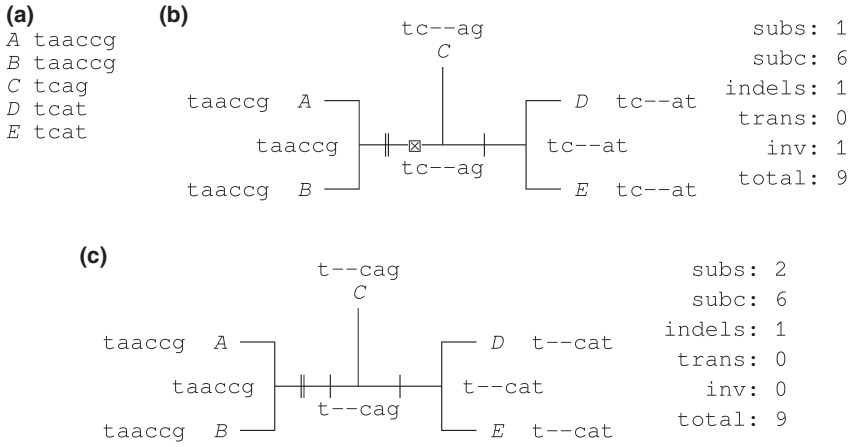

Fig. 4. A sequence character (a) and two tree alignments (b, c). Extending the analysis of unaligned sequence data to translocation and inversion events, the better of two tree alignments is the one with minimal (weighted) sum of indels, subcharacters, substitutions, translocations, and inversions. Under equal weighting, (b) and (c) explain the data equally well. Subs, subc, indels, trans, and inv are total numbers of substitutions, subcharacters, indel events, translocation events, and inversion events, respectively. Single bars across branches indicate substitutions, double bars indel events, crossed boxes inversion events.

principle as it arises from the views of Baker (2003), a principle that in their opinion provides sufficient grounds to conclude that "explanatory power is maximized by minimizing the number of transformation events required to explain the character-states of the terminal taxa as hypotheses of homology" (Kluge and Grant, 2006, p. 285).

Both also took minimization of equally weighted transformations to lead to a preference for setting the unit gap and substitution costs to one in analyses of unaligned sequence data, a step that they left unargued. But this cost set assigns a cost of n to an indel of a subsequence of n residues, the same cost as assigned to n substitutions. It follows that their preference for this cost set to minimize unweighted transformations implies the strong and unrealistic claim that insertion/deletion events affect only single bases at a time (see De Laet, 2005, pp. 111–114 for a discussion).

Under the more realistic view that indel events can affect multiple residues at once, the hypothesized indel of a subsequence of n residues across a branch of a tree, however long, would get the same cost as a single substitution. Clearly, this is not achieved by setting substitution and unit gap costs equal. In addition, such equal weighting of transformations would enable most sequences to be explained best by trivial alignments that, irrespective of the tree being considered, require just as many insertions as there are terminals. So under the realistic assumption that indels can affect multiple residues at once, Frost et al.'s (2001) and Kluge and Grant's (2006) recommendation to assign the same weight to all transformations, including indels, leads to a methodological breakdown. To all intent and purposes, it makes nucleotide-level phylogenetic analysis of unaligned sequence data impossible.

Baker (2003, p. 257) discusses the case where perturbations in the orbit of some planet can be explained by postulating a single hitherto unobserved planet. If this is the case, those perturbations can also be explained by postulating multiple unobserved planets. Baker argues that the explanation with the least postulated objects should be preferred. Applying Baker's theoretical framework to indels—as Kluge and Grant do—it would seem then that it leads to a preference for explaining an indel of length n by one indel event of length n, rather than by n indel events of length one as Kluge and Grant do. So their parsimony rationale in terms of Baker's (2003) theoretical framework is, at best, also internally inconsistent.

This is not to say that one cannot use equal substitution and unit gap costs in such analyses, and this cost set clearly does not suffer from methodological breakdown. But it implies either the unrealistic assumption that indels affect only single bases at a time, or giving up the notion that all transformations should be weighted equally. What is weighted equally instead are just the parameters being considered. Yet, equal weighting of all transformations remains explicitly being invoked as the rationale for preferring equal substitution and unit gap costs as the sole cost set used in POY analyses of empirical data. Some recent examples are Peloso et al. (2012, p. 3), Faivovich et al. (2012, p. 464), Blotto et al. (2013, p. 116), and Jungfer et al. (2013, p. 355).

Trivial alignments

Figure 5 reproduces the example that I used in 2005 to illustrate that Farris's (1983) framework, expressed

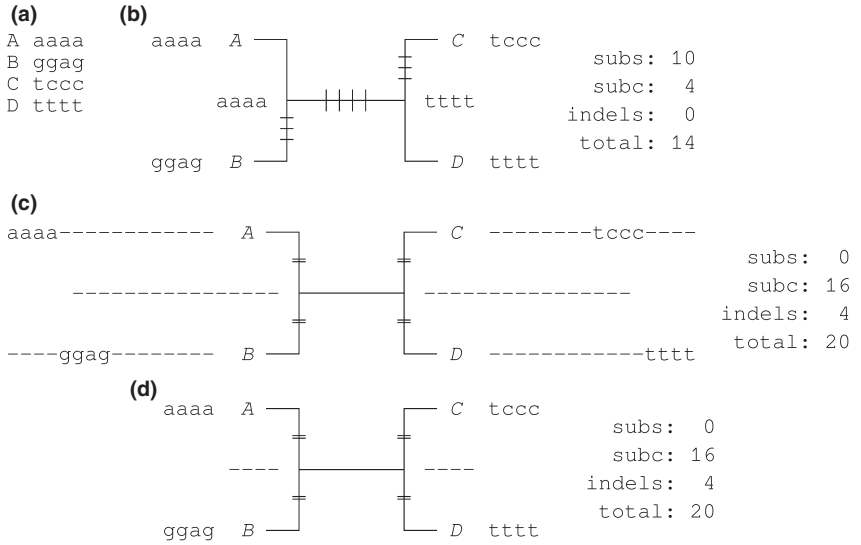

Fig. 5. A sequence character (a) and three tree alignments (b–d) on its optimal tree (A B)(C D), after De Laet (2005, p. 111, Fig. 6.13). Tree alignment (b) is optimal: it maximizes homology or, equivalently, it minimizes *ad hoc* hypotheses of homoplasy in the sense of Farris (1983). Tree alignment (c), the trivial alignment that is obtained by juxtaposing all observed sequences, is suboptimal by six units. It has as many subcharacters as there are positions in the alignment. Tree alignment (d) is an equivalent but more compact representation of the trivial alignment, obtained by putting multiple subcharacters in single columns of the tree alignment. Subs, subc, and indels are total numbers of substitutions, subcharacters, and indel events, respectively. Single bars across branches indicate substitutions, double bars indel events.

as maximization of homology, does not lead to a preference for trivial alignments (De Laet, 2005, p. 111, Fig. 6.13). The optimal tree alignment on the optimal tree (Fig. 5b) has a total of 14 subcharacters, indels, and substitutions. The trivial alignment (Fig. 5c), with a total of 20, is suboptimal by six units. This is also the number of additional *ad hoc* hypotheses of homoplasy that the trivial alignment requires relative to this optimal solution. Four of these are the indel events of the trivial alignment. The remaining two indicate that the trivial alignment can explain two fewer similarities in the base composition of the unaligned sequences than the optimal solution. As the trivial alignment does not explain any such similarity, it is the number of similarities in base composition that the optimal solution can explain as homology: the *a* in the third position of raw sequences A and B; and the *t* in the first position of raw sequences C and D. Direct minimization of *ad hoc* hypotheses in the sense of Farris (1983) clearly does not lead to a preference for the trivial alignment either.

A trivial tree alignment as in Fig. 5c can be represented in a more compact form, by packing subcharacters more tightly where possible. As an example, a most compact representation of that same trivial alignment is shown in Fig. 5d, with four subcharacters per position of the tree alignment. Numerically and biologically these two representations are equivalent, and still many other equivalent representations exist. They can be seen to be equivalent because tree alignments include the reconstructed sequences at the inner nodes. As a result, their subcharacters and the boundaries of their subcharacters are unambiguously defined, and in these two representations they are identical. Ambiguity does arise though when considering the implied alignment of just the observed sequences. At that point, it is no longer possible to check if a single column contains one or several subcharacters.

Figure 6 provides a reanalysis of the example of Kluge and Grant (2006, their Fig. 1). According to Kluge and Grant, the optimizations of Figs. 6b and 6c both require six transformations and so are equally good explanations when minimizing unweighted transformations. The trivial alignment (shown in Fig. 6d, using a compact representation), in their view, requires 25 transformations (Kluge and Grant, 2006, pp. 276–277) and is therefore highly suboptimal. But as discussed, they arrive at that conclusion by assuming that an indel of length *n* arises by *n* transformations. Under the more realistic assumption that single indel events can affect multiple residues at once, the trivial alignment only requires five transformations (five indels of a sequence of length five). So the trivial alignment of their example is optimal under their own rationale of minimizing unweighted transformations, illustrating the methodological breakdown that comes with that rationale.

Kluge and Grant's second error is in their count of *ad hoc* hypotheses of homoplasy. In their view, no such hypotheses are present in the optimization of Fig. 6c or in the trivial alignment of Fig. 6d. The only

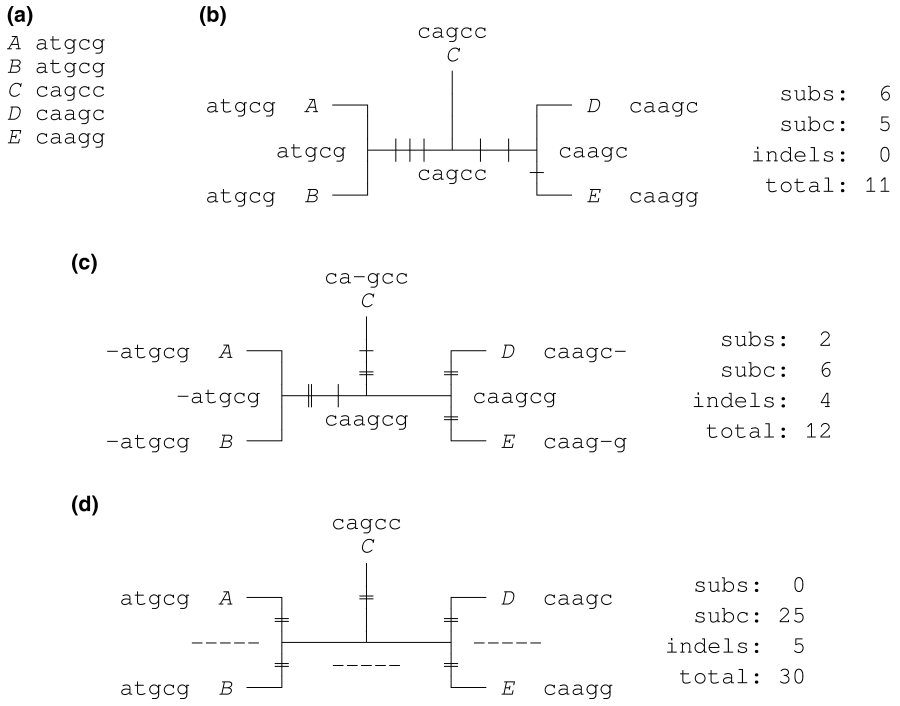

Fig. 6. A reanalysis of Kluge and Grant's (2006, their Fig. 1) example of alleged optimality of trivial alignments when minimizing *ad hoc* hypotheses of homoplasy in the sense of Farris (1983). The three optimizations of their example are shown here as three tree alignments (b–d) of a sequence character (a). The trivial alignment, in its most compact representation, is (d). Subs, subc, and indels are total numbers of substitutions, subcharacters, and indel events, respectively. Single bars across branches indicate substitutions, double bars indel events. See text for further explanation.

instance of homoplasy that they do count is an instance of an extra step within a subcharacter in the optimization of Fig. 6b, making it suboptimal by one. So, they conclude, trivial alignments are among optimal explanations when minimizing *ad hoc* hypotheses of homoplasy (Kluge and Grant, 2006, p. 277). That conclusion, however, is based on their implicit and erroneous assumption that extra steps within subcharacters is all there is to homoplasy when comparing explanations with different alignments for the same set of sequences. As seen above, this does not necessarily account for all homoplasy in such cases.

When homoplasy is properly counted, the tree alignment of Fig. 6b is optimal (11 indels, subcharacters, and substitutions) and that of Fig. 6c is suboptimal by one (12 indels, subcharacters, and substitutions), exactly the opposite of Kluge and Grant's conclusion with respect to *ad hoc* hypotheses of homoplasy in these two tree alignments. This level of suboptimality can be given a precise meaning. Compositional homology in a tree alignment can be compared using the sum of substitutions and subcharacters: the lower that number, the higher the amount of compositional homology. So the difference in compositional homology between two tree alignments is their difference in the sum of substitutions and subcharacters. On this count, Fig. 6c is optimal, and Fig. 6b is suboptimal by three. However, to get that relative gain in compositional homology, the explanation of Fig. 6c has to postulate four *ad hoc* indel events. The net result is that it is suboptimal by one. Turning to the trivial tree alignment (Fig. 6d), it is suboptimal by 19 units compared with the optimal explanation of Fig. 6b. As above, five of these are the five indel events that are required. The other 14 measure the amount of compositional homology in that optimal explanation, none of which is retained in the trivial alignment.

A reanalysis of Wheeler's example (Wheeler, 2012; his Fig. 1) is presented in Fig. 7. In this case, the putative orthologues that constitute the data are all exactly one base long, and the issue of how to deal with indels of length greater than one is evaded. But Wheeler's further analysis, just as Kluge and Grant's, tacitly assumes that extra steps within subcharacters is all there is to homoplasy when comparing different alignments for the same set of sequences. As a result, the only instance of homoplasy that he counts is in the tree alignment that has no indels (Fig. 7b), making it a worse explanation than the two other alignments of his example, including the trivial alignment. As with Kluge and Grant's example, when *ad hoc* hypotheses of homoplasy are properly counted by minimizing the sum of

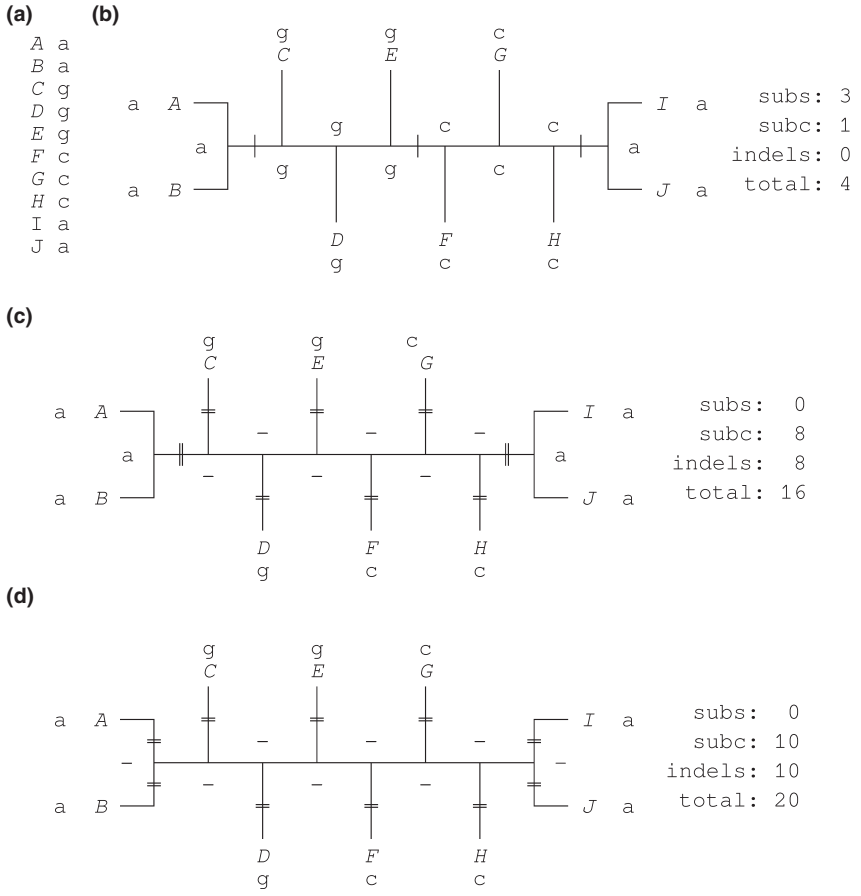

Fig. 7. A reanalysis of Wheeler's (2012, his Fig. 1) example of alleged optimality of trivial alignments when minimizing *ad hoc* hypotheses of homoplasy in the sense of Farris (1983). The three optimizations of his example are shown here as three tree alignments (b–d) of a sequence character (a). The trivial alignment, in its most compact representation, is (d). Subs, subc, and indels are total numbers of substitutions, subcharacters, and indel events, respectively. Single bars across branches indicate substitutions, double bars indel events. See text for further explanation.

subcharacters, substitutions, and indels (see Fig. 7), the explanation without indels is the best of the three explanations, the trivial solution the worst.

So it is not just that Wheeler's (2012) conclusion that parsimony must signify minimization of total cost does not follow from his premise that minimization of *ad hoc* hypotheses always leads to a preference for trivial alignments. The premise itself is not correct to start with. Wheeler (2012) did not provide other argumentation for his conclusion, so his preference for minimization of total cost is, at best, an operational guideline in search of deeper ground.

Approximations

Exact maximization of homology in an analysis of unaligned sequence data involves tree alignment algorithms that keep track of substitutions, indels, and subcharacters (De Laet, 2005, p. 108). In general, the computational complexity of finding an optimal tree alignment on a given tree (see Jiang et al., 1994) is such that heuristic approximations for tree evaluation are unavoidable in practice. Heuristic tree alignment programs such as POY (Wheeler et al., 2003; Varón et al., 2010, 2013) provide parameters for substitution costs and gap opening and extension costs, but they do not take into account subcharacters. It can be shown that, in POY, an optimal approximation for maximization of homology under equal weighting of all instances of homology is obtained by setting the gap opening cost to three, all substitution costs to two, and the gap extension cost to one (De Laet, 2005, p. 109). The approximation is optimal in the sense that it cannot be improved with algorithms that are currently available in POY. The argument that leads to this conclusion is straightforward.

Maximization of homology in tree alignments amounts to minimizing loss of subsequence homology due to indel events while simultaneously maximizing compositional homology or base matches in positions that remain homologous after these indels have been taken into account (De Laet, 2005, p. 106). As to subsequence homology, each indel event of a subsequence

across a branch of a tree constitutes a statement of non-homology compared with the situation where the indel event did not occur. Homology of subsequences is then optimized relatively and indirectly by minimizing indel events or instances of such non-homology across branches. Technically, the number of indel events in a tree alignment is obtained by setting the gap opening cost to one, the gap extension cost to zero, and all substitution costs to zero.

Next consider compositional homology. In pairwise comparisons of sequences, base matches can be maximized by setting the unit gap cost to half the substitution cost in a cost minimization (Smith et al., 1981, p. 39; equation 4b with $w_k = 0$). This result also holds for comparisons of three terminal sequences on a tree, but it breaks down for four sequences related by a (resolved) tree, as can be shown by example (De Laet, 2005, Fig. 6.11: from four sequences on, subcharacters have to be taken into account). With integer costs, the simplest way to set the unit gap cost to half the substitution cost is to set the substitution cost to two and the unit gap cost to one. In terms of gap opening and extension costs, a unit gap cost of one amounts to setting both to one. So base matches for up to three sequences related by a tree can be maximized by setting the gap opening cost and the gap extension cost to one and the substitution cost to two.

To obtain equal weighting of subsequence homology and compositional homology the loss or gain of a subsequence, however long, across a branch of a tree must be assigned the same cost as a substitution of a base.[8] So if, as in the previous paragraph, a substitution is assigned a cost of two, then indels—measured by the gap opening cost—must be assigned a cost of two as well. Adding it all up, for up to three sequences maximization of equally weighted sequence homology is then obtained exactly by setting substitution costs to two, gap opening cost to three, and gap extension cost to one in a cost mimimization.[9] For more than three sequences, this cost set can be considered a heuristic approximation for maximization of homology (subcharacters would have to be factored in to get exact results). As POY's tree evaluation heuristics are all based on simultaneous comparisons of up to three sequences at a time at most, this approximation can be considered optimal with currently available algorithms.

Edgecombe and Giribet (2006, p. 515) were the first to refer to the cost set that results from this argument as 3221, short for gap opening cost three, transversion cost two, transition cost two, and gap extension cost one. Following my 2005 proposal they included 3221, the only cost set with different values for gap opening and extension costs that they applied, in a sensitivity analysis with 16 cost sets. Using an ILD-derived index to assess congruence among partitions, it performed second best, after 111 (all parameters equal). Since then, 3221 has become the name that stuck, to the point of identification with the entire approach (see e.g. Varón et al., 2013, p. 165). This is somewhat unfortunate, because emphasizing a technical implementation—values of parameters to be set in POY to approximate the approach under equal weighting—rather than the underlying rationale comes at the risk of losing sight of that rationale. Parsimony by itself, for example, does not prescribe that all parts of the data should receive equal weight[10] (Farris, 1983), and maximization of homology in analyses of unaligned sequence data lends itself to such differential weighting (De Laet, 2005, p. 91).

As a simple example, compositional homology and subsequence homology can be weighted differentially by assigning the desired relative values to those components of the cost set that determine either. If, for example, loss of subsequence homology due to an indel is to be downweighted by half relative to loss of compositional homology due to a substitution, the resulting cost set is 2221, not 3221.

Within compositional homology, differential weighting can be achieved as follows. In a tree alignment one can assign a similarity score of two to observed and comparable identical bases (*a*-*a*, *c*-*c*, *g*-*g*, *t*-*t*), a similarity score of one to paired non-identical purines (*a*-*g*) and to paired non-identical pyrimidines (*c*-*t*), and a similarity score of zero to purine–pyrimidine pairs.[11] Using similar logic as above (the relevant equations from Smith et al., 1981 are now 6a and 6b), and giving equal weight to an instance of subsequence homology

[8]The same result is obtained when weighting sets of indel events by the amount of total loss of equally weighted homology (subsequence and compositional) that they imply relative to explanations that do not have them. Doing so amounts to equal weighting of homology itself.

[9]For these two cases—two sequences and three sequences on a star tree—Fredman (1984) provided algorithms that optimize the same criterion directly as a maximization of a similarity measure, a measure that he proposed precisely because of its biological relevance. He pointed out that his technique can be generalized to more than three sequences, but such a generalization would seem to involve a star tree, not a resolved tree. Minimization of subcharacters, indels, and substitutions in tree alignments can be seen as a generalization of Fredman's alignment approach to resolved trees with more than three sequences. Or, the other way around, as a generalization of tree alignments to accommodate Fredman's similarity metric.

[10]Sharma et al.'s (2011) recent case for exploring mixed parameter sets in sensitivity analysis can be seen to exemplify this point at the level of different sequences in analyses of unaligned sequence data.

[11]This implies a concomitant breakdown of homology: two observed and comparable bases can be non-homologous at the level of base identity but homologous as purines or as pyrimidines.

and a match between identical bases, the resulting optimal approximation in terms of parameters available in POY is cost set 3211 (gap opening cost three, transversion cost two, transition cost one, and gap extension cost one), a cost set that recently started being used successfully in sensitivity analyses using POY (e.g. Andrade et al., 2012, p. 150; Vahtera et al., 2012, p. 12). If subsequence homology were to be downweighted by half, the resulting cost set would be 2211.

Two practical considerations are worth mentioning. First, all costs in these cost sets have been multiplied by two to keep costs integer. Therefore, if a match between two identical states of a morphological character is to count as much as a match between two identical bases, morphological characters have to be assigned a prior weight of two in simultaneous analyses of morphological data and unaligned sequences with such cost sets for the sequence data. Assigning prior weights of one to the morphological data in such analyses amounts to downweighting morphology by half relative to the sequence data. An example of equal weighting, including a discussion of the issue, can be found in Giannini and Simmons (2005, p. 416).

Second, POY versions 4 (Varón et al., 2010) and 5 (Varón et al., 2013) use a different definition of gap opening cost than both POY version 3 (Wheeler et al., 2003) and De Laet (2005). In the latter two, as in this paper, the cost of the first unit gap of a gap of a given length is the opening cost, whereas in POY versions 4 and 5 it is the opening cost plus the extension cost.[12] So to have the desired effect in terms of maximization of equally weighted homology, the opening cost in POY versions 4 and 5 should be set to two, not to three. Failure to point out this difference may lead to problems of interpretation. So it is best, as for example in Giribet et al. (2010, p. 411), to be explicit about it. In this particular case, they did so by mapping my 2005 cost set to the command invocation that they used to set the parameters in POY 4.

Empirical data

Together with equal costs according to the rationale of Frost et al. (2001) and parameter sensitivity analysis as proposed by Wheeler (1995), maximization of homology following my 2005 rationale for using 3221 is among the three general approaches that POY practitioners most commonly use to select cost regimes in nucleotide-level analyses (Varón et al., 2013, p. 165). To a degree, it seems that the rationale for using that cost set is often no longer cited. Edgecombe et al. (2012, pp. 772–773), for example, included 3221 in their sensitivity analysis of six cost sets because it is among parameter sets that are "routinely used in other similar studies" (it minimized incongruence between morphology and the six genes that they studied). Aktipis and Giribet (2012, p. 17), as another example, cite Varón and Wheeler (2008) when discussing 3221 as one of the ten cost sets that they included in their sensitivity analysis (it minimized incongruence among five molecular datasets). Varón and Wheeler, however, did not discuss rationales for using cost sets, they reported a bug with gap opening and extension costs in POY version 3 and early builds of POY version 4.

Giannini and Simmons (2005) and Faivovich et al. (2005), both referring to and discussing the theoretical appeal of the underlying rationale, were the first to use 3221 in analyses of empirical data. Giannini and Simmons compared results obtained with 3221 and equal cost set 111 (transition and transversion costs one, unit gap cost one) and found that "overall congruence favored the maximization of homology by a narrow margin" (Giannini and Simmons, 2005, p. 411). Faivovich et al. (2005, p. 47) performed a two-step analysis in which they first applied equal costs, leading to a single tree. To examine the effect of gap treatment, they then submitted that tree to TBR using cost set 3221 for the sequences and prior weights two for the morphological data, thus effectively giving equal weight to all homologies. They found that the resulting tree differed in the position of two small clades (Faivovich et al., 2005, p. 49).

Other early papers that followed my 2005 proposal to use 3221 are Lindqvist et al. (2006; on theoretical grounds they only used 3221) and Edgecombe and Giribet (2006; see above). Giribet et al. (2006), an exception until recently, were the first to use 3221 without mentioning its theoretical underpinnings. Referring to papers on sensitivity analysis, their methods section just mentions that the POY analyses were done "under different analytical parameter sets" (Giribet et al., 2006, p. 7728). That 3221 was among those is only clear from the tree they included, the strict consensus of two trees obtained with 3221 (Giribet et al., 2006, p. 7724, legend of their Fig. 2).

These were also among the first papers that used the parameters for gap opening and extension costs after they first became available in POY (version 3, Wheeler et al., 2003). The very first to explore these parameters in POY were Petersen et al. (2004). Based on a sensitivity analysis in which the gap opening cost was kept equal to the substitution costs while the extension cost was successively lowered, they suggested that lower extension costs decreased incongruence and

[12]Strictly, usage as in Varón et al. (2010, 2013) applies the gap opening cost for an indel to that indel as a whole. Tying it to the first unit gap of that indel is a way to achieve this.

that a ratio of four to one might be optimal in Triticeae grasses (Petersen et al., 2004, p. 739). Aagesen (2005) and Aagesen et al. (2005a) performed sensitivity analyses that in addition also explored the effect of varying gap opening costs relative to substitution costs and extension cost. Both discussed my approach but neither included 3221 in the analyses that they performed. Other papers in that time frame that applied different gap opening and extension costs in POY, a list that is probably not exhaustive, are Aagesen et al. (2005b), Arnedo and Gillespie (2006), and Pons and Vogler (2006). All three performed extensive sensitivity analyses but none included 3221, and none of the cost sets with different gap opening and extension costs that they did include got as widely adopted later as 3221.

For a systematic and detailed assessment of results with empirical data, one would have to consider several and potentially confounding factors. For example, the early applications of the approach used POY version 3, a version that suffered from the bug that Varón and Wheeler (2008) reported. In analyses using POY versions 4 and 5, on the other hand, it is not always clear which definition of gap opening cost authors are adhering to. There are also methodological concerns to be dealt with when assessing topological congruence among partitions (see, for example, Sharma et al., 2011 and references therein for some discussion), or more generally how well one cost set behaves relative to another with empirical data. But as exemplified by the studies cited in this brief review, there is little reason to go into that level of detail at this point because the overall picture that is emerging is clear enough: when assessing congruence among partitions, cost regime 3221, the optimal approximation to maximize equally weighted homology in POY, performs in general quite well with empirical data.

Sensitivity analysis

The underlying rationale makes that cost set a natural starting point or base of comparison when exploring other cost sets, as done in sensitivity analysis. Reasons for such exploration may be several. For one, it is never harmful to find out how stable results are when parameters are slightly perturbed, especially because 3221 is only a heuristic approximation to maximize equally weighted homology. In addition, as pointed out above, parsimony by itself does not require that all parts of the data should be assigned the same weight. Based on empirical findings such as molecular mechanisms that lead to point mutations, or prior studies in related groups, one may well choose to explore the direction in which *a*-*g* and *c*-*t* matches across a branch get a non-zero similarity score, especially in coding sequences, leading to cost sets such as 3211, as discussed.

Similarly, one can explore cost sets that assign length-dependent penalties to subsequence homology losses that come with indel events. Under equal weighting, every indel event, however long the subsequence involved, is assigned the same penalty on the subsequence homology score as a single substitution. It is feasible, however, to weight the loss of subsequence homology that comes with an indel event that involves fewer positions less than the loss of subsequence homology that comes with an indel event that involves more positions, and there may be empirical grounds for doing so. One way to achieve this is to give the same weight to a single substitution and to an indel that involves just a single position, and to increase the indel penalty linearly with that same amount as the indel gets longer. Using similar logic as above, this results in best approximation cost set 322 for use in POY (unit gap cost three, transition and transversion cost two).

That logic can be applied the other way around as well: given a cost set used in POY, what is the underlying weighting scheme? As an example, consider cost set 111, the same cost for a unit gap, a transition, and a transversion. To avoid non-integer costs later on, this can also be expressed as 222. As pointed out, homologies in base matches are maximized by setting the unit gap cost to half the substitution cost. The simplest way to achieve this with integer costs is cost regime 122. The complement of 122 to achieve 222 (the part of 222 not explained by 122) is 100, or a unit gap cost equal to one and no additional costs for substitutions. Because there are no additional costs for substitutions, 100 is the part of cost regime 222 that describes penalties assigned to indels. In terms of gap opening and extension costs, a unit gap weight of one amounts to a gap extension and opening cost that are both set to one. Considering that a substitution is assigned a cost of two, this amounts to the following length-dependent gap penalty when maximizing homology: an indel that involves a single position gets half the cost of the loss of homology that comes with a single substitution; from there on, the indel cost linearly increases with that same amount as the indel gets longer. So with 111, a single substitution has the same effect on the homology score as an indel that involves two positions, or as two indels of only a single position.

Such examples illustrate how maximization of homology can be embedded in a sensitivity analysis. One does not preclude the other. Maximization of homology rather provides the general framework to understand cost sets used in sensitivity analyses in terms that are biologically relevant and meaningful: homology, base substitutions, and indel events that are not restricted to single residues at a time. No

such integrated and coherent view of the parameters that determine cost is apparent in Wheeler's (2012) operationalist stance on minimization of total cost.

Acknowledgements

My thanks to Dennis Stevenson, Jim Carpenter, Pablo Goloboff, and an anonymous reviewer for their constructive criticism during review. Their comments were of great help to improve the manuscript. This obviously does not need imply that they agree with the views expressed, nor are they to be held responsible for any errors that may remain.

References

Aagesen, L., 2005. Direct optimization, affine gap costs, and node stability. Mol. Phylogenet. Evol. 36, 641–653.

Aagesen, L., Petersen, G., Seberg, O., 2005a. Sequence length variation, indel costs, and congruence in sensitivity analysis. Cladistics 21, 15–30.

Aagesen, L., Medan, D., Kellerman, J., Hilger, H.H., 2005b. Phylogeny of the tribe Colletieae (Rhamnaceae)—a sensitivity analysis of the plastid region *trn*L-*trn*F combined with morphology. Plant Syst. Evol. 250, 197–214.

Agolin, M., D'Haese, C.A., 2009. An application of dynamic homology to morphological characters: direct optimization of setae sequences and phylogeny of the family Odontellidae (Poduromorpha, Collembola). Cladistics 25, 353–385.

Aktipis, S.W., Giribet, G., 2012. Testing relationships among the vetigastropod taxa: a molecular approach. J. Molluscan Stud. 78, 12–27.

Andrade, S., Strand, M., Schwartz, M., Chen, H., Kajihara, H., von Döhren, J., Sun, S., Junoy, J., Thiel, M., Norenburg, J.L., Turbeville, J.M., Giribet, G., Sundberg, P., 2012. Disentangling ribbon worm relationships: multi-locus analysis supports traditional classification of the phylum Nemertea. Cladistics 28, 141–159.

Arnedo, M.A., Gillespie, R.G., 2006. Species diversification patterns in the Polynesian jumping spider genus *Havaika* Prószyński, 2001 (Araneae, Salticidae). Mol. Phylogenet. Evol. 41, 472–495.

Baker, A., 2003. Quantitative parsimony and explanatory power. Br. J. Philos. Sci. 54, 245–259.

Blotto, B.L., Nuñez, J.J., Basso, N.G., Úbeda, C.A., Wheeler, W.C., Faivovich, J., 2013. Phylogenetic relationships of a Patagonian frog radiation, the *Alsodes* + *Eupsophus* clade (Anura: Alsodidae), with comments on the supposed paraphyly of *Eupsophus*. Cladistics 29, 113–131.

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., Williams, S.E., Gadek, P.A., Quinn, C.J., Eguiarte, L.E., Golenberg, E., Learn, G.H. Jr, Graham, S.W., Barrett, S.C.H., Dayanandan, S., Albert, V.A., 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Mo. Bot. Gard. 80, 528–580.

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

De Laet, J., 2013. Anagallis: a program for minimization of homoplasy in characters that are inapplicable in some terminals (Abstract). XXXIInd Meeting of the Willi Hennig Society. Rostock, Germany, 3–7 August 2013. Conference abstracts, p. 112.

Edgecombe, G.D., Giribet, G., 2006. A century later—a total evidence re-evaluation of the phylogeny of scutigeromorph centipedes (Myriapoda: Chilopoda). Invertebr. Syst. 20, 503–525.

Edgecombe, G.D., Vahtera, V., Stock, S.R., Kallonen, A., Xiao, X., Rack, A., Giribet, G., 2012. A scolopocryptopid centipede (Chilopoda: Scolopendromorpha) from Mexican amber: synchrotron microtomography and phylogenetic placement using a combined morphological and molecular data set. Zool. J. Linn. Soc. 166, 768–786.

Faivovich, J., Haddad, C.F.B., Garcia, P.C.A., Frost, D.R., Campbell, J.A., Wheeler, W.C., 2005. Systematic review of the frog family Hylidae, with special reference to Hylinae: phylogenetic analysis and taxonomic revision. Bull. Am. Mus. Nat. Hist. 294, 6–228.

Faivovich, J., Ferraro, D.P., Basso, N.G., Haddad, C.F.B., Rodrigues, M.T., Wheeler, W.C., Lavilla, E.O., 2012. A phylogenetic analysis of *Pleurodema* (Anura: Leptodactylidae: Leiuperinae) based on mitochondrial and nuclear gene sequences, with comments on the evolution of anuran foam nests. Cladistics 28, 460–482.

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

Farris, J.S., 2008. Parsimony and explanatory power. Cladistics 24, 825–847.

Fredman, M.L., 1984. Algorithms for computing evolutionary similarity measures with length independent gap penalties. Bull. Math. Biol. 46, 553–566.

Frick, H., Scharff, N., 2014. Phantoms of Gondwana?—phylogeny of the spider subfamily Mynogleninae (Araneae: Linyphiidae). Cladistics 30, 67–106.

Friedemann, K., Spangenberg, R., Yoshizawa, K., Beutel, R.G., 2014. Evolution of attachment structures in the highly diverse Acercaria (Hexapoda). Cladistics 30, 170–201.

Frost, D.R., Rodrigues, M.T., Grant, T., Titus, T.A., 2001. Phylogenetics of the lizard genus *Tropidurus* (Squamata: Tropiduridae: Tropidurinae): direct optimization, descriptive efficiency, and sensitivity analysis of congruence between molecular data and morphology. Mol. Phylogenet. Evol. 21, 352–371.

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

Giribet, G., Okusu, A., Lindgren, A.R., Huff, S.W., Schrodl, M., Nishiguchi, M.K., 2006. Evidence for a clade composed of molluscs with serially repeated structures: Monoplacophorans are related to chitons. Proc. Natl Acad. Sci. USA 103, 7723–7728.

Giribet, G., Vogt, L., González, A.P., Sharma, P., Kury, A.B., 2010. A multilocus approach to harvestman (Arachnida: Opiliones) phylogeny with emphasis on biogeography and the systematics of Laniatores. Cladistics 26, 408–437.

Grant, T., Kluge, A.G., 2009. Perspective. Parsimony, explanatory power, and dynamic homology testing. Syst. Biodivers. 7, 357–363.

Jiang, T.L., Lawler, E.L., Wang, L., 1994. Aligning sequences via an evolutionary tree: computational complexity and approximation. In Proc. 26th ACM Symposium on the Theory of Computing. ACM, New York, pp. 760–769.

Jungfer, K.-H., Faivovich, J., Padial, J.M., Castroviejo-Fisher, S., Lyra, M.L., Berneck, B.V.M., Iglesias, P.P., Kok, P.J.R., MacCulloch, R.D., Rodrigues, M.T., Verdade, V.K., Torres Gastello, C.P., Chaparro, J.C., Valdujo, P.H., Reichle, S., Moravec, J., Gvoždík, V., Gagliardi-Urrutia, G., Ernst, R., De la Riva, I., Means, D.B., Lima, A.P., Señaris, J.C., Wheeler, W.C.,

Haddad, C.F.B., 2013. Systematics of spiny-backed treefrogs (Hylidae: Osteocephalus): an Amazonian puzzle. Zoolog. Scr. 42, 351–380.
Kluge, A.G., Grant, T., 2006. From conviction to anti-superfluity: old and new justifications of parsimony in phylogenetic inference. Cladistics 22, 276–288.
Lindqvist, C., De Laet, J., Haynes, R.R., Aagesen, L., Keener, B.R., Albert, V.A., 2006. Molecular phylogenetics of an aquatic plant lineage, Potamogetonaceae. Cladistics 22, 568–588.
Maddison, W.P., 1993. Missing data versus missing characters in phylogenetic analysis. Syst. Biol. 42, 576–581.
Martínez, A., Di Domenico, M., Rouse, G.W., Worsaae, K., 2014. Phylogeny and systematics of Protodrilidae (Annelida) inferred with total evidence analyses. Cladistics doi: 10.1111/cla.12089.
Monod, L., Prendini, L., 2014. Evidence for Eurogondwana: the roles of dispersal, extinction and vicariance in the evolution and biogeography of Indo-Pacific Hormuridae (Scorpiones: Scorpionoidea). Cladistics doi: 10.1111/cla.12067.
Peloso, P.L.V., Faivovich, J., Grant, T., Gasparini, J.L., Haddad, C.F.B., 2012. An extraordinary new species of *Melanophryniscus* (Anura, Bufonidae) from southeastern Brazil. Am. Mus. Novit. 3762, 31.
Petersen, G., Seberg, O., Aagesen, L., Frederiksen, S., 2004. An empirical test of the treatment of indels during optimization alignment based on the phylogeny of the genus *Secale* (Poaceae). Mol. Phylogenet. Evol. 30, 733–742.
Platnick, N.I., 2013. Less on homology. Cladistics 29, 10–12.
Pons, J., Vogler, A.P., 2006. Size, frequency, and phylogenetic signal of multiple-residue indels in sequence alignment of introns. Cladistics 22, 144–156.
Sankoff, D., 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28, 35–42.
Sankoff, D., Cedergren, R.J., 1983. Simultaneous comparison of three or more sequences related by a tree. In: Sankoff, D., Kruskal, J. (Eds.), Time Warps, String Edits, and Macromolecules. The Theory and Practice of Sequence Comparison. CSLI Publications, Stanford, CA (1999 reprint), pp. 253–263.
Schwikowski, B., Vingron, M., 1997. The deferred path heuristic for the generalized tree alignment problem. J. Comput. Biol. 4, 415–431.
Sharma, P.P., Vahtera, V., Kawauchi, G.Y., Giribet, G., 2011. Running WILD: the case for exploring mixed parameter sets in sensitivity analysis. Cladistics 27, 538–549.
Simmons, M.P., Müller, K.F., Webb, C.T., 2011. The deterministic effects of alignment bias in phylogenetic inference. Cladistics 27, 402–416.
Smith, T.F., Waterman, M.S., Fitch, W.M., 1981. Comparative biosequence metrics. J. Mol. Evol. 18, 38–46.
Vahtera, V., Edgecombe, G.D., Giribet, G., 2012. Evolution of blindness in scolopendromorph centipedes (Chilopoda: Scolopendromorpha): insight from an expanded sampling of molecular data. Cladistics 28, 4–20.
Varón, A., Wheeler, W.C., 2008. Application note: on extension gap in POY version 3. Cladistics 24, 1070.
Varón, A., Wheeler, W.C., 2012. The tree alignment problem. BMC Bioinformatics 13, 293.
Varón, A., Wheeler, W.C., Bar-Noy, A., 2008. An efficient heuristic for the Tree Alignment problem. CUNY PhD Program in Computer Science technical report 2008015.
Varón, A., Vinh, L.S., Wheeler, W.C., 2010. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 26, 72–85.
Varón, A., Lucaroni, N., Hong, L., Wheeler, W.C., 2013. POY 5.0. Black Sabbath Development build 4818bc4e6323 R.C. Program documentation (file commands.pdf in poy_5.0.0beta-docs.zip, downloaded from research.amnh.org on 21 August 2013). American Museum of Natural History, New York, 216 pp.
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., 2003a. Implied alignment: a synapomorphy-based multiple sequence alignment method and its use in cladogram search. Cladistics 19, 261–268.
Wheeler, W.C., 2003b. Iterative pass optimization of sequence data. Cladistics 19, 254–260.
Wheeler, W.C., 2012. Trivial minimization of extra-steps under dynamic homology. Cladistics 28, 188–189.
Wheeler, W.C., Gladstein, D., De Laet, J., 2003. POY, Phylogeny Reconstruction via Optimization of DNA and Other Data, Version 3.0.11. American Museum of Natural History, New York.
Wolfe, J.M., Hegna, T.A., 2014. Testing the phylogenetic position of Cambrian pancrustacean larval fossils by coding ontogenetic stages. Cladistics 30, 366–390.