Syst. Biol. 0(0):1–18, 2021

DOI:10.1093/sysbio/syab005

Incorporating Hierarchical Characters into Phylogenetic Analysis

Melanie J. Hopkins[1] and Katherine St. John[2,3,*]

[1]*Division of Paleontology (Invertebrates), American Museum of Natural History, New York, NY, USA;* [2]*Department of Computer Science, Hunter College, City University of New York, New York, NY, USA; and* [3]*Division of Invertebrate Zoology, American Museum of Natural History, New York, NY, USA*
Correspondence to be sent to: Department of Computer Science, Hunter College, City University of New York, New York, NY, USA;
E-mail: katherine.stjohn@hunter.cuny.edu

Received 25 June 2020; reviews returned 28 December 2020; accepted 1 February 2021
Associate Editor: April Wright

Abstract.—Popular optimality criteria for phylogenetic trees focus on sequences of characters that are applicable to all the taxa. As studies grow in breadth, it can be the case that some characters are applicable for a portion of the taxa and inapplicable for others. Past work has explored the limitations of treating inapplicable characters as missing data, noting that this strategy may favor trees where internal nodes are assigned impossible states, where the arrangement of taxa within subclades is unduly influenced by variation in distant parts of the tree, and/or where taxa that otherwise share most primary characters are grouped distantly. Approaches that avoid the first two problems have recently been proposed. Here, we propose an alternative approach which avoids all three problems. We focus on data matrices that use reductive coding of traits, that is, explicitly incorporate the innate hierarchy induced by inapplicability, and as such our approach extend to hierarchical characters, in general. In the spirit of maximum parsimony, the proposed criterion seeks the phylogenetic tree with the minimal changes across any tree branch, but where changes are defined in terms of dissimilarity metrics that weigh the effects of inapplicable characters. The approach can accommodate binary, multistate, ordered, unordered, and polymorphic characters. We give a polynomial-time algorithm, inspired by Fitch's algorithm, to score trees under a family of dissimilarity metrics, and prove its correctness. We show that the resulting optimality criteria is computationally hard, by reduction to the NP-hardness of the maximum parsimony optimality criteria. We demonstrate our approach using synthetic and empirical data sets and compare the results with other recently proposed methods for choosing optimal phylogenetic trees when the data includes hierarchical characters. [Character optimization, dissimilarity metrics, hierarchical characters, inapplicable data, phylogenetic tree search.]

Phylogenetic inference based on morphology requires that systematists translate the variation observed across taxa into semi-quantitative characters. The different states of a character represent categories of biological expression of that character, and complex traits are often translated into multiple characters in order to capture all of the variation. At broad enough taxonomic scales, some complex traits may only be observed in a subset of the taxa, and thus any character describing some aspect of variation in that trait will not be applicable to all of the taxa. Characters with this type of dependence have been referred to as "inapplicable characters" (e.g., Strong and Lipscomb 1999) or "hierarchical characters" (e.g., Simmons 1993).

Several strategies exist for coding and handling such data in the tree reconstruction process. Some coding strategies have been proposed that make it possible to assign character states to all characters in a matrix. For example, each expression of a character, when present, can be described as a single character which is either present or absent (variously referred to as nominal variable coding (Pimentel and Riggins 1988); "Method D" of Pleijel (1995); "reductive coding" sensu Wilkinson (1995); and most recently as nonadditive binary coding (Strong and Lipscomb 1999)). For example, myriapods have antennae on the tritocerebral segment while other arthropods do not (some might have chelicera instead). These antennae may be filiform (more than twice the length of the head) or attenuate (less than twice the length of the head). Following the nonadditive binary coding approach, one would create two characters, the first describing the absence or presence of filiform antennae, and the second describing the absence or presence of attenuate antennae. For complex traits that require description of multiple attributes (e.g., color as well as length; or the presence of additional organs, like antennal apical cones), such binary characters could be structured to describe each individual attribute or each combination of attributes (e.g., the absence or presence of blue, filiform antennae with apical codes). One might also add an additional character describing the absence or presence of antennae, or the nature of the appendage on the tritocerebral segment (see Hawkins et al. 1997; Forey and Kitching 2000). Another strategy is to code all of the variation into one or more multistate characters, where either all expressions of an attribute are coded in one character ("Method B" of Pleijel (1995)), or where each observed combination of attributes is a state and the absence of the trait is also a state (referred to as composite coding (Maddison 1993; Wilkinson 1995), fused coding (Lee and Bryant 1999) and "Method A" of Pleijel (1995)). One might also specify that change within a multistate character must occur in a particular direction (Simmons 1993). All of these approaches have been criticized for overweighting absences, creating logical interdependencies among characters, making assumptions about the direction of evolution, or conflating homology statements (Pimentel and Riggins 1988; Maddison 1993; Simmons 1993; Wilkinson 1995; Lee and Bryant 1999; Strong and Lipscomb 1999; Fitzhugh 2006; Brazeau 2011).

A different strategy is to code the presence of the trait and the variable expression of that trait into separate characters (conventional coding (Hawkins et al. 1997), contingent coding (Forey and Kitching 2000), or reductive coding sensu Strong and Lipscomb (1999)). For example, one character would code the absence or presence of antennae on the tritocerebral segment in arthropods, while three additional characters would code for the relative length (filiform or attenuate), color, and presence of apical cones on the antennae when present. This coding strategy makes explicit homology statements separating the presence of a trait from its expression, does not overweight absences, and is hierarchical when done properly (see Hopkins and St. John (2018) for some common practices that violate this). For this coding strategy, we refer hereafter to characters that are applicable to all taxa as "primary characters," and to those that describe variation in primary characters and may not be applicable to all taxa as "secondary characters" (Kendrick 1965; McNeill 1972). Not all primary characters have secondary characters that are dependent on them. We refer to those that do as "controlling primaries"; these are usually, but not always, presence–absence characters. We note that this coding strategy accommodates "nested" characters such as tertiary characters (as well as further nesting such as quaternary and quinary characters) and can be defined by applying recursively the schema above. For matrices with tertiary characters describing variation in secondary characters, we would refer to the latter as "controlling secondary characters."

The reductive coding strategy pushes the analytical problem of having inapplicable cells in the matrix to the handling of the data during inference. Treating the inapplicable states as a separate state (absence coding of Strong and Lipscomb (1999)), not only does not solve the problem, but re-introduces the issues of overweighting absences and conflating homology statements. More commonly, inapplicable characters are simply treated as missing data during inference. This practice can skew a maximum parsimony analysis in several important ways. For example, this approach may favor phylogenetic trees where internal nodes are assigned impossible states (e.g., the absence of antennae but presence of antennal apical cones or antenna color) (Platnick et al. 1991, p. 342). Further, taxa that otherwise share most primary characters can nonetheless be grouped distantly if there are many secondary characters that they do not share (Fig. 1, right-hand tree; see also Methods section).

A more subtle problem occurs when a complex controlling primary character evolves independently in two clades, because the optimization of secondary traits describing variation in that character in one clade can influence the optimization (and therefore the "best" tree) in the other clade, even if there is no a priori reason to expect nonindependence in trait evolution across disparate clades. We refer to this scenario as Maddison's "red/blue tail" problem, referencing the example given by Maddison (1993). Contrary to some examples, Strong and Lipscomb (1999) posit that composite coding does not ameliorate this problem if absence of the character is the result of secondary loss.

Recent work has sought to address issues that arise when inapplicable characters are treated as missing data. For example, De Laet (2005) and Brazeau et al. (2019) propose introducing new characters ("sequence characters" and "regions," respectively) into the analysis to capture where secondary characters are applicable. The implementation of Brazeau et al. (2019) modifies the traditional Fitch parsimony scoring algorithm (Fitch 1971) for phylogenetic trees. Their aim was to prevent the assignment of impossible states (Platnick et al. 1991) and the influence of secondary character distribution across phylogenetically distant clades (Maddison 1993). Their approach also seeks to account for character contingency by penalizing trees by the number of "regions" where each secondary character is applicable. The choice was made to implicitly "simulate" a primary controlling character independently for each secondary character (Brazeau et al. 2019). If the secondary character is distributed homoplastically, the tree length is increased by 1 for any intervening regions where the character is inapplicable (Fig. S1 of the Supplementary material available on Dryad at http://dx.doi.org/10.5061/dryad.m37pvmd06). However, by ignoring the relationships between controlling primary characters and the secondary characters, this approach effectively overweights absences, more so for controlling primary characters with more secondary characters, and can still favor trees that group otherwise disparate clades. In this respect, it retains some of the same drawbacks to treating inapplicable characters as missing using the Fitch approach (Fig. 1; see also Fig. S1 and Table S1 of the Supplementary material available on Dryad). Another recent approach to the problem uses Markov models within a Bayesian framework to model inapplicable characters as hidden states (Tarasov 2019). This approach encourages the use of ontology languages. However, it requires a fundamentally different way of coding characters and, as such, is not included in our study, since it includes the co-option of polymorphic coding to represent variable potential expression as hidden states when the controlling primary character is absent, and states that represent combinations of secondary characters (similar to composite coding) when the controlling primary character is present.

In order to address inapplicable characters and, more generally, hierarchical characters, we propose to use the ubiquitous approach of Fitch (1971) to score candidate phylogenetic trees within a framework of parsimony, with a simple, but powerful, change. The parsimony score of a phylogenetic tree, where all nodes are labeled, is the sum of the changes across the branches (Farris 1970; Fitch 1971). Our proposed scoring criteria also seeks the phylogenetic tree that minimizes the changes across the branches, but where changes are counted via

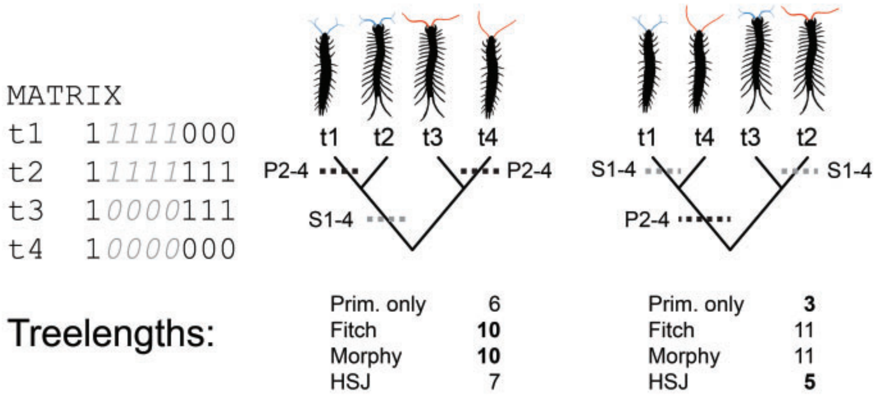

FIGURE 1. The influence of hierarchical characters on inferred phylogenetic relationships. Example matrix has four taxa (t1 and t4 represent Symphyla myriapods and t2 and t3 represent Chilopoda myriapods), four primary characters (character 1, 6–8), and four secondary characters (2–5, gray italics) that describe variation in the first character (here, variation in the antennae on the tritocerebral segment). These would be inapplicable for any taxa where the first primary character was absent (coded as "0," none shown here). If there were no secondary characters included in the matrix ("Prim. only"), the right-hand tree would be favored since the minimum number of character state transitions is smaller. Both the Fitch (Fitch 1971) and Morphy (Brazeau et al. 2019) algorithms consider secondary characters independently of the other characters (the Morphy algorithm uses the inapplicable coding symbol to recognize secondary characters independent of the other coded characters). In this example, the number of secondary characters describing the antennae is greater than the number of additional primary characters shared by the Symphyla or Chilopoda, respectively, and Fitch and Morphy will favor trees where the number of character state changes in the secondary characters are minimized relative to the additional primary characters (left-hand tree). The approach proposed in this study ("HSJ") uses a family of dissimilarity metrics to weight the influence of the secondary characters, and would still favor the right-hand tree. Silhouettes modified from images sourced from phylopic.org (*Scutigerella immaculata* by Ralf Janssen, Nikola-Michael Prpic, Wim G. M. Damen, and T. Michael Keesey; *Lithobius forficatus* by Birgit Lang).

a dissimilarity metric that explicitly handles hierarchical characters. Following the approach of Fitch (1971), for a fixed phylogenetic tree, we extend the character labelings of the leaves to the internal nodes of the tree and compute the minimal score.

Several dissimilarity metrics and semi-metrics have been proposed as concise quantitative descriptions of both morphological and ecological relative differences (Legendre and Legendre 2012; Lloyd 2016; Hopkins and St. John 2018; Wills 2001). Similarly to phylogenetic analysis, inapplicable characters have almost always been treated as missing in disparity analyses. Noting that this can lead to situations where taxa that share more primary character states are assigned larger dissimilarity values than taxa that share fewer, Kendrick and Proctor (Kendrick and Proctor 1964; Kendrick 1965), Gower (1971), and more recently Hopkins and St. John (2018) proposed methods that weight controlling primary characters according to the (dis)similarity among the secondary characters that describe them. These approaches were described and formalized in the context of Gower's coefficient (Hopkins and St. John 2018), but it is possible to count per-character differences using the same approach within the context of other common dissimilarity metrics as well (e.g., Wills' GED, see implementation in the R package Claddis v. 0.3.4, (Lloyd 2016, 2018)).

We provide a proof-of-concept of our tree scoring function through comparisons with traditional maximum parsimony and the more recent approach of Brazeau et al. (2019) (hereafter referred to as "Morphy") for both synthetic and empirical data sets. We show that our Fitch-style approach for computing scores for candidate trees can be computed in linear time in the number of taxa and characters in the analysis (see the Appendix for the proof of correctness). This modified approach for computing scores for candidate trees can then be used in any search strategy for finding the optimal tree in the space of possible trees. Further, any dissimilarity metric that equally weights characters when no inapplicable characters are present (i.e., performs identically to the regular "changes across the branches" used in parsimony when all characters are applicable) yields an optimality criterion that is computationally hard to compute (this NP-hardness result follows almost immediately from the classic proof of Foulds and Graham (1982) and is outlined in the Appendix). An implementation of the algorithm we propose (hereafter referred to as "HSJ") is provided in the statistical computing language R (R Core Team 2020).

We analyze a selection of synthetic and empirical data sets under the proposed and existing methods. The underlying design of the experiments is to test how the inclusion of secondary characters into phylogenetic studies affects which trees are considered most parsimonious. For the synthetic data sets, we look at character matrices for eight taxa, since the size allows all possible rooted trees (135,135 trees in all) to be scored. We examine the effects of increasing the number of secondary characters under the Fitch algorithm, with inapplicable characters treated as both missing and as an additional state, the Morphy algorithm, and the HSJ approach proposed herein. We also examine these four scoring methods on two empirical data sets: a moderately sized study of fossil brachiopod species (Cusack et al. 1999) and a large study of living and fossil myriapods (Fernández et al. 2016). Since the number of possible trees grows exponentially in the number of taxa, it is not possible to score every possible tree for each of the empirical data sets. Instead, we find the most parsimonious trees using only the primary characters (using PAUP; Swofford 2002, 2020), and then score these trees with the secondary characters included under the four methods above.

Materials

We created two synthetic data sets with only one controlling primary character but different numbers of secondary characters relative to the total number of primary characters. Each has eight taxa, a size we chose deliberately because it is computationally tractable to apply the Fitch, Morphy, and HSJ approaches to the full set of rooted trees with eight tips ($N = 135{,}135$, generated using the `allTrees()` function in the phangorn R package (Schliep 2010)). The matrices have 10 primary characters, where the distribution of variation in the first 10 characters is structured so that there are two unrooted most parsimonious trees (MPTs) under the Fitch approach (Chai and Housworth 2011; Radel et al. 2013). For all the examples and data sets, we assume that controlling primaries are binary characters where the "0" token represents the state for which all contingent secondary characters are inapplicable (typically, this is the character state "absent" for a trait). The tenth primary is the controlling primary character, is described by either three or ten secondary characters, and is present in half of the taxa but absent in the other half, in a way that is inconsistent with the other characters (Fig. 2). The character matrices are available in nexus format and the character types in text files (see the Implementation section).

We also selected two empirical data sets which were different from one another in size but which met the following criteria: 1) the character matrix included a large proportion of secondary characters that were broadly applicable across taxa; 2) the documentation was sufficient for reconstructing the character hierarchy; and 3) the documentation was sufficient for recoding characters to meet the assumptions of reductive coding and our approach. The first data set comes from a moderately sized study of fossil brachiopod species (Cusack et al. 1999). In the original matrix, characters were coded using a mixture of nonadditive binary coding and composite coding, and we recoded them using redundant coding; the new matrix has 23 taxa and 49 characters, 39% of which are secondary characters. The second data set comes from a large study of living and fossil myriapods (Fernández et al. 2016). The original matrix was previously modified for use as a case study for managing inapplicable characters for disparity analyses (Hopkins and St. John 2018). We have modified it further for this study by removing tertiary characters ($N = 5$) and editing the coding so that for all controlling primaries, the state that elicits inapplicable tokens in secondary characters is coded as "0." The matrix has 47 taxa representing three classes of myriapods (Symphyla, centipedes, and millipedes) and 200 characters, 35% of which are secondary characters. The character matrices are available in nexus format and the character types in text files (see the Implementation section).

Methods

We consider four methods for computing the score of a phylogenetic tree given a character matrix: 1) the standard Fitch algorithm (Fitch 1971) where inapplicable tokens are treated as missing data; 2) the standard Fitch algorithm where inapplicable tokens are treated as an additional character state; 3) the Morphy approach of Brazeau et al. (2019); and 4) the HSJ approach proposed herein.

To compute Fitch scores, we used the `fitch()` function from the R package `phangorn` v. 2.5.5 (Schliep 2010). The `fitch()` function treats all tokens other than "?" as unique characters states (method 2 above), so the token "-," typically used to identify gaps or inapplicable characters, was replaced with "?" in the character matrices before computing Fitch scores for method 1. Hereafter, we refer to the first method as "FitchM" and the second method as "FitchS."

The Morphy approach is implemented in the R package `TreeSearch` v. 0.4.0 (Smith 2018), which in turn is dependent on a the C implementation `MorphyLib` (Brazeau et al. 2017). The R function that returns Morphy scores is called `Fitch()`, for which we use the default parameters and character matrices with the "-" token for all inapplicable characters. To avoid any confusion that might stem from similarity in the function names, we refer to the algorithms of Fitch (1971) as the FitchM and FitchS approaches and that of Brazeau et al. (2019) as the Morphy approach throughout.

New Method for Scoring Phylogenetic Trees (HSJ)

The HSJ approach proposed herein extends the maximum parsimony criteria in a natural way by measuring the change across the branches in terms of a dissimilarity metric. In traditional parsimony, we seek the tree that minimizes the sum of the changes across branches. That is, given an assignment of values to the internal nodes, we sum the changes, or pairwise mismatches, witnessed across each branch. For example, in Figure 2, the characters for taxa $t1$ and $t2$ are:

$t1$ 10000**10**001 **11111111**11
$t2$ 10000**01**001 **00000000**11

with the pairwise differences in bold. The first 10 characters are primary, and the last 10 are secondary characters dependent on the tenth primary character.

MATRIX

	10 primary	+3	+10
t1	100001000**1**	*111*	*1111111111*
t2	100000100**1**	*100*	*0000000011*
t3	010000010**0**	---	----------
t4	010000001**0**	---	----------
t5	001011000**0**	---	----------
t6	001010100**0**	---	----------
t7	000110010**1**	*110*	*1111111100*
t8	000110001**1**	*000*	*0000011111*

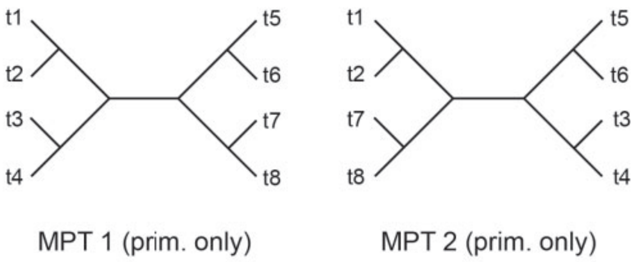

FIGURE 2. Synthetic trees with 8 taxa and 10 primary characters. The first nine characters prescribe one most parsimonious tree under Fitch ("MPT 1") with a tree score of 13 (Chai and Housworth, 2011; Radel et al., 2013). Adding the tenth character results in two most parsimonious trees under Fitch ("MPT 1" and "MPT 2") with tree scores of 15. The tenth character is inconsistent with the other characters, and requires additional character changes to be accommodated. Each tree can be rooted 13 ways, so in the set of all rooted trees, analysis of this matrix would identify 26 MPTs. Also shown are the additional secondary characters (+3 and +10) analyzed, which describe the tenth character (bolded) in the synthetic datasets.

If the characters are considered independently of one another, the dissimilarity between them is the sum of the pairwise differences: 2 changes on the first 10 primary characters plus 8 on the additional 10 secondary characters. For the taxon *t*3, the tenth primary is coded as absent and all of the secondary characters are inapplicable. The pairwise differences between *t*1 and *t*3 are bolded:

*t*1 **10**000**1**0**0**0**1** 1111111111
*t*3 **01**000**0**0**1**0**0** ----------

If the inapplicable character states are treated as missing data and not included, the dissimilarity is 5 between *t*1 and *t*3. Thus, the variation in the secondary characters describing the controlling primary has significantly increased the estimated dissimilarity between *t*1 and *t*2, which otherwise agree on most of the primary characters. Because the secondary characters are inapplicable to *t*3, they have no influence on the estimated dissimilarity, and the dissimilarity between *t*1 and *t*3 is estimated to be smaller than that of *t*1 and *t*2, even though *t*1 shares fewer primary characters with *t*3 than with *t*2. Since the tree score is the sum of these differences (Farris 1970; Fitch 1971), treating inapplicable characters as missing data can skew the analysis to favor trees that would place taxa with similar primary characters distant from each other. Treating inapplicable characters as a new, separate state, will also similarly skew the analysis, because having a new separate state increases the dissimilarity of all pairwise comparisons like that of *t*1 and *t*3, by 10. This results in overweighting the tenth primary character and favors trees with clades that separate taxa with secondary characters from those without.

To explicitly account for hierarchical characters in the analysis, we estimate the contribution of each character to the tree score using a new family of dissimilarity metrics ("HSJ" of Hopkins and St. John (2018), see Appendix for detailed definition). The possible per-branch change of noncontrolling primary characters is the same as that obtained using traditional parsimony (Farris 1970; Fitch 1971). The possible per-branch change of controlling primaries, however, is weighted by the differences among the secondaries (when applicable) and scaled using a parameter α that ranges from 0 to 1. Specifically, the contribution of a controlling primary to the tree score is calculated as $\alpha \cdot \frac{d}{m}$, where d is the number of nonmatching secondary characters dependent on that primary, and m is the total number of secondary characters dependent on that primary character. Thus, a single controlling primary and its associated secondary characters can contribute at most 1 per branch (the same as any noncontrolling primary) to the tree score.

As an example, consider again the first three taxa from Fig. 2. The contribution to the tree score of noncontrolling primary characters is the same as that for traditional parsimony (Farris 1970; Fitch 1971), so along the branches connecting *t*1 and *t*2, the noncontrolling primaries differ at the sixth and seventh characters and contribute 2 to the tree score. The tenth character is present in both *t*1 and *t*2 and has ten additional secondary characters describing it. *t*1 and *t*2 disagree on eight of them, so the contribution would be $\alpha \cdot \frac{8}{10}$. As a result, when $\alpha=0$, the secondaries do not contribute at all, but when $\alpha>0$, the secondaries contribute a fractional amount associated with the number of differences among the secondary characters. For *t*1 and *t*2, the largest the contribution of the controlling primary could be is 0.8, when $\alpha=1$ (in comparison to 8 if the secondary characters had been considered independently, see above). For comparison, the noncontrolling primaries along the branches connecting *t*1 and *t*3 contribute 4 to the tree score. However, since the tenth primary character is coded as present in *t*1 and as absent in *t*3, the secondary characters are only applicable to *t*1. In this case, for all values of α, 1 is added to the overall score to represent the difference in the controlling primary. Using this approach, the contribution of this character due to variation between *t*1 and *t*3 is greater than the contribution due to variation between *t*1 and *t*2 for any value of α. Thus, the ranking of the contributions from the two pairs of taxa reflects the shared primary characters and is not unduly modified by the differences in secondary characters that describe an otherwise shared controlling primary.

To assign a score to a tree, we follow the basic framework of the classic algorithm of Fitch (1971). We sketch the ideas behind the algorithm here and provide a precise definition and proof of correctness in the Appendix. If a tree already has labelings on all the leaves and the internal nodes, then the score of the tree is the sum of the HSJ dissimilarities across all the branches. However, most trees have only the leaves labeled by the characters and the internal nodes unlabeled and as such, we must first assign labelings to the internal nodes. If a primary character has no secondary characters describing it, then the labeling that the Fitch algorithm produces will also be optimal for these characters. For controlling primary characters, we augment the Fitch algorithm, due to the asymmetry of the contributions of choices: agreeing on a primary character can add 0 to the overall score (in the case where both are coded as absent) or can add some positive difference (where both are present, differ on some secondary traits, and have the scaling parameter, $\alpha > 0$). As such, for each set of secondary characters and their controlling primary character, we keep track of both the possible score when the controlling primary character is present and when it is absent. First, we give a preliminary labeling to all internal nodes, using the Fitch approach. We then, starting at the leaves, compute the possible score for when the controlling primary character is absent and when it is present at the parent node. Moving up the tree, we use the already computed numbers for the children nodes and use the best possible scoring of the possible combinations and update the labels to reflect the choices that optimized the score. We continue until we reach the root of the tree and return the minimal score.

As an example, consider $t1$ and $t2$ which have the same parent in Figure 2; call it p. We make a first pass and give p the tentative label of:

10 Primary Characters:	Secondary Characters:
1 0 0 0 0 0/1 0/1 0 0 **CP**	0/1 0/1 0/1 0/1 0/1 0/1 0/1 0/1 1 1

where the bolded CP is a placeholder for the primary associated with the secondary characters. We compute and store two scores for the parent node, p: if CP is coded as absent, and if CP is coded as present. If CP is 0 (absent), then the distance to the two leaves, both which have that character present, is $1+1=2$ plus the change in the remaining characters (also 2, for a total of 4). If CP is 1 (present), then the total distance from p to the two leaves is $\alpha \cdot \frac{8}{10} + 2$, as described above. At this point, labeling CP as present at node p yields the minimum score. For nodes whose children are internal nodes, we store the minimum values for all possible labels (absence and presence for each child for each possibility for the parent) as we work through the tree in postorder.

Comparison of Methods

Our strategy for comparing the HSJ approach with the FitchM, FitchS, and Morphy approaches depended on the size of the character matrix. For the 8-tip synthetic data sets, we scored the entire space of trees using all four approaches and then compared the subsets of trees with optimal scores. For the empirical data sets, computational limitations prevented us from scoring the entire tree space for any approach. Instead, we compared the performance of each approach within a subset of the tree space. The subset was defined as the set of optimal trees if only the primary characters are known, because if there are no inapplicable characters in the matrix, then the Morphy and HSJ approaches are equivalent to the Fitch approaches. Similarly, the HSJ approach is equivalent to the Fitch approaches when $\alpha = 0.0$, as the contributions of the secondary characters are ignored in this situation. To efficiently identify the subset of optimal trees based on just primary characters, heuristic searches were conducted in PAUP*4.0a167 (Swofford 2002, 2020); taxa were added by random sequence addition with 1000 replicates and branch swapping was performed using the tree bisection reconstruction option (TBR). These sets of noncollapsed trees were then scored for the full set of characters (i.e., including secondary characters) using FitchM, FitchS, Morphy, and HSJ with $\alpha = 0.1, 0.2, 0.3, \ldots 1.0$. Thus, for the empirical data sets, we compared the different approaches in the context of the question: which trees are considered optimal when secondary characters are added to the analysis?

Because the number of trees within the subsets of optimal trees is still quite large, we employed Landmark Multi-Dimensional Scaling (LMDS) in order to visualize the tree space. LMDS is a computationally efficient approximation of classical multi-dimensional scaling. The first step is to run a classical MDS on a matrix of pairwise RF distances for a random subset of trees; these data are referred to as the "landmarks." The second step is to place each remaining tree within the resulting lower-dimensional space based on its RF distances from the "landmark" trees that define the space (de Silva and Tenenbaum 2004).

Results

Synthetic Data Sets with Eight Taxa and Balanced Variation in Characters

Two unrooted (26 rooted) optimal trees are supported when the 8 taxa are coded for just 10 primary characters (Fig. 2). When secondary characters are added but remain low in number relative to the total number of primary characters (Fig. 2, "+3"), the FitchM, Morphy, and HSJ approaches all identify the same two optimal unrooted trees (26 rooted trees), and all trees comprise the same sister taxon pairs (Fig. 3). The FitchS approach identifies one of the unrooted trees as optimal ("MPT 2" of Fig. 2).

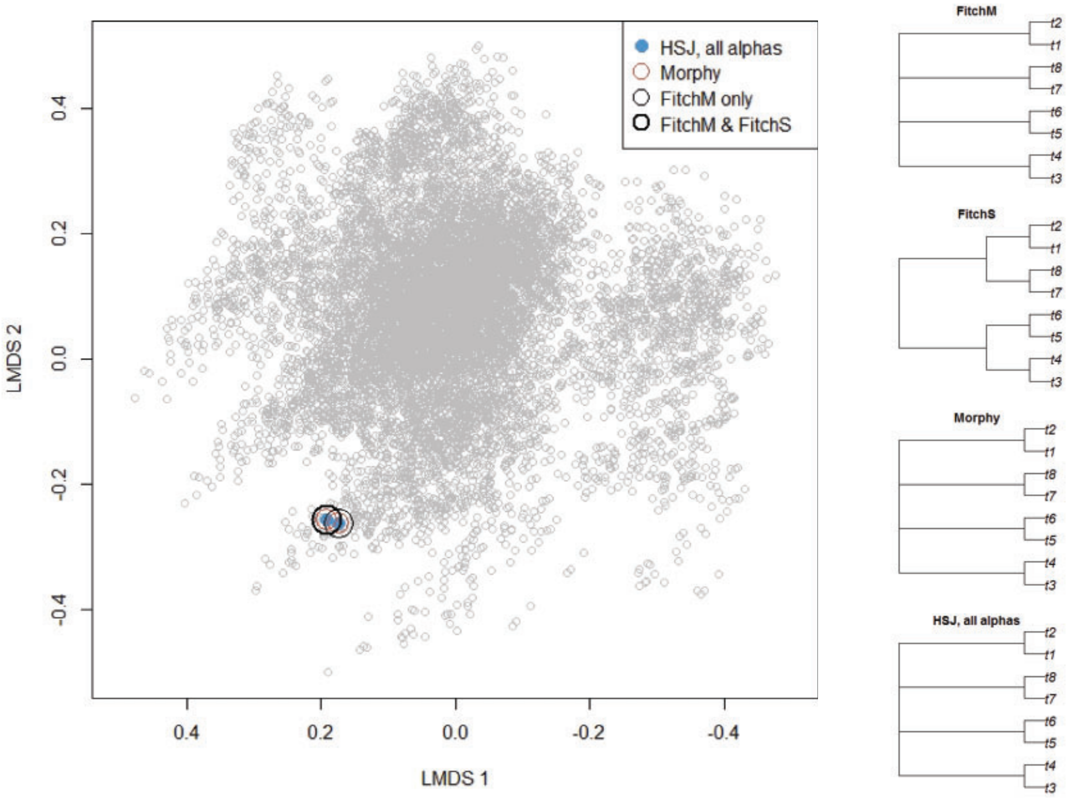

FIGURE 3. Best-scoring trees found by each algorithm using the synthetic data set matrix with 13 characters and fewer secondary characters than primary characters. Left panel: LMDS (de Silva and Tenenbaum 2004) of all trees with eight tips with the best-scoring trees for each treatment color-coded. The same two trees were assigned optimal scores using the HSJ approach for all values of α. Right panels: 50% majority rule consensus trees for the best-scoring trees for each approach. LMDS based on 500 randomly selected trees; implemented using the R package `lmds` (Cannoodt and Saelens 2019). See Figure S2 of the Supplementary material available on Dryad for the histograms of scores.

However, when the number of secondaries is high (Fig. 2, “+10”), the HSJ algorithm identifies the same set of optimal trees as with the smaller matrix, but the FitchM and Morphy approaches identify a different, larger set of trees (273 rooted/21 unrooted; Fig. 4, left panel), of which the FitchS approach identifies a subset as optimal (65 rooted/5 rooted). As expected, the primary difference between the trees that the HSJ approach identifies compared with the trees that the FitchM, FitchS, and Morphy approaches identify is the grouping of taxa 1, 2, 7, and 8 (Fig. 4, right panels). Moreover, since none of the trees that HSJ identifies are the same as any of the trees that FitchM, FitchS, and Morphy identify, we can infer that none of the latter comprise (*t*1, *t*2) and (*t*7, *t*8) as sister taxa, even though these groupings are supported by the variation in the primary characters (and identified by all approaches when the number of secondary characters is low). The subset of trees identified as optimal by the FitchS approach comprise the sister taxa (*t*1, *t*7) and (*t*2, *t*8), which together make up one of two subclades.

Note also that for both matrices, the HSJ approach identifies the same trees regardless of the value of scaling parameter α, even when $\alpha = 0$, which disregards the secondary characters entirely (this set of trees is the same as what would be identified using the Fitch approach on a matrix comprised of only the primary characters).

Empirical Data Set 1: Fossil Brachiopods

The PAUP* heuristic search on the brachiopod data set restricted to primary characters yielded 54,524 noncollapsed MPTs with a score of 78. We scored these candidate trees using the complete data set, including secondaries, under the FitchM, FitchS, Morphy, and HSJ approaches, with the scaling parameter ranging from $\alpha = 0.1, 0.2, 0.3, \ldots 1.0$ for the latter. Sorting the trees by the HSJ tree score when $\alpha = 1.0$ results in a curve that reflects a complex distribution of tree scores (Fig. 5a). As α decreases, the rank order amongst a small number of similarly scoring trees shifts, but the same basic ranking of the trees is retained (Fig. 5b, c). Regardless of the α value, the HSJ approach found the same 68 best-scoring trees (Fig. 5g), at least within the subset of 54,524 trees examined herein. The distribution of Morphy scores for the set of 54,524 trees is asymmetrically skewed towards smaller values, and the scores are not correlated with the HSJ scores (Fig. 5d). FitchM scores for the set of 54,524 trees are asymmetrically skewed towards larger values

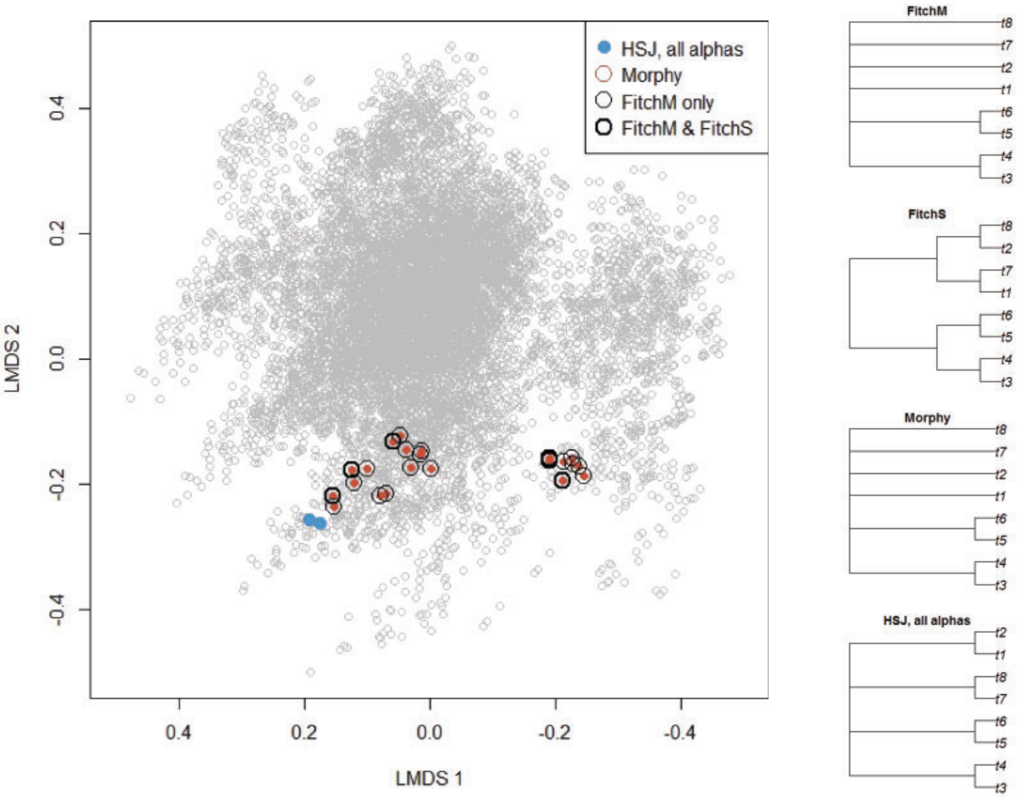

FIGURE 4. Best-scoring trees found by each approach using the synthetic data set matrix with 20 characters and as many secondary characters as primary characters. Left panel: LMDS (de Silva and Tenenbaum 2004) of all trees with eight tips and the best-scoring tree for treatment color-coded. The same two trees were assigned optimal scores using the HSJ approach for all values of α; the same two trees are found as for the other synthetic matrix (see Fig. 3). Right panels: 50% majority rule consensus trees for the best-scoring trees for each approach. Note that the sister taxon arrangements for taxa 1, 2, 7, and 8 vary across the consensus trees. LMDS is the same as that shown in Figure 3. See Figure S2 of the Supplementary material available on Dryad for histogram of scores.

and are also not correlated with the HSJ scores (Fig. 5e). FitchS scores are the most normally distributed and tend to correlate with HSJ scores although there is a large variance (Fig. 5f). FitchM and Morphy yield more similar rankings of scores to each other than either to FitchS or HSJ (Fig. S5 of the Supplementary material available on Dryad). In contrast to the other approaches, the FitchS approach identified a single tree of the set of 54,524 trees as optimal.

Although all four approaches identify a small number of optimal trees within the set of 54,524 trees, each set is distinct from the other (Fig. 5h,i and Fig. S4 of the Supplementary material available on Dryad). This is true even for the FitchS tree, which shares tree space with the subset of trees found optimal by the HSJ approach (Fig. 5g), but is not one of the latter set. Amongst the best-scoring trees, shared features, such as the subclade (*Aulanotreta*, (*Bicarinata*, *Pseudolingula*), (*Glottidia*, *Lingula*)) and the sister group (*Lingulasma*, *Pseudolingula q*), are present in the entire set of 54,524 trees (Fig. S4c of the Supplementary material available on Dryad). The HSJ and Morphy approaches both identify a subclade (*Broeggeria*, *Zhanatella*, *O. eichwaldi*, *Lingulella*, *Oepikites*, *Experilingula*, *Eoobolus*, *Ungula*, *O. apollonis*, *Lingulellotreta*, *Lingulasma*, *Pseudolingula q*) that is not resolved in the consensus of the best-scoring trees using either FitchM or FitchS (Fig. S4a of the Supplementary material available on Dryad). Only the FitchS and HSJ best-scoring trees place the original outgroup, *Acrothele* and *Acrotreta*, sister to the rest of the taxa (Fig. 5h and Fig. S4b of the Supplementary material available on Dryad). In summary, although the different approaches select similar optimal trees, the variation between them could be enough to differentially influence any systematic taxonomy and evolutionary inferences based on them. Comparisons between these results and the original tree figured in Cusack et al. (1999) are limited by the fact that the character matrix was recoded using a different coding strategy and that we had difficulties replicating the original results (Fig. S3 of the Supplementary material available on Dryad).

Empirical Data Set 2: Myriapods

The PAUP* heuristic search on the myriapods data set restricted to primary characters resulted in 34,560 noncollapsed MPT trees with a score of 215. We scored these candidate trees using the complete data set, including secondaries, under the FitchM,

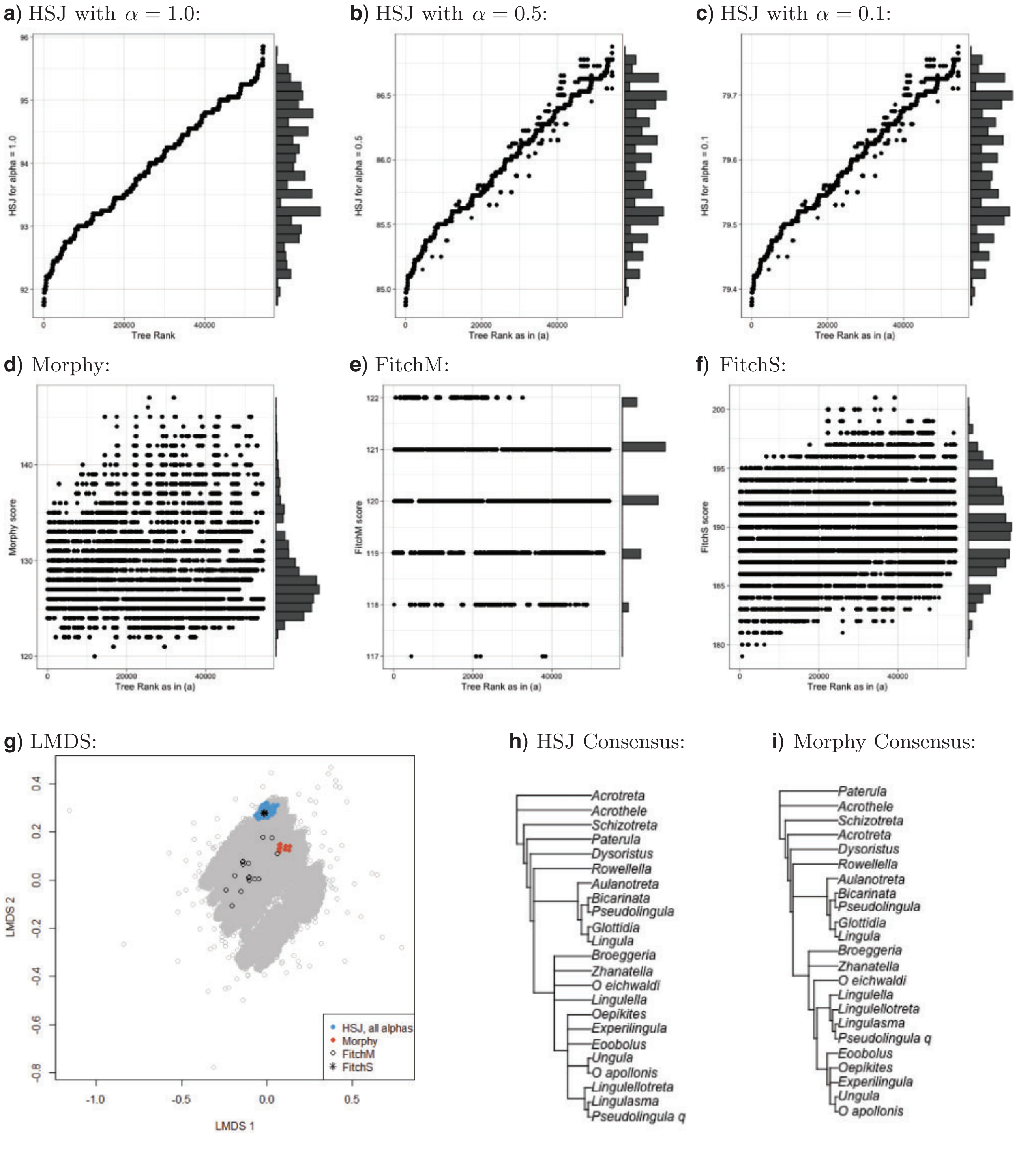

FIGURE 5. Analysis of the 54,524 MPTs returned by PAUP on the brachiopod dataset restricted to primary characters. a) HSJ scores (entire data set, $\alpha = 1.0$), arranged from best to worst scoring, with histogram of scores on right margin. b) HSJ scores (entire data set, $\alpha = 0.5$), ordered as in a), with histogram of scores on right margin. c) HSJ scores (entire data set, $\alpha = 0.1$), ordered as in a), with histogram of scores on right margin. d) Morphy scores (entire data set), ordered as in a), with histogram of scores on right margin. e) FitchM scores (entire data set), ordered as in a), with histogram of scores on right margin. f) FitchS scores (entire data set), ordered as in a), with histogram of scores on right margin. g) LMDS of 54,524 MPTs with best-scoring trees using Fitch, Morphy, and HSJ approaches (see Fig. 3 for information on implementation). Note that although the tree found by FitchS is in the same part of treespace as the trees identified by HSJ, it is not one of them. h) The strict consensus of the 68 best scoring tree for HSJ. The same trees had the lowest score for all values of $\alpha > 0$. i) The strict consensus of the 10 best scoring trees for Morphy. See also Figures S4 and S5 of the Supplementary material available on Dryad.

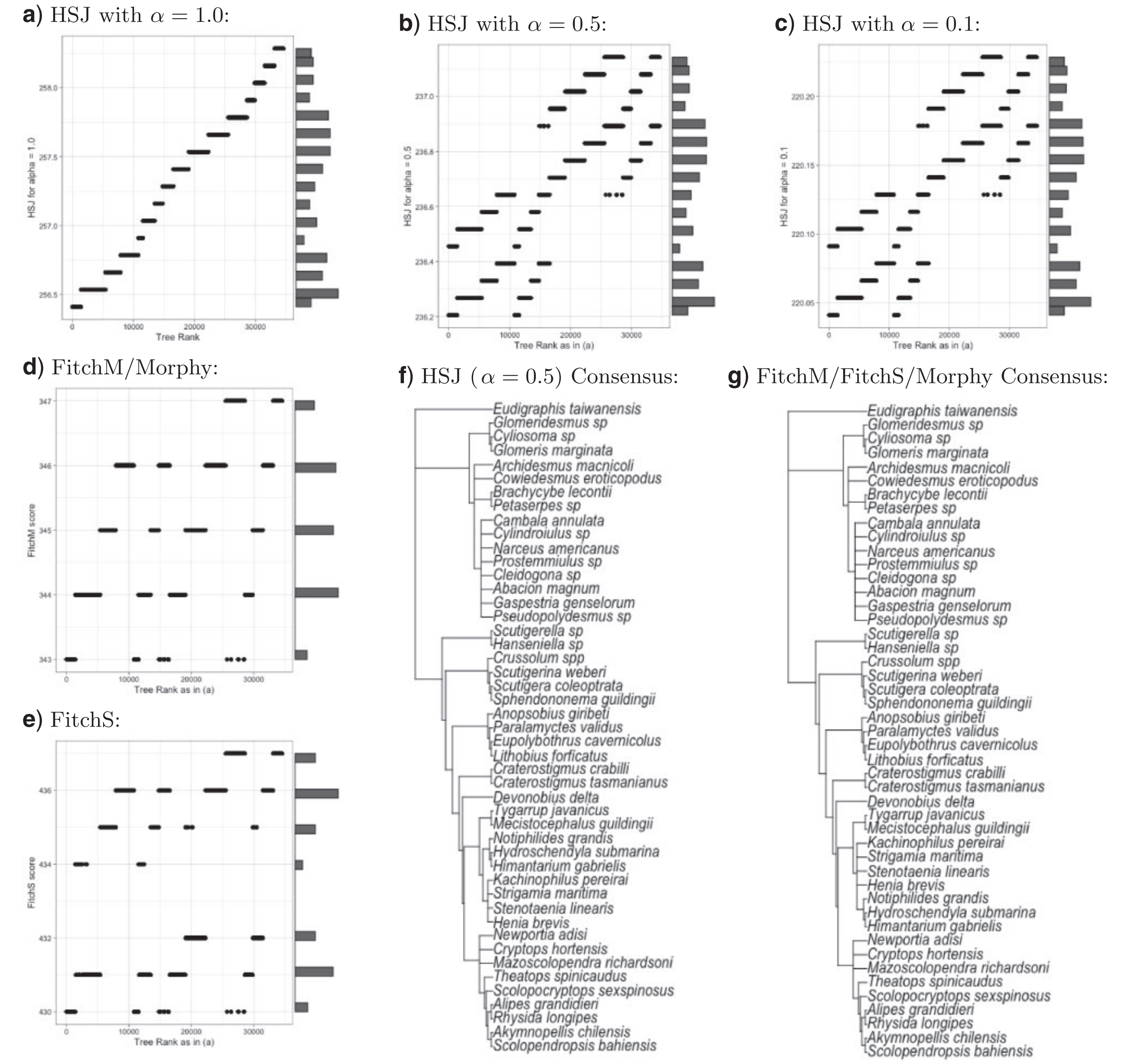

FIGURE 6. The 34,560 best scoring trees returned by PAUP on the Myriapod data set restricted to primary characters. a) HSJ scores (entire data set, $\alpha = 1.0$) for the 34,560 MPTs returned by PAUP for the data set restricted to primary characters. The marginal plot shows a histogram of the scores. b) HSJ scores for $\alpha = 0.5$ for the 34,560 MPTs, ordered as in a), with histogram of scores on the right margin. c) HSJ scores for $\alpha = 0.1$ for the 34,560 MPTs, ordered as in a), with histogram of scores on the right margin. d) Morphy scores for the 34,560 MPTs, ordered as in a), with histogram of scores on the right margin. The FitchM approach yielded identical scores. e) Scores under the FitchS approach for the 34,560 MPTs, ordered as in a), with histogram of scores on the right margin. f) The strict consensus of the 1464 best scoring trees for HSJ with $\alpha = 0.5$. g) The strict consensus tree of the 2700 best scoring trees for Morphy, FitchM, and FitchS (all approaches yielded the same set of optimal trees). The consensus was identical to the majority rule consensus of all 34,560 trees.

FitchS, Morphy, and HSJ approaches, with the scaling parameter ranging from $\alpha = 0.1, 0.2, 0.3, \ldots 1.0$ for the latter. Sorting the trees by the HSJ tree score when $\alpha = 1.0$ results in a nearly uniform distribution of the trees (Fig. 6a). As α decreases, the rank order amongst similarly scoring trees shifts but the same basic ranking of the trees is retained (Fig. 6b,c). However, the magnitude of the shift is greater for the myriapod data set compared to the brachiopods. The other notable difference between the results for each empirical data set is that for the myriapods, the number of best scoring trees for the HSJ approach varied for values of α, ranging from 1414 trees ($\alpha = 1.0$) to 1464 trees ($\alpha = 0.5$), totaling 2700 unique trees. Figure 5e shows the strict consensus tree of the 1464 best scoring trees for $\alpha = 0.5$.

The Fitch and Morphy approaches identified the same 2700 optimal trees, but with a score of 343. The FitchM/Morphy trees in Figure 6d are ordered the same as the previous panels Fig. 6a-c but see only five possible scores of 343 (2700 trees), 344 (9720 trees), 345 (8640 trees), 346 (9180 trees), and 347 (4320 trees). Although the FitchS approach identified the same optimal trees, the remaining trees were distributed across six additional distinct scores.

The strict consensus trees for the best-scoring trees under any approach were similar in overall structure (Fig. 6e,f), most notably monophyly among major subclades, because this structure is shared across the subset of 34,560 trees. In this example, then, the addition of the secondary characters served only to increase resolution within subclades; this service was slightly enhanced using the HSJ approach for any given value of α compared to the Morphy or Fitch approaches.

Discussion

The inclusion of secondary characters into phylogenetic studies have the potential to provide deeper insight into the relationship between the taxa. Our focus is on data matrices that use reductive coding (Strong and Lipscomb 1999), where the assumptions of the hierarchical organization of the characters is incorporated via explicitly denoting which secondary characters are dependent on which primary characters. We explored three approaches to scoring phylogenetic trees: the ubiquitous approach of Fitch (1971), the recent approach of Brazeau et al. (2019), and a new HSJ approach described herein. The Fitch approach computes the sum of the changes across the branches of the tree. Missing characters are not included in the total, and inapplicable characters are treated as missing or a separate, new character state. The scoring of a single tree takes linear time in the number of taxa, while finding the optimal tree among all possible trees is computationally hard (Foulds and Graham 1982). The Morphy approach of Brazeau et al. (2019) seeks to include inapplicable characters in a more informative way by grouping secondary characters that are present into "regions" where similaries can be explained by homologies (De Laet 2015), and each new region adds to the overall score of the tree. While the computational complexity is not explicitly stated in Brazeau et al. (2019), analysis of their tree scoring algorithm yields an upper bound of linear time in the number of taxa and finding the optimal tree is computationally hard by reduction to the optimal parsimony problem (Foulds and Graham 1982). Our proposed approach follows the original spirit of the maximum parsimony criteria that seeks to minimize the number of changes across branches. Secondary characters contribute, not independently, but as a fraction of the primary character they describe. Building on the canonical approach allows us to extend existing algorithms and proofs to give correctness and complexity results yielding linear time algorithms for scoring individual trees and the computational hardness (NP-completeness) of finding the optimal tree (see Appendix).

These four different approaches to scoring phylogenetic trees behave in disparate ways when the proportion of secondary characters is increased. Due to the size of the synthetic data sets examined, we were able to generate all 135,135 possible rooted trees on those eight taxa and exhaustively find the optimal trees for each method. While it has been suggested that secondary characters do not affect the phylogeny (Lockhart and Koenig 1965; Caira et al. 1999), we note a more nuanced situation, where, not surprisingly, both the approach used for the inclusion of the secondary characters and the phylogenetic informativeness of the secondary characters, play significant roles in the analysis. The approach used for the inclusion of secondary characters in the analysis can affect the clades supported in trees with optimal scores, but this depends on the relative proportion of secondary characters as well as the character consistency among both primary and secondary characters. For the synthetic data sets where all primaries were consistent *except* for the controlling primary, increasing the amount of secondary data from 23% (3 secondaries of 13 total characters) to 50% (10 secondaries of 20 characters) resulted in a different set of optima using the Morphy and Fitch approaches (though the same as, or subsets of, each other), while the HSJ approach continued to favor the same two trees, for all values of the scaling parameter ($\alpha = 0, 0.1, 0.2, \ldots, 1.0$). Further, the number of optimal unrooted trees increased using the Fitch and Morphy approaches and the resulting consensus tree retained unresolved clades where the secondary characters are applicable.

While the results of applying the Fitch and Morphy approaches to the synthetic data sets are the same, the reasons are different. For the FitchM approach, the inapplicable characters are coded as missing data, and, as a result, the differences in the secondary characters dominate the overall scores of the trees. For the FitchS approach, the inapplicable characters are treated as the same additional character state for all taxa where the controlling primary character is coded as absent, and as a result, the weight of that absence is increased as the number of secondary characters contingent on that primary character increases. In contrast, the Morphy approach is designed to increase the scores of topologies where primary characters described by secondary characters are homoplastic, by adding a point to the total tree score for each "region" where a subset of taxa for which the character in question is inapplicable separates two subsets of taxa for which the character is applicable (see the mapping of "R" in Fig. S1 of the Supplementary material available on Dryad). Because character contingency is ignored, this penalty is assessed separately for each secondary character,

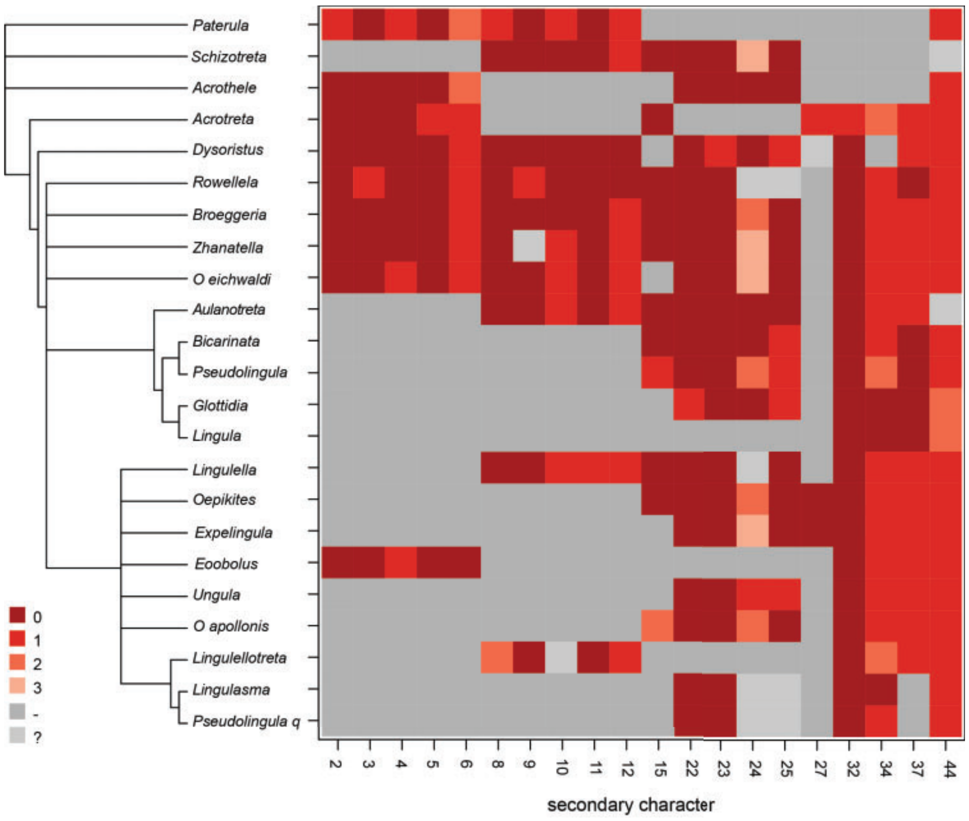

FIGURE 7. Map of secondary characters in brachiopod data set. The tree is the majority rule consensus of 54,524 MPTs found by PAUP based only on primary characters. Elements are shaded on a red scale by the state of the secondary character. Characters coded as missing (i.e., '?') are light gray; those coded as inapplicable (i.e., '-') are dark gray; all others are red scale (see legend).

quickly increasing the length of the tree when many secondary characters associated with a homoplastic primary character are included in the matrix.

Favoring trees that group characters into regions of homology (De Laet 2015; Brazeau et al. 2019) is intuitive and appealing but hard to capture in practice. As noted above, for the synthetic data set, secondary characters that are homoplastic across the tree unduly influenced the results. A more subtle issue is identifying the regions of homoplasy, and for this, comparison of the two empirical data sets is illuminating. Both data sets have moderately large percentages of secondary characters (35% and 39%). For the brachiopods, primary and secondary characters are inconsistently distributed among the taxa. Further, almost all combinations of controlling primary characters are expressed across the set of taxa, making secondary characters homoplastic across many trees (Fig. 7) and prompting additional penalties for "regions" using the Morphy approach. This leads to differences in the Fitch and Morphy scores, with the Morphy score higher on average than the FitchM scores, and the FitchS scores higher on average than the Morphy scores (Fig. 5d–f). The correlation between scores for the methods was low ($|r| < \frac{1}{3}$, see Fig. S5 of the Supplementary material available on Dryad) but was the strongest between the Morphy and FitchM approaches. Thus across all approaches, the inclusion of the secondary characters adds significant phylogenetic information (Fig. 7), but because of the wide distribution of the secondary characters, the treatment of the secondary characters yields different trees favored by the different methods. In such situations, the approach employed for handling inapplicable characters can have a substantial impact on the results of the analysis.

For the myriapods data set, in contrast, the Fitch and Morphy approaches gave the same score(s) to all 34,560 trees comprising the subset produced by the PAUP search based on just primary characters (Fig. 6d). Closer examination of the distribution of secondary characters across the consensus tree shows that each can be mapped in a homologous fashion to the tree (Fig. 8), and as such, does not incur a penalty for having multiple regions of homology under the Morphy approach. With no penalties for multiple regions, the Morphy score is identical to that computed by the FitchM approach, neither of which provide more resolution than the analysis on the primary characters alone. Further, the states of the secondary characters are nearly constant in the subclades and, as such, add little additional information over that provided by the primary characters. Under the HSJ approach, only two more polytomies are

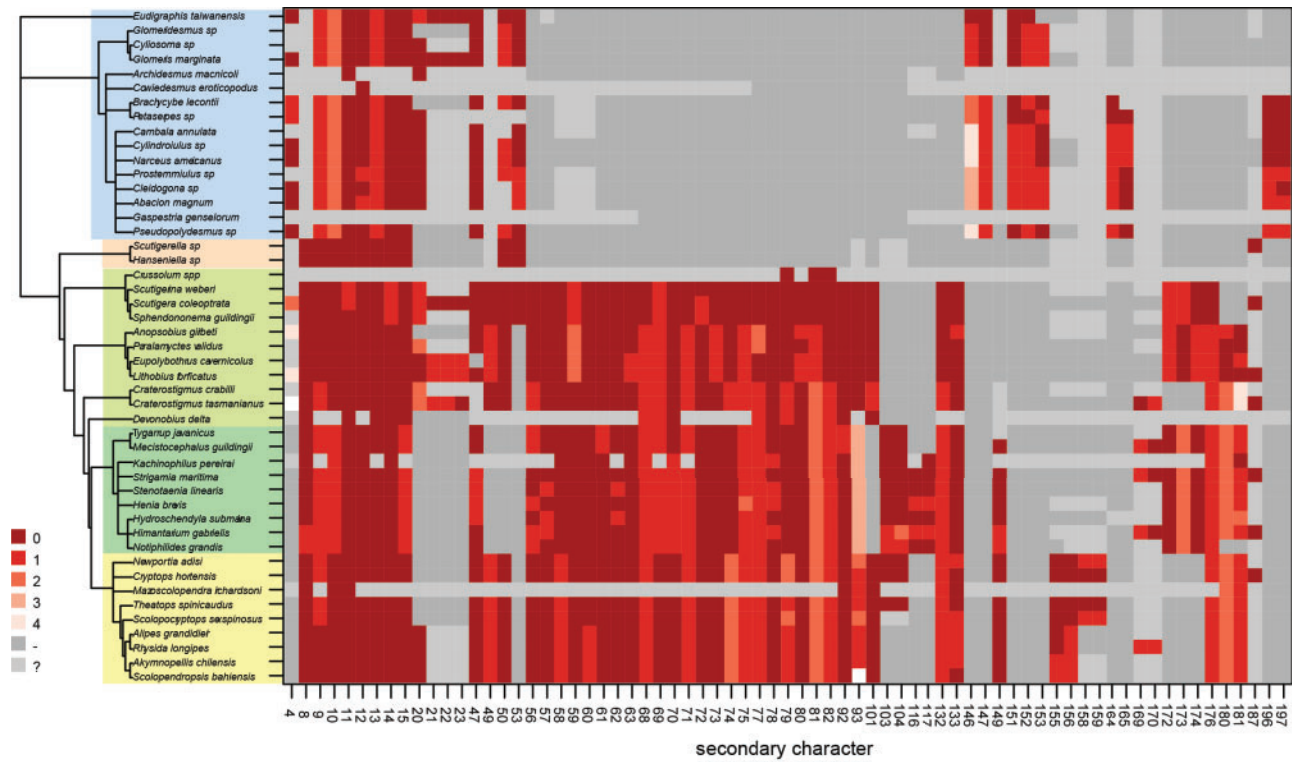

FIGURE 8. Map of secondary characters in myriapod data set. The tree is the strict consensus of 34,560 MPTs found by PAUP based on just primary characters. The tree is colored by subclade: Blue = Diplopoda (millipedes); orange = Symphyla; lime green = Chilopeda (centipedes) subclades Scutigeromorpha, Craterstigmomorpha, and Lithobiomorpha; grass green = Chilopoda (centipedes) subclade Geophilomorpha; yellow = Chilopoda (centipedes) subclade Scolopendromorpha. Elements are shaded on a red scale by the state of the secondary character. Characters coded as missing (i.e., "?") are light gray; those coded as inapplicable (i.e., "-") are dark gray; all others are red scale (see legend). Secondary characters with minor homoplasy are not sufficient to prompt "region" penalties when applying the Morphy approach, which is why the tree scores are the same for those two approaches.

resolved, seen in the Geophilomorpha (Chilopeda) (Fig. 6e). To summarize, adding secondary characters to the empirical data sets can reduce the number of optimal trees (and thus increase resolution in the consensus tree), because including secondary characters includes potential synapomorphies for subclades; this is analogous to the observation that excluding characters with a high proportion of missing data can still contribute to low resolution (Kearney and Clark 2003). However, if the variation in secondary characters is low or highly consistent with the phylogenetic information contained within the primary characters, the consensus tree will not be greatly affected.

Reductive coding sensu Strong and Lipscomb (1999) allows the "nested" relationship among hierarchical characters to be described. In addition to the primary and secondary characters that are the focus on this study, further nesting, such as tertiary, quaternary, and quinary characters, can be defined as those describing the variation within the level above. While theoretically possible, this further nesting of characters is relative rare in coded matrices. For example the myriapod data set of Fernández et al. (2016), we analyze contained 2.4% tertiary characters overall and no quaternary or quinary characters. Due to their rarity and the fact that other methods do not consider nested characters explicitly (the Morphy algorithm, e.g., recognizes hierarchical characters based on the presence of at least three inapplicable tokens, but not whether those characters are secondary, tertiary, etc), we focused our analysis on the comparison of results for matrices that have only primary and secondary characters. We note, however, that the scoring of tertiary characters follows the similar nesting that we applied to give scores to a controlling primary by their associated secondary characters (see Hopkins and St. John (2018) for details), and thus our approach can be extended to accommodate any number of hierarchical levels. The proof of correctness can be similarly extended and the running time of computing the tree score would remain the same.

The inclusion of inapplicable characters into a phylogenetic analysis is an intriguing challenge from the choice and coding of characters in the matrix to the methods used to score the candidate phylogenetic trees. The reductive coding sensu Strong and Lipscomb (1999) yields a hierarchical structure allowing the analysis of the relative contributions of the primary characters that apply to all taxa and the secondary characters that describe variation in primary characters in some taxa. However, treating secondary characters as missing data when they are inapplicable has limitations: it may favor trees where internal nodes are assigned impossible states, where the arrangement of taxa within subclades is unduly influenced by variation in distant parts of the tree, or where taxa that otherwise share most primary characters are grouped distantly. The first two problems have been addressed by recent work of Brazeau et al. (2019). Of those examined herein, only the HSJ approach consistently gave a feasible solution to the problem of including inapplicable characters in a way that resolves subclades without biasing the overall tree favored by the search. This suggests that optimality criteria based on a parsimony framework, albeit with a more refined way to compute change across branches, provides a discerning approach to including inapplicable characters into broader phylogenetic analysis.

Implementation and Data

All methods have been implemented in R. Data and code are available on github (`stjohn/hsj`). Data is available on Dryad and Morphobank. Our code builds on the framework set up in `TreeSearch` (Smith 2018) and `TreeTools` (Smith 2019) that rely on `phangorn` (Schliep 2010) for additional phylogenetic analysis. The `TreeSearch` package includes two methods, `TreeSearch()` and `Ratchet()` that allow the tree scoring function to be specified as a parameter for which we implemented our tree scoring function, `hsjTS()`. It relies on the HSJ dissimilarity R functions (Hopkins and St. John 2018) that are included, which in turn rely on the `Claddis` (Lloyd 2016, 2018) R package.

Acknowledgments

We would like to thank Frank Burbrink and Brian Smith, for organizing the Explorations in Ecological-Evolutionary Modeling (ECEM) discussion group that led to this project. We also thank Matthew Wills, Graeme Lloyd, and the organizers and participants of the 2018 Royal Society Disparity Workshop in Milton Keynes, UK, for discussion about dissimilarity metrics, and to Ward Wheeler for discussions about parsimony. We also thank Graeme Lloyd, Martin Brazeau, Thomas Guillerme, and Martin Smith for their extendable software frameworks. We would also thank Jan de Laet, Thomas Guillerme, and Will Pett for their reviews and April Wright and Bryan Carstens for editorial support.

Funding

This work was supported by the Simons Foundation (K.S.) and the National Science Foundation for funding (M.J. & K.S.).

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.m37pvmd06.

Appendix

We prove that our scoring algorithm computes the optimality criteria and further show that finding the optimal tree under this criteria is computationally hard.

Definitions and Notations

A phylogenetic tree $\mathcal{T}$ consists of vertices or nodes, $V(\mathcal{T})$, connected by branches or edges $E(\mathcal{T})$. The vertices consist of nodes of degree 1 or *leaves*, $L(\mathcal{T})$, and internal nodes. A *two-state* or *binary character* χ for $L(\mathcal{T})$ is a function $\chi: L(\mathcal{T}) \to \{0,1\}$. We say that the function χ': $V(\mathcal{T}) \to \{0,1\}$ *extends* the character χ if χ' is identical to χ on $L(\mathcal{T})$. For a set of characters, $S = \{\chi_1,\ldots,\chi_k\}$ and node x, we call $(\chi_1(x),\ldots,\chi_k(x))$ the *label* of x. We will call a function, $s(\mathcal{T},S)$ that assigns nonnegative real numbers to input tree, $\mathcal{T}$ and set of characters, S, a *(tree) scoring function*. We assume a hierarchical coding of characters where *primary characters* describe traits. We say a character is *controlling* if its value indicates the presence or absence of additional characters states. *Secondary characters* describe variation within such traits. A *tertiary character* is similarly defined, referring back to a secondary character.

Dissimilarity Metrics

Hopkins and St. John (2018) introduced a family of metrics that extend Gower's weighting solution (Gower 1971) and allow the contribution of the primary character to be scaled, via a parameter α that can range from 0 to 1. We define:

$$S_\alpha(i,j) = \begin{cases} \dfrac{(1-\alpha)\cdot 1 + \alpha \cdot \dfrac{s}{m} + k}{n+1} & \text{if the primary character} \\ & \quad \text{is present for both} \\ \dfrac{k}{n+1} & \text{if the primary character} \\ & \quad \text{is present for only one} \\ \dfrac{1+k}{n+1} & \text{otherwise (neither has} \\ & \text{the primary character)} \end{cases} \tag{A.1}$$

where m is the number of secondary characters, s is the number of secondary characters that agree between the ith and jth taxa, and k be the number of primary characters that agree out of the remaining n primary characters. The dissimilarity metric is defined as $d_\alpha(i,j) = 1 - S_\alpha(i,j)$.

Optimal Internal Labelings

Given a tree and sequence of characters, we present an algorithm that extends the characters to the internal nodes and yields an optimal score where the change across branches is computed by the HSJ measure in Equation A.1. Our approach follows Fitch (1971): we make a postorder traversal of the tree to assign initial internal node labelings and then make a second pass to score the tree. We divide the characters into "blocks." We set the first block to be all of the noncontrolling primary characters. Each subsequent blocks consists of a controlling primary character and its associated secondary characters. For the algorithm of Fitch (1971), the score of each character can be computed first and then added to the overall score. Similarly, we compute the score for each of the blocks of characters, and then added to the overall score.

When there are no controlling primary characters, Equation A.1 becomes: $S_\alpha(i,j) = \dfrac{k}{n}$ where k is the number of (primary) characters that agree out of the n characters. The resulting dissimilarity metric, $d_\alpha(i,j) = 1 - \dfrac{k}{n} = \dfrac{d}{n}$, where d is the number of differences between ith and jth taxa. When scaled by the total number of characters, this is exactly the measure used for traditional parsimony. More formally:

Lemma 1. *Let $s_F(\mathcal{T},S)$ be the tree scoring function associated with Fitch's parsimony and $s_{hsj,\alpha}(\mathcal{T},S)$ be the tree scoring function that assigns the minimal score to $\mathcal{T}$ if change across branches is computed with d_α. If a set of characters, S, is applicable to all taxa (i.e., there are no secondary characters), then for all α and for all $\mathcal{T}$, $s_F(\mathcal{T},S) = s_{hsj,\alpha}(\mathcal{T},S)$.*

Algorithm 1: Optimal tree score using dissimilarity metrics

```
  Input: Tree, T, with root r, character sequence, S,
         and metric, m.
  Output: Optimal score for the T and S.
1 Initialization: For each leaf, l, set L(l) to its
                   character sequence in S,
                   a(l) := 0, and p(l) := 0.
2 Let L(n) be the first-pass Fitch labeling for node n.
3 for n in a post-order traversal of T:
4     if n is not a leaf:
5         Let c1 and c2 be the children of n.
6         Let a(n) := min( a(c1)+a(c2),
                           a(c1)+p(c2)+1,
                           p(c1)+a(c2)+1,
                           p(c1)+p(c2)+2 )
7         Let p(n) := min( a(c1)+a(c2)+2,
                           m(L(n),L(c2))
                           +a(c1)+p(c2)+1,
                           m(L(n),L(c1))
                           +p(c1)+a(c2)+1,
                           m(L(n),L(c1))
                           +(m(L(n),L(c2))
                           +p(c1)+p(c2) )
8         Update L(n) to L(c1) or L(c2) if it yielded
              the minimal score for p(n).
9 Return min(a(r),p(r)).
```

For the remaining blocks that consist of a controlling primary character and the associated secondary characters, we apply Algorithm 1. In lines 1 and 2, we initialize the scores for the leaves and make a preprocessing step of labeling the secondary characters at each node by that computed in the first pass of Fitch's algorithm (Fitch, 1971), treating the inapplicable states (often coded as "-" in the matrix) as a separate character state. After the initialization is done, we traverse the tree in postorder (lines 3–8), and compute the score for each nonleaf node, n, for the controlling character state being absent (stored in $a(n)$) and for it being present (stored in $p(n)$). To simplify the computation, we multiply scores by the length of the sequence. We use the scores of the children nodes to compute the minimum possible score for the parent node when its primary character is absent and the minimum possible score for when its primary character is present. When the character state for a node and for a child are both absent, there is no mismatch, and the contribution is the score of the child node. If the character states disagree, then we are in the second clause of Equation A.1 and the mismatch contributes 1. If the primary character is present for both, the mismatch contributes the distance between the two sequences, computed by the dissimilarity metric, multiplied by the length. We use the first clause of Equation A.1, the dissimilarity metric, since there is only one primary and m secondaries, $d_\alpha(i,j) = (1 - S_\alpha(i,j)) = \alpha\frac{d}{m}$, where m is the number of secondary characters, s is the number of matches of the secondary characters, and $d = m - s$ is the number of secondary characters that do not match. After the postorder traversal is complete, we return the minimum of the scores of the root, r (i.e., $\min(a(r), p(r))$). We now prove that our algorithm returns the minimal score on a block of a controlling primary and associated secondary characters:

Theorem 2. *Given a fixed tree, $\mathcal{T}$, with character sequence S consisting of a single primary character and a sequence of secondary characters that only apply if the primary character is present, our algorithm returns the minimal score for $\mathcal{T}$.*

Proof: Towards a contradiction, assume that *SL* is a minimal scoring labeling that gives a strictly lower score than our algorithm and that SC is the numerical score assigned to each node under that labeling. Let $a(n)$ and $p(n)$ be the scores assigned and $L(n)$ be the label to each node by Algorithm 1. Let n be a node with the fewest descendants that has a lower score under *SC* than that computed by our algorithm. We have:

$$SC(n) < \min(a(n), p(n)).$$

Since the labels of leaves are given as input, n must be an internal node with children nodes, say c_1 and c_2. By the hypothesis that n has this property and no node with fewer descendants does, we have that c_1 and c_2 under the

SL algorithm have labels and scores computed by our algorithm. The score for the *SL* algorithm is:

$$\begin{aligned} SC(n) = {} & d_\alpha(SL(n), SL(c_1)) + SC(c_1) \\ & + d_\alpha(SL(n), SL(c_2)) + SC(c_2) \end{aligned}$$

Case 1: Assume $a(n) < p(n)$. Then, $SC(n) < a(n)$. If $SL(n)$ assigns the primary character state to be absent, then $SC(n)$ is one of the following:

- $SC(c_1) + SC(c_2)$ if the primary character state is absent for both c_1 and c_2,
- $SC(c_1) + SC(c_2) + 1$ if the primary character state is absent for c_1 and present for c_2,
- $SC(c_1) + SC(c_2) + 1$ if the primary character state is present for c_1 and absent for c_2, or
- $SC(c_1) + SC(c_2) + 2$ if both the primary character state is present for both c_1 and c_2.

Since, by hypothesis, the nodes c_1 and c_2 under the *SL* algorithm have labels (and scores) computed by our algorithm, $SC(n)$ takes one of the following values:

$$\begin{aligned} & \{a(c_1) + a(c_2), a(c_1) + p(c_2) + 1, p(c_1) + a(c_2) + 1, \\ & p(c_1) + p(c_2) + 2\} \end{aligned}$$

By definition, $a(n)$ computes the minimal score when the primary character is absent:

$$\begin{aligned} a(n) = \min(& a(c_1) + a(c_2), 1 + a(c_1) + p(c_2), 1 + p(c_1) + a(c_2), \\ & 2 + p(c_1) + p(c_2)) \end{aligned}$$

so, $a(n)$ is less than or equal to all the possible values of $SC(n)$ which contradicts the assumption $SC(n) < a(n)$.

We assumed that $SL(n)$ has assigned the primary character state to be absent. Instead, assume that it is present and $a(n) < p(n)$. $SC(n)$ takes one of the following values:

- $SC(c_1) + SC(c_2) + 2 = a(c_1) + a(c_2) + 2 \geq a(n)$ if the primary character state is absent for both c_1 and c_2,
- $SC(c_1) + SC(c_2) + 1 + d_\alpha(SL(n), SL(c_2)) = a(c_1) + p(c_2) + 1 + d_\alpha(SL(n), L(c_2)) \geq a(n)$ if the primary character state is absent for c_1 and present for c_2,
- $SC(c_1) + SC(c_2) + 1 + d_\alpha(SL(n), SL(c_1)) = p(c_1) + a(c_2) + d_\alpha(SL(n), L(c_1)) + 1 \geq a(n)$ if the primary character state is present for c_1 and absent for c_2, or
- $SC(c_1) + SC(c_2) + d_\alpha(SL(n), SL(c_1)) + d_\alpha(SL(n), SL(c_2)) = p(c_1) + p(c_2) + d_\alpha(SL(n), L(c_1)) + d_\alpha(SL(n), L(c_2))$ if both the primary character state is present for both c_1 and c_2.

In the first three cases, $SC(n) \geq a(n)$ which contradicts the hypothesis.

If $SC(n)$ takes the fourth value, then, unlike the previous cases, more work is needed since $a(n) \leq a(c_1) + p(c_2) + 2$ which does not necessarily imply $a(n) \leq SC(n)$. But, by definition, $p(n) \leq p(c_1) + p(c_2) + d_\alpha(SL(n), L(c_1)) + d_\alpha(SL(n), L(c_2))$. This implies $p(n) \leq SC(n) < a(n)$ which contradicts the assumption that $a(n) < p(n)$.

Case 2: Consider the second case: $SC(n) < p(n) \leq a(n)$. If scoring labeling function, $SL(n)$, assigns the primary character state to be absent, $SC(n)$ is one of the following:

$$\begin{aligned} & \{a(c_1) + a(c_2), a(c_1) + p(c_2) + 1, p(c_1) + a(c_2) + 1, \\ & p(c_1) + p(c_2) + 2\}. \end{aligned}$$

As in the case above, $a(n) \leq SC(n)$. But, $SC(n) < p(n) \leq a(n)$ yields a contradiction.

So, assume that that $SL(n)$ assigns the primary character state to be present. $SC(n)$ takes one of the following values:

- $SC(c_1) + SC(c_2) + 2 = a(c_1) + a(c_2) + 2$ if the primary character state is absent for both c_1 and c_2,
- $SC(c_1) + SC(c_2) + 1 + d_\alpha(SL(n), SL(c_2)) = a(c_1) + p(c_2) + 1 + d_\alpha(SL(n), L(c_2))$ if the primary character state is absent for c_1 and present for c_2,
- $SC(c_1) + SC(c_2) + 1 + d_\alpha(SL(n), SL(c_1)) = p(c_1) + a(c_2) + 1 + d_\alpha(SL(n), L(c_1))$ if the primary character state is present for c_1 and absent for c_2, or
- $SC(c_1) + SC(c_2) + d_\alpha(SL(n), SL(c_1)) + d_\alpha(SL(n), SL(c_2)) = p(c_1) + p(c_2) + d_\alpha(SL(n), L(c_1)) + d_\alpha(SL(n), L(c_2))$ if both the primary character state is present for both c_1 and c_2.

If $SC(n)$ takes the first value, $a(c_1) + a(c_2) + 2$, then, by definition, $p(n) \leq a(n) \leq a(c_1) + a(c_2) < a(c_1) + a(c_2) + 2 = SC(n)$ which contradicts the hypothesis.

If $SC(n)$ takes the second value, $a(c_1) + p(c_2) + 1 + d_\alpha(SL(n), SL(c_2))$, then to show $SC(n) < p(n)$ it suffices to show that

$$\begin{aligned} a(c_1) + p(c_2) + 1 + d_\alpha(SL(n), SL(c_2)) < a(c_1) + p(c_2) + 1 + \\ d_\alpha(L(n), L(c_2)) \end{aligned}$$

which simplifies to $d_\alpha(SL(n), SL(c_2)) < d_\alpha(L(n), L(c_2))$. Or equivalently,

$$1 - [(1-\alpha) \cdot 1 + \alpha \cdot \frac{s'}{m}] < 1 - [(1-\alpha) \cdot 1 + \alpha \cdot \frac{s''}{m}],$$

where s' is the number of characters in agreement between $SL(n)$ and $SL(c_2)$ and s'' is the number of characters in agreement between $L(n)$ and $L(c_2)$. This simplifies to: $s'' < s'$. By hypothesis, $SL(c_2) = L(c_2)$. If $SL(n) = S(n)$, then $s' = s''$ which is a contradiction. So, assume that the labels $SL(n)$ and $L(n)$ differ. Since we are in the second case where the primary character is absent for c_1, the algorithm sets $L(n)$ to be exactly the label, $L(c_2)$. So, $s'' = 0$, since $0 \leq s' < s''$ this contradicts the hypothesis.

The case where $SC(n)$ takes the third values follows similarly to the case above.

That leaves the case where primary character state is present for both c_1 and c_2 and $SC(n)$ takes the fourth value: $p(c_1)+p(c_2)+d_\alpha(SL(n),L(c_1))+d_\alpha(SL(n),L(c_2))$. If $SL(n)$ assigns the same label to n as our algorithm, then they would return equal values. So, assume that the $SL(n)$ assigns a difference label than our algorithm $L(n)$ does for node n. As in the second case, the different in score reduces to the number of characters in agreement between the label assigned to the n and the children nodes. The label $L(n)$ is computed by the Fitch parsimony algorithm if both are present, so, if $L(c_1)$ and $L(c_2)$ agree at the ith position, then $L(n)$ takes that value and contributes 0 to the sum. Otherwise, the value of $L(n)$ agrees with one of $L(c_1)$ or $L(c_2)$ and disagrees with the other at position i and contributes 1 to the sum. If at some position, i, $SL(n)$ yields a lower score than $L(n)$, then it must agree with both $SL(c_1)$ and $SL(c_2)$ at the ith position while $L(n)$ disagrees with $L(c_1)$ and $L(c_2)$ at the ith position. Since $SL(c_1)=L(c_1)$ and $SL(c_2)=L(c_2)$ this is not possible, and the scores must be equal which contradicts the hypothesis. □

The correctness follows immediately as a corollary to Theorem 2 and Lemma 1:

Corollary 3. *Given a fixed tree, $\mathcal{T}$, with character sequence S, our algorithm returns the minimal score for $\mathcal{T}$.*

We note that the optimal score using dissimilarity metrics can be computed in polynomial time in the number of nodes of the tree:

Theorem 4. *Given a fixed tree, $\mathcal{T}$, on n leaves with character sequence S of length k, the algorithm computes the minimal score in $O(nk)$ time.*

Proof: This follows by analyzing the algorithm in terms of the number of computations done for each branch in the tree and the number of characters. The initialization steps take $O(nk)$ time where n is the number of taxa and k is the number of characters. Step 2 labels the tree via Fitch's algorithm and is also $O(nk)$ time. The algorithm then loops through each branch in the tree. Since there are $O(n)$ branches and the computations in Steps 5–8 take $O(k)$ time, these contribute $O(kn)$ time. Overall, the running time is $O(kn)$. □

Computational Hardness of Finding the Optimal Tree

The computational hardness of finding the optimal tree for a set of data follows almost directly from the original proof of Foulds and Graham (1982) for maximum parsimony. Towards a contradiction, assume that finding the optimal tree for our revised parsimony criteria could be done in polynomial time. For any instance of the traditional parsimony problem, we can embed that instance in one for our problem with all characters set to primary. Since, when working with only primary characters, the branch distances for our approach is the same as traditional parsimony, we could run the polynomial approach for ours, which, would yield a solution to traditional parsimony in polynomial time, contradicting the results of Foulds and Graham (1982).

References

Brazeau M. 2011. Problematic character coding methods in morphology and their effects. Biol. J. Linnean Soc. 104:489–498.

Brazeau M.D., Guillerme T., Smith M. R. 2019. An algorithm for morphological phylogenetic analysis with inapplicable data. Syst. Biol. 68:619–631.

Brazeau M.D., Smith M.R., Guillerme T. 2017. MorphyLib: a library for phylogenetic analysis of categorical trait data with inapplicability. Version 0.0.1-alpha (`http://www.morphyproject.org/`).

Caira J., Jensen K., Healy C. 1999. On the phylogenetic relationships among tetraphyllidean, lecanicephalidean and diphyllidean tapeworm genera. Syst. Parasitol. 42:77–151.

Cannoodt R., Saelens W. 2019. lmds: landmark multi-dimensional scaling. R package version 0.1.0 (`CRAN.R-project.org/package=lmds`).

Chai J., Housworth E.A. 2011. On the number of binary characters needed to recover a phylogeny using maximum parsimony. Bull. Math. Biol. 73:1398–1411.

Cusack M., Williams A., Buckman J.O. 1999. Chemico-structural evolution of linguloid brachiopod shells. Palaeontology 42:799–840.

De Laet J. 2005. Parsimony and the problem of inapplicables in sequence data. In: Albert V.A., editor. Parsimony, phylogeny, and genomics. Oxford: Oxford University Press p. 81–116.

De Laet J. 2015. 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. Cladistics 31:550–567.

de Silva V., Tenenbaum J.B. 2004. Sparse multidimensional scaling using landmark points. Technical Report, Stanford University 41.

Farris J.S. 1970. Methods for computing Wagner trees. Syst. Zool. 19:83–92.

Fernández R., Edgecombe G., Giribet G. 2016. Exploring phylogenetic relationships within Myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. Syst. Biol. 65:871–889.

Fitch W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20:406–416.

Fitzhugh K. 2006. The philosophical basis of character coding for the inference of phylogenetic hypotheses. Zool. Scr. 35:261–286.

Forey P., Kitching I. 2000. Experiments in coding multistate characters. chap. 10. In: Scotland R., Pennington R. T., editors, Homology and systematics: coding characters for phylogenetic analysis. London & New York: Taylor & Francis, p. 266–290.

Foulds L.R., Graham R.L. 1982. The Steiner problem in phylogeny is NP-complete. Adv. Appl. Math. 3:43–49.

Gower J. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–871.

Hawkins J.A., Hughes C.E., Scotland R.W. 1997. Primary homology assessment, characters and character states. Cladistics 13:275–283.

Hopkins M.J., St. John K. 2018. A new family of dissimilarity metrics for discrete character matrices that include inapplicable characters and its importance for disparity studies. Proc. R. Soc. B 285:20181784.

Kearney M., Clark J.M. 2003. Problems due to missing data in phylogenetic analyses including fossils: a critical review. J. Vertebrate Paleontol. 23:263–274.

Kendrick W. 1965. Complexity and dependence in computer taxonomy. Taxon 14:141–154.

Kendrick W., Proctor J. 1964. Computer taxonomy in the Fungi Imperfecti. Can. J. Bot. 42:65–88.

Lee D.-C., Bryant H.N. 1999. A reconsideration of the coding of inapplicable characters: assumptions and problems. Cladistics 15:373–378.

Legendre P., Legendre L. 2012. Numerical ecology. Oxford, UK: Elsevier.

Lloyd G. 2016. Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. Biol. J. Linnean Soc. 118:131–151.

Lloyd G.T. 2018. Journeys through discrete-character morphospace: synthesizing phylogeny, tempo, and disparity. Palaeontology 61:637–645.

Lockhart W., Koenig K. 1965. Use of secondary data in numerical taxonomy of the genus Erwinia. J. Bacteriol. 90:1638–1644.

Maddison W. 1993. Missing data versus missing characters in phylogenetic analysis. Syst. Biol. 42:576–581.

McNeill J. 1972. The hierarchical ordering of characters as a solution to the dependent character problem in numerical taxonomy. Taxon 21:71–82.

Pimentel R.A., Riggins R. 1988. The nature of cladistic data. Cladistics 3:201–209.

Platnick N.I., Griswold C.E., Coddington J.A. 1991. On missing entries in cladistic analysis. Cladistics 7:337–343.

Pleijel F. 1995. On character coding for phylogeny reconstruction. Cladistics 11:309–315.

R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria (`www.R-project.org`).

Radel D., Sand A., Steel M. 2013. Hide and seek: placing and finding an optimal tree for thousands of homoplasy-rich sequences. Mol. Phylogenet. Evol. 69:1186–1189.

Schliep K. P. 2010. phangorn: phylogenetic analysis in r. Bioinformatics 27:592–593.

Simmons N.B. 1993. The importance of methods: archontan phylogeny and cladistic analysis of morphological data. In: MacPhee R.D.E., editor. Primates and their relatives in phylogenetic perspective. New York: Springer. p. 1–61.

Smith M.R. 2018. TreeSearch: phylogenetic tree search using custom optimality criteria. R package version 0.4.0.

Smith M.R. 2019. TreeTools: create, modify and analyse phylogenetic trees. R package version 1.0.0.

Strong E.E., Lipscomb D. 1999. Character coding and inapplicable data. Cladistics 15:363–371.

Swofford D. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.

Swofford D. 2020. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0a167. (`paup.phylosolutions.com`).

Tarasov S. 2019. Integration of anatomy ontologies and evo-devo using structured markov models suggests a new framework for modeling discrete phenotypic traits. Syst. Biol. 68:698–716.

Wilkinson M. 1995. A comparison of two methods of character construction. Cladistics 11:297–308.

Wills M. 2001. Morphological disparity: a primer. In: Adrain J., Edgecombe G., Lieberman B, editors. Fossils, phylogeny, and form. New York: Springer. pp. 55–144.