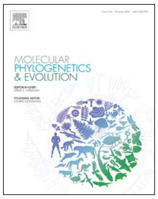
Molecular Phylogenetics and Evolution 122 (2018) 95–101

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev



On defining a unique phylogenetic tree with homoplastic characters

Pablo A. Goloboff[a,*], Mark Wilkinson[b]

[a] Unidad Ejecutora Lillo, Fundación Miguel Lillo, CONICET, Miguel Lillo 251, 4000 San Miguel de Tucumán, Argentina
[b] Department of Life Sciences, Natural History Museum, London SW7 5BD, United Kingdom

ARTICLE INFO

Keywords:
Phylogeny
Homoplasy
Parsimony
Likelihood
Tree-searches

ABSTRACT

This paper discusses the problem of whether creating a matrix with all the character state combinations that have a fixed number of steps (or extra steps) on a given tree *T*, produces the same tree *T* when analyzed with maximum parsimony or maximum likelihood. Exhaustive enumeration of cases up to 20 taxa for binary characters, and up to 12 taxa for 4-state characters, shows that the same tree is recovered (as unique most likely or most parsimonious tree) as long as the number of extra steps is within 1/4 of the number of taxa. This dependence, 1/4 of the number of taxa, is discussed with a general argumentation, in terms of the spread of the character changes on the tree used to select character state distributions. The present finding allows creating matrices which have as much homoplasy as possible for the most parsimonious or likely tree to be predictable, and examination of these matrices with hill-climbing search algorithms provides additional evidence on the (lack of a) necessary relationship between homoplasy and the ability of search methods to find optimal trees.

1. Introduction

This paper explores the problem of whether all (or some) of the possible character state distributions with different numbers of steps on a given tree, produce a matrix for which that tree is the unique most parsimonious or likely tree. The results shed some additional light on problems recently discussed by Radel et al. (2013) and Goloboff (2014), connects to results of Goloboff (1991) and Steel and Charleston (1995), and illustrates both similarities and differences between two important criteria for phylogenetic reconstruction, maximum parsimony (MP) and maximum likelihood (ML).

The most common type of discrete morphological character are binary characters. This paper is primarily concerned with discrete binary data, for which it is easier to establish conclusions; discussion of DNA sequence data is included at the end. While ML methods are most commonly used for DNA sequence data, and discrete morphological characters are usually analyzed with parsimony, the Mkv model (Lewis, 2001) is being used increasingly for ML analysis of morphological data, and is implemented in recent releases of MrBayes (Ronquist et al., 2012) and PAUP* (Swofford, 2002). Thus, the behaviour of the Mkv model is studied also for the categorical (binary) datasets. Our main finding is that it is possible to univocally define any phylogenetic tree (both under ML and MP) by using characters with much larger amounts of homoplasy than allowed by previous methods to generate datasets with known optimal trees (e.g., Chai and Housworth, 2011).

2. Matrices with all *s*-step combinations

For any binary tree *T* with *t* leaves, there is a number *n* of ways to assign states to the terminal branches such that the resulting character has a number *s* of steps. For binary characters, this number *n* does not depend on the shape of *T*, only on *t* and *s* (see formula in Steel and Charleston, 1995: 370). Let $M_{T,s}$ be the matrix with a copy of every possible character-state distribution with exactly *s* steps on tree *T*; define $M_{T,e}$ analogously, but for extra steps *e* instead of absolute steps *s*. Note that in non-constant binary characters $e = s{-}1$, so that $M_{T,s=1} = M_{T,e=0}$, but this does not hold for 3 or more states. Let *P* and *L* be respectively the (set of) most parsimonious and likely tree(s) for the set of characters in question.

It is well known that (for binary characters and any value of *t*), when $s = 1$, then *P* equals *T*: $M_{T,s=1}$ is the Baum-Ragan "matrix-representation with parsimony" (or MRP) of tree *T* (Baum, 1992; Ragan, 1992). For the MRP, *T* is also the unique most likely tree *L*, under the Mkv model of Lewis (2001; in this paper, all the ML Mkv analyses were carried out with PAUP*, vers. 4.0a151, with *lset nstates = mkv mkstatespace = fixed*).

The problem of whether *T* can be uniquely retrieved from the matrix $M_{T,s}$when $s > 1$ has not been so far considered in the literature. That is, the problem of whether $T = P$, $T \neq P$, or $T \subset P$ (i.e. *T* is a most parsimonious tree, but not the only one, so that $M_{T,s}$ does not univocally define *P*). In the case of $t = 4$, it is easy to verify that for $M_{T,s=2}$, $P \neq T$ for any topology *T*. The length of *T* in that case is $ns = 4$ (2 characters of

[*] Corresponding author.
E-mail addresses: pablogolo@csnat.unt.edu.ar, pablogolo@yahoo.com.ar (P.A. Goloboff).

https://doi.org/10.1016/j.ympev.2018.01.020
Received 26 June 2017; Received in revised form 30 December 2017; Accepted 26 January 2018
Available online 31 January 2018

Table 1
Results of maximum parsimony (MP) analysis for matrices containing all the possible binary character with different numbers ***s*** of steps for each of the tree-shapes, for different numbers of taxa. Cases where the matrix for every tree-shape produces the same tree T as single most parsimonious tree P (i.e. $T = P$) are indicated with +; cases where every one of the shapes produces most parsimonious trees different from T are indicated with a dash (–). In the rest of cases, the number of shapes for which $T = P$ is indicated first, followed by the number of shapes where T is one among multiple equally parsimonious trees, ending with the number of shapes where T is not a most parsimonious for the corresponding matrix. Cases marked as (*) were checked with a random sample of 200 shapes instead of exhaustively.

TAXA	s = 2	s = 3	s = 4	s = 5	s = 6	s = 7	s = 8	s = 9
4	–							
5	–							
6	–	–						
7	0,3,3	–						
8	10,0,1	–	–					
9	+	–	–					
10	+	–	–	–				
11	+	80,3,15	–	–				
12	+	+	–	–	–			
13	+	+	–	– –	–			
14	+	+	813,0,170	–	–	–		
15	+	+	+	–	–	–		
16	+	+	+	–	–	–	–	
17	+	+	+	8911,0,1994	–	–	–	
18	+	+	+	+	–	–	–	–
19	+	+	+	+	–	–	–	–
20	+	+	+	+	102676,0,25236	– (*)	– (*)	– (*)

2 steps each), but the length of P is 3 (i.e. there are two most parsimonious trees, where one of the two characters perfectly matches a group).

For the case of $M_{T,s=2}$ and $t = 5$, predicting whether $T = P$ is harder. A simple TNT script (*testequality.run*, available as Supplementary Material) can be used to generate the matrix $M_{T,s}$ for each of the tree-shapes (2 in the case of $t = 5$), and check whether $T = P$. The results are shown in Table 1. While the number n of binary characters with s steps does not depend on the shape of T, whether $T = P$ may well depend on the shape (see Felsenstein, 2004: 29-32 for a general discussion of rooted tree-shapes, and Goloboff et al., 2017 for a recent application). However, the equality between T and P will hold (or not), for every one of the trees of a given shape, thus simplifying the task of checking whether $T = P$ for each of the possible trees for t taxa –it is only necessary to check every tree-shape, not every tree (specially convenient, since the number of shapes is vastly smaller than the number of possible tree topologies; see, e.g. Felsenstein, 2004: 30, Table 3.3). The script used for this paper automatically generates $M_{T,s}$ for each of the relevant tree-shapes for t taxa and checks whether $T = P$, using TNT (Goloboff et al., 2008; Goloboff and Catalano, 2016) to find P. The script generates all the rooted shapes for t−1 taxa, since TNT always uses one of the taxa as root or outgroup; rooting the trees differently would change neither the MP nor ML scores. For both tree-shapes for $t = 5$ and $s = 2$, $T \neq P$.

Is it possible that $T = P$, for some number t, and $s = 2$? As t increases to 8, then for 10 of the 11 possible tree-shapes, $T = P$. For the 11th shape (the shape that results from rooting the fully symmetrical 8-taxon tree in one of the terminal taxa), $T \neq P$. This shows that it is indeed possible for P to be identical to T for some shapes, and different for others. For $t \geq 9$ and $s = 2$, it can be verified that $T = P$ for every tree-shape.

What about larger values of s? The number of steps a binary character can have on a most-parsimonious reconstruction is bounded by the number of 0s and 1s (whichever is minimum; the maximum number of steps an MP reconstruction can have is the integer $\frac{t}{2}$). Thus, only for $t \geq 6$ there are some characters with $s = 3$. For $t = 6$ to $t = 10$, and $s = 3$, $T \neq P$. For $t = 11$, some tree-shapes produce MRP matrices for which $T = P$, but others produce matrices where T is not a shortest tree (i.e. $T \neq P$) or where T is simply one of multiple MPTs (i.e. $T \subset P$) (see Table 1). When $s = 3$, $T = P$ for every possible tree-shape only if $t \geq 12$.

The previous case, $s = 3$, fulfilled the condition of $T = P$ for every possible tree-shape for all cases where $s \leq 0.25t$. For larger values of t and s, it can be verified that the same is true; for example, $M_{T,s=5}$ when $t = 20$ fulfills the condition that $T = P$ for every tree T, despite the fact that all the characters have significant amounts of homoplasy. The only exception to the relationship $s \leq 0.25t$ is the case of $t = 8$, where one of the 11 shapes does not fulfill $T = P$. Some cases where $s > 0.25t$ have some shapes producing a matrix $M_{T,s}$ for which T is a most parsimonious tree (i.e. $T \subset P$ when $s > 0.25t$), but there is no case where $s \geq 0.3t$for which T is a most parsimonious tree (i.e. $T \neq P$ for all cases of $s \geq 0.3t$).

For $t > 20$, exhaustively checking whether all tree-shapes fulfill the condition that $T = P$ becomes difficult. First, the script works by generating all possible binary state distributions (2^{t-1}, given that the root only has 0 states), and for $s = 6$, T could be expected to equal P for all shapes only when $t \geq 24$–this requires generating matrices with 8,388,608 characters or more. Second, the number of shapes also increases rapidly; for $t = 24$, the number of shapes is 3,626,149, and the 990,080 characters with 6 steps on each of those shapes would have to be identified (deactivating the rest), followed by a search for P. Taking samples of the shapes, for t up to 28 (the maximum matrix size we could handle with TNT), suggests that the same relationship between t and s (i.e. $s \leq 0.25t$) continues being valid for T to be equivalent to P.

The same bounds seem to hold for ML under the Mkv model, regarding the most likely tree L (i.e. $T = L$ for all trees when $s \leq 0.25t$, and $T \neq L$ for all trees when $s \geq 0.3t$). When $0.25 < s < 0.3$, the cases where $T \neq L$ or $T \subset L$ seem to be more common than the cases where $T \neq P$ or $T \subset P$; there are, however, some cases where $T \neq P$ or $T \subset P$ while $T = L$.

Note that, for these matrices, application of implied weighting (which downweights characters with more homoplasy; Goloboff, 1993), is guaranteed to produce exactly the same results as equal weights MP, given that all the characters have the same homoplasy.

3. Undecisive matrices

Of course, adding possible binary state distributions with fewer steps than $0.25t$ also produces matrices where $T = P$ and $T = L$. However, adding those possible characters does not make it possible to also add at the same time characters with $s > 0.25t$ and still have $T = P$. In the end, if all the characters $(1, \leq ,s, \leq ,\frac{t}{2})$ are included in the matrix, the matrix becomes completely undecisive (Goloboff, 1991): every possible partition of the data is represented exactly once, and thus every possible (binary) tree is equally parsimonious (making $T \subset P$). The amount of homoplasy in undecisive matrices (with average character length more than $0.3t$ for $t \geq 7$) is larger than in the matrices for which $s = 0.25t$ explored in the preceding section; but in this case P (the set of all binary trees) does not equal T. Given that the number n of characters with s steps for t taxa is constant (because, from Steel and Charleston, 1995, n depends only on s and t, not on tree-shape), then all the possible trees have the same numbers of characters with 1, 2, … $\frac{t}{2}$ steps, and therefore these matrices are also fully undecisive under implied weighting, regardless of weighting strength (i.e. regardless of concavity constant $\boldsymbol{k}$; see Goloboff, 1993 for discussion of $\boldsymbol{k}$).

These matrices, undecisive under MP, also seem to be fully undecisive under ML with the Mkv model: every possible tree has exactly the same likelihood, as calculated by PAUP*. The undecisiveness, however, only occurs when the possible state space is fixed at 2 for all characters. When considering a state space of 4 (e.g., replacing the 0s and 1s by 2s and 3s in odd-numbered characters, which still defines the same partitions, and continues being undecisive for parsimony), some tree-shapes have a worse likelihood than others. Fully pectinate trees are always among the trees of worst scores; the score difference among trees increases with number of taxa; as the number of characters for the matrix to be undecisive under MP or 2-state Mkv grows rapidly, the maximum number of taxa which could be checked with PAUP* was 16 –in that case, the score difference between best and worst trees (from 2706.17×10^3 to 2708.36×10^3) approaches significance (according to the *khtest* = *normal* of PAUP*) at the 10% level (P = .1313).

Using a state space of 4 instead of 2 implies that, as branch length increases, the probability of state transformations (as well as stasis) along the branch converges to 0.25 instead of 0.5 –i.e. a ratio of 1:2. This would suggest that the results for 2- and 4-state spaces would be equivalent, simply with correspondingly lower likelihoods. However, for very short branches, the probability of stasis (i.e. the same state for ancestor and descendant) tends to 1 regardless of the size of the state space, and the 1:2 correspondence does not hold for branches of intermediate lengths; the ratio actually varies with branch length. This means that, as character changes are distributed along branches of intermediate (and different) lengths, the correspondence between results for 2-state and 4-state spaces no longer holds. It also means that (in empirical Mkv analyses) whether states not recorded in the matrix exist may make a difference, even when those states would otherwise be "uninformative" (e.g., restricted to a single taxon).

Parsimony is often used to combine trees (e.g., producing supertrees). In this context, it is important to note that a fully undecisive MRP matrix will result from single-group trees that represent each of the possible taxon partitions, not from all possible binary trees for the taxa at hand. For 6 taxa, combining the 105 possible trees for 6 taxa in an MRP, does not produce the same 105 trees as result. In the 105 possible binary trees for 6 taxa, every 2-taxon groups occurs in 15 of those, while 3-taxon groups occur in only 9. The imbalance in the representation of these characters leads to an MRP matrix for which some trees are shorter than others (producing, for 6 taxa, only 15 distinct trees, instead of the 105 that would result from full undecisiveness).

4. Spread of changes in $\boldsymbol{M_{T,s}}$

A character with homoplasy on some tree, from the perspective of parsimony, provides evidence *against* the tree –i.e. there is necessarily an alternative tree with a better fit for that character. How is it possible that selecting precisely those characters that seem to provide evidence against the tree, uniquely determines that tree, both under MP and ML, and that (as long as t increases) this can continue happening for larger and larger numbers of extra steps?

Fig. 1a-b illustrates the situation of $s = 2$, showing some of the characters that can support two of the groups in the tree (clades numbered as 9 and 11). Fig. 1a shows the characters that (at least in some MP reconstructions) have a change at branch 9 (marked with a white square). The grey and black squares indicate other branches of the tree where there could be changes such that the character has two steps. Once a change occurs on branch 9, the second change could occur in 5 branches (0, 2, 3, 7, 12, 13, marked with black) such that the character can be unambiguously mapped as synapomorphies of the two groups; it could also occur in 2 other branches (6, 11, marked in grey) such that the character provides no clear-cut synapomorphy (i.e. the famous acctran-deltran case). Fig. 1b shows the case of a character occuring in branch 11; even when the branches where the second change could be located such that the MP mapping is ambiguous (branches 2, 3, 4, 7, 9) outnumber the branches where the MP mapping is unambiguous (branches 0, 1, 5), there are still three unambiguous synapomorphies for the clade corresponding to branch 11.

What is more important: for $s = 2$, the second change can occur in a number of alternative locations, such that an alternative tree would better explain that character, but *individually*: those alternative locations result in characters that are *in conflict with each other*. This conflict between subsequent character changes is related to the condition $s \leq 0.25t$. As s is smaller relative to t, for each change to be acting as a synapomorphy of a branch, the alternative changes in the other characters can be spread more evenly on the tree (many of those, also providing synapomorphies for some clades). The empirically observed factor ¼ results evidently from the fact that changes, to be unambiguous, have to be distributed (a necessary but not sufficient condition, see next section) on branches that are at least 3 nodes apart –more than $0.25t$ changes will clutter the tree, and synapomorphies will not outnumber ambiguous changes. That the shape corresponding to the symmetric 8-taxon tree is the only case where $s = 0.25t$ does not result in $T = P$ reinforces this idea, for in that case (and only in that case) no branch can be 3 or more nodes apart from the middle branches.

5. Large datasets, lots of homoplasy

Given that the characters providing unambiguous synapomorphies (and with parallel changes conflicting in turn with each other) are a significant majority, taking a sample of characters from $M_{T,s}$, forming a subset $M'_{T,s}$, generally leads to $T = P$ for all tree-shapes. The fraction of characters from $M_{T,s}$ that need to be included for this to be the case is smaller as s is smaller relative to t (of course, when $s = 1$, every single character must be included, or the tree will have unresolved groups, so that $T \subset P$).

The fact that only some characters with s steps need to be included in $M'_{T,s}$ for $T = P$, and the consideration of the branch distances between changes for the sample of characters to provide unambiguous support for T, suggests a way to create matrices that have large amounts of homoplasy, but for which the optimal tree P can be known in advance with (almost) certainty. Matrices for which P is known are useful to study the ability of tree-search methods to actually find the optimal tree. Finding the optimal tree, for both MP and ML, is a difficult, NP-complete problem (Foulds and Graham, 1982; Roch, 2006), and the problem of tree-searches continues being investigated (e.g., Goloboff and Simmons, 2014; Ford et al., 2015; Goloboff, 2015; St. John, 2016). Radel et al. (2013) were the first to use datasets for which P can be known exactly (generated with Chai and Housworth's, 2011 methods), but where homoplasy is very abundant, to empirically test the effectiveness of heuristic tree-search algorithms; inspired by Radel et al. (2013), Goloboff (2014) also examined datasets for which P can be known exactly but for which standard methods perform poorly (with improved methods performing significantly better).

Phylogenetic programs cannot analyze all the numerous characters with s steps for a large t due to memory limitations, so the fact that subsamples of c characters with s steps also define $T = P$ is specially convenient. The considerations of the previous section, on the distances needed between changes for the characters to support T as P, can be used to generate datasets with very large amounts of homoplasy.

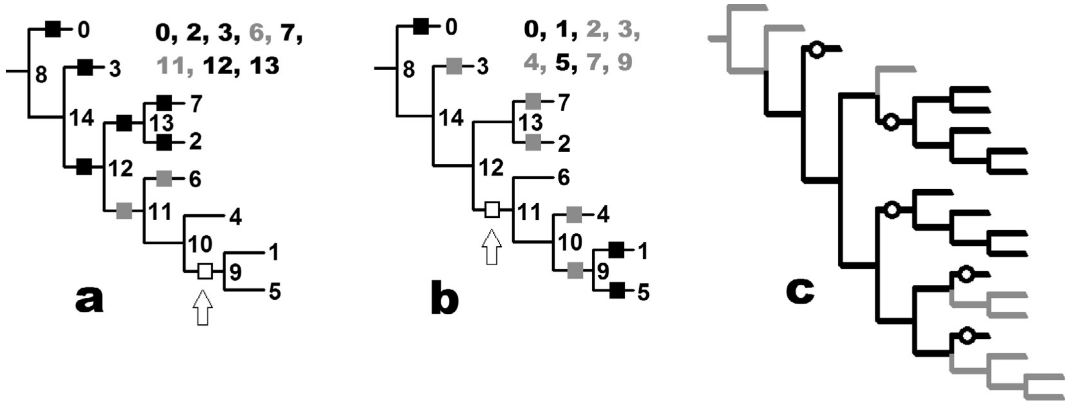


Fig. 1. (a) The possible binary character-state distributions that have 2 steps and a synapomophy at group 9. The second step can occur in two branches (6, 11) such that the optimization is ambiguous, and in 6 branches (0, 2, 3, 7, 12, 13) such that the character provides an unequivocal synapomophy for group 9. (b) The same, for a synapomorphy at group 11. Although there are more places of change producing ambiguous optimizations (2, 3, 4, 7, 9) than unequivocal synapomophies for group 9 (0, 1, 5), the characters resulting from changes at nodes 2, 3, 4, 7 or 9 are also in conflict with each other, so that they cannot simultaneously have a better fit on an alternative tree. (c) A case where placing points of change on the tree, spread 3 nodes apart, still produces character-state distributions in the terminals that can be parsimoniously mapped with fewer steps than points of change. The terminal branches receive a grey or black color depending on the points of change (with initial state grey, for the root); the internal branches receive their color in a most parsimonious optimization (based on the terminal colors). See text for additional discussion.

Guaranteeing that $T = P$ would require that characters are selected as synapomorphies for each (internal) branch of T, and then that the additional changes are uniformly spread along all the branches (in such a way that no single branch concentrates parallelisms on different characters, and that there is a distance of at least 3 nodes between each point of change). A simpler script can generate datasets randomly choosing for each character points of change such that each point is 3 or more nodes from previous points of change. This does not guarantee that $T = P$, but the probability that parallelisms in the characters supporting a branch of T can be concentrated on another branch decreases with the number c of characters, and therefore the probability that $T = P$ asymptotically approaches unity as c grows (and as long, of course, as s remains within $0.25t$). This script (*maxhomo.run*, available in the Supplementary Material) first sets, for each character, the points of change, then sets the root to state 0, and travels the tree in an uppass (inverting the state as it crosses each point of change). After "evolving" the character in this way, the character is mapped onto the tree to ensure that the number s of steps has been effectively obtained (see Fig. 1c for a case where the character can be mapped on the tree with fewer steps than the points of change, even if all points of change are separated by 3 nodes); otherwise, the character is re-evolved by spreading points of change again. As the first points of change are assigned randomly, it may well be that the last points of change cannot be assigned so that they are 3 or more nodes away from existing points of change; in that case, the process is started again from scratch for that character. For every character, the initial point of change is assigned to one of the internal tree branches; after all internal branches have been assigned one such character, the branches start being revisited (so that each branch has several characters that act as "synapomorphies"). Since the rest of the changes gets randomly distributed over the tree, it becomes increasingly improbable –as the number c of characters grows– that several changes can be concentrated on some tree branches (thus leading to a different most parsimonious tree). This informal argument strongly suggests that, for the datasets generated with this TNT script (used in the experiments summarized in Table 2), even when P is not known with certainty, there is a high probability that $T = P$; simulation of many datasets also suggests that this is indeed the case.

In addition to the method just described, of maximum homoplasy, there are 4 additional methods (other than the trivial MRP) to generate a matrix for which P can be known with certainty. Ordered from lowest to highest amounts of homoplasy these 5 methods are:

(a) ***Quad-Char***: datasets where P is uniquely defined by four (i.e. $c = 4$) multistate characters with no homoplasy (Huber et al., 2005; see also Steel and Penny, 2005); T can have any topology.

(b) ***Max-Missing***: datasets with missing entries and no homoplasy, such that each character has only two 0s and two 1s, providing a synapomorphy for each of the branches of T; studied in detail by Goloboff (2014), based on Steel (1992), and Bocker et al. (2000). T can have any topology, $c = t–3$.

(c) ***Modular***: a "fractal" expansion of a pattern of binary characters (Goloboff, 2014) for which P is known; for the pattern expanded to a level p, the resulting dataset has $t = 1 + 4^p$; if c_p is the number of characters for level p (with $c_1 = 9$), then $c_p = (4c_{p-1}) + 9$. These datasets have low amounts of homoplasy (the consistency index for all the expansions of this pattern is 0.6923; see Goloboff, 2014:123) but P can be easily missed by hill-climbing searches.

(d) ***Chai-Housworth***: two types of construct (Chai and Housworth, 2011), composed of the minimum number of binary characters that support a fully pectinate or fully symmetrical tree (and with significant amounts of homoplasy). Radel et al. (2013) studied the symmetrical case in more detail.

(e) ***Max-Homoplasy***: the $M'_{T,s}$ construct defined above; T can have any topology, and each of the characters has $\frac{t}{4}$ steps. For T to be the same as P with high probability, c must be at least $3t$.

The Chai-Housworth datasets Radel et al. (2013) studied in more detail have a much lower average number of steps than the datasets constructed with the Max-Homoplasy method. Symmetrical Chai-Housworth datasets are built by expanding a pattern, based on an integer $p \geq 2$, such that s steps per character are obtained for a matrix with $t = 2^p$ and $c = 4p–3$, with $s = \frac{2^{p+1}-3}{4p-3}$ (see Radel et al., 2013: 1188; note that their formula is for homoplasy, subtracting 1 to the value given here, since these are binary characters). Chai and Housworth's (2011) goal was to find sets of binary characters that would define a uniquely most parsimonious tree with the minimum number of characters, not with the maximum homoplasy; homoplasy is thus high in their datasets, but far from the maximum possible to still have $T = P$. For the same numbers of taxa, generating datasets with the Max-Homoplasy method described here, $s = \frac{t}{4} = \frac{2^p}{4}$, and thus the ratio of average number of steps with the method of Chai-Housworth, is $\frac{t(4p-3)}{2^{p+3}-12}$. This ratio increases with number of taxa; for $p = 4$ (32 taxa), the average character length for Max-Homoplasy datasets is 2.2295 times that of Chai-Housworth datasets, but for $p = 15$ (32,768 taxa) the ratio of average character lengths is 7.1253. While Radel et al. (2013)

Table 2
Results of applying hill-climbing searches to five types of data set where the most parsimonious tree can be known in advance. “Mult = hold 1” is a random addition sequence Wagner tree followed by TBR branch-swapping saving a single tree; “Mult = hold 10” saves up to 10 trees. The probability of finding the minimum length (known in advance) with a search is indicated, as well as the number of rearrangements examined on average by each type of search. Both types of searches, as implemented in TNT. The datasets where finding minimum length is easiest are those with the most homoplasy.

					mult = hold 1		mult = hold 10	
Method	Taxa	Chars	Min. length	Homoplasy (CI)	P (min. length)	Rearrangs per repl.	P(min length).	Rearrangs per repl.
Max-Hom	64	300	4800	Very high (0.0625)	> 99%	250 × 10ˆ3	> 99%	300 × 10ˆ3
Chai-Hous.	64	21	125	High (0.168)	25%	180 × 10ˆ3	62.5%	630 × 10ˆ3
Quad-Char	64	4	61	None (1.00)	1.2%	180 × 10ˆ3	30%	1.2 × 10ˆ6
Modular	65	189	273	Low (0.692)	0.2%	65 × 10ˆ3	0.2%	100 × 10ˆ3
Max-Missing	64	61	61	None (1.00)	≪0.01%	160 × 10ˆ3	≪0.01%	1.3 × 10ˆ6

show (their Theorem 1, p. 1187) that the maximum homoplasy in datasets with binary characters that determine a uniquely most parsimonious tree is unbounded and quickly grows with t (which they illustrate by means of Chai-Housworth datasets), the homoplasy grows more quickly in the present Max-Homoplasy datasets than in Chai-Housworth constructs, and yet it is still bounded by $\frac{t}{4}$ (as otherwise the dataset cannot determine a uniquely most parsimonious tree).

Radel et al. (2013) examined only Chai-Housworth datasets; they considered it remarkable that tree-searches in TNT could consistently find the optimal trees (up to 32,768 taxa), despite the large amounts of homoplasy, which indicates that homoplasy does not necessarily erase “phylogenetic signal” (Radel et al., 2013: 1186–1187). Goloboff (2014) examined Chai-Housworth datasets in more detail, as well as Modular and Max-Missing datasets; the present paper adds Max-Homoplasy and Quad-Char datasets. Of all these, the datasets where tree-search algorithms perform best are, surprisingly, those with the maximum homoplasy (Table 2). The tree-landscape for the homoplasy-free Quad-Char datasets is (for comparable numbers of taxa) even less rugged than for the highly homoplastic Chai-Housworth datasets. Quad-Char datasets still seem to have a single local optimum (when saving numerous multiple trees), but swapping on a single tree often gets stuck in a local optimum –for Max-Homoplasy datasets, even TBR swapping on a single tree finds the optimal tree in the vast majority of cases.

The optimal tree is so well defined by the Max-Homoplasy datasets, that bootstrapping (Felsenstein, 1985) or other measures of group support based on resampling (e.g., Farris et al., 1996; Goloboff et al., 2003) find that all groups of the tree have supports of 99% or more (even when analyzing each resampled data sets with a single random addition sequence followed by TBR without saving multiple trees, which makes it more likely that the optimal tree will be missed, and thus tends to produce lower frequencies for groups actually supported by the data).

This does not mean, of course, that finding optimal trees will always be harder for datasets with less homoplasy –only that the relationship between homoplasy and ease of analysis is not a lineal one, with the distribution of homoplasy among the different terminals coming into play as well when determining the shape of the tree-landscape. In addition, none of the methods (a)-(e) listed above consist of evolving characters on a tree –rather, a tree topology T is being used to select characters in such a way that the optimal tree P follows. In a sense, this is a “model-less” method for generating experimental datasets. One of the obvious difficulties in translating the present results to real biological datasets is that the homoplasy in those cases tends to be concentrated in certain parts of the tree, or in specific characters, or with particular patterns. In addition, all the studies of tree search algorithms based on the empirical performance on real datasets show that even the best tree-search algorithms often have difficulty converging to the same tree-scores for datasets with even a few hundred taxa (e.g., see Goloboff, 1999; Roshan et al., 2004; Goloboff and Pol, 2007, for parsimony-based comparisons).

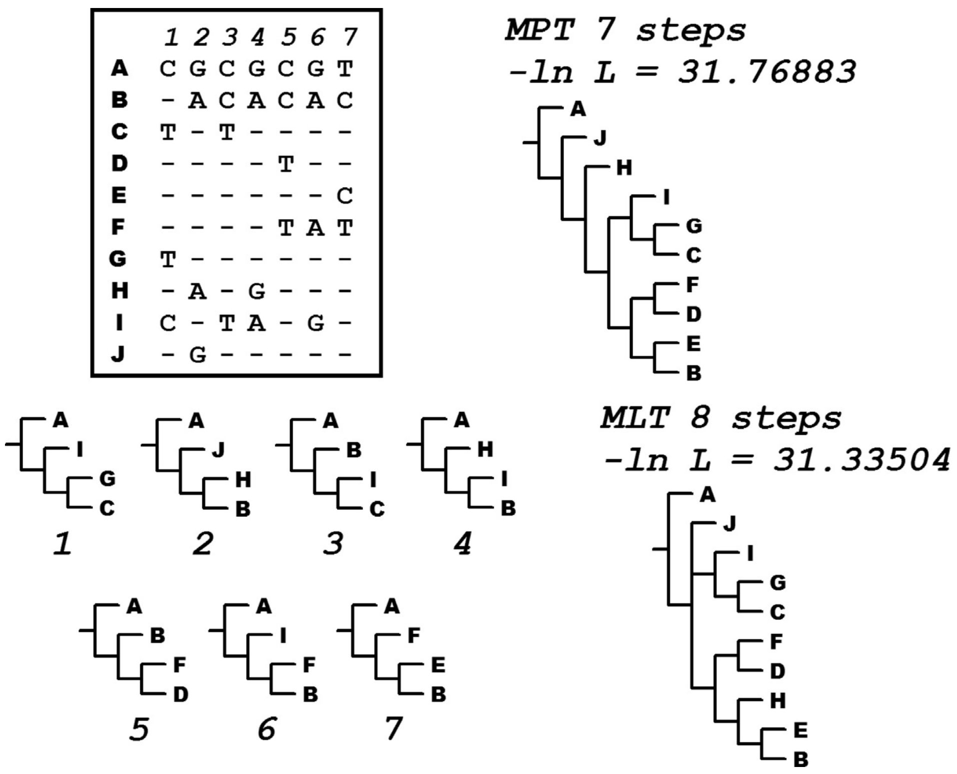
6. Equivalence (or not) under ML

For all the types of dataset examined in the previous section, except Max-Missing, the results obtained under MP or ML are always the same –repeated analysis with PAUP* (e.g., branch-swapping with T as starting point, trying to find trees of better likelihood) suggests that indeed $T = P = L$ for those datasets.

In the case of Max-Missing, which have no homoplasy, for some datasets $P = L$, and for others $P \neq L$. Fig. 2 illustrates one case where $P \neq L$, with 7 compatible characters, each of which determines a 4-taxon tree. All these 4-taxon trees are compatible, and thus P displays all 7 trees as subtrees (with a length of 7 steps). Although the partitions determined by the data are compatible, L is a different tree, 8 steps long (this analysis used the JC69 model, Jukes and Cantor, 1969, as implemented in PAUP*, with short branches collapsed); instead of displaying the subtree (A(H(IB))), L displays the subtree (A(I(BH))). The subtrees corresponding to characters 1–7 do not have conflict in terms of partitions, but their branch lengths cannot be combined in a single tree; thus, ML perceives a “conflict”, which in terms of the parameteres used by the method (i.e. branch lengths) is legitimate. Such a situation illustrates possible problems of using quartets to reconstruct the ML tree (e.g., Schmidt et al., 2002; Yang et al., 2014); optimality cannot be guaranteed even in the case of fully compatible quartets. The example also shows that partitions and homoplasy are not necessarily comparable in MP and ML, and that the latter method –if so required for better fitting branch lengths– may prefer trees that require some homoplasy for the data, even when some other tree would allow fitting the data with no homoplasy whatsoever.

7. DNA sequences

The previous analyses used all the n combinations (or a sample thereof) of the binary characters that have s steps on a given tree T. For characters with more than two states, the number n of characters with s steps changes with tree-shape (Steel and Charleston, 1995). It is possible that matrices which determine that $T = P$ can be constructed with even larger amounts of homoplasy, when the characters can take more than 2 alternative states. For example, the maximum amount of homoplasy in a binary character is the integer $e_{max} = \frac{1}{2}t{-}1$, but for a character with 4 states it is $e_{max} = \frac{3}{4}t{-}3$, a larger number, suggesting the possibility that larger amounts of homoplasy in 4-state characters still allow constructing matrices for which $T = P$. However, this seems not to be the case, based on the results shown in Table 3. Table 3 summarizes the results of creating, for up to 12 taxa, the matrix $M_{T,s}$ for each tree-shape (as for binary characters, $M_{T,s}$was created by deactivating in turn all the characters except those having s steps, from a matrix containing all 4^{t-1}characters for t taxa, with the first taxon having the first state for each character; this is 4,194,304 characters for 12 taxa). In the case of 4-state characters for low numbers of taxa, $T = P$ for $s \leq 0.5t$ –twice the number of steps as in binary characters! But the



	1	2	3	4	5	6	7
A	C	G	C	G	C	G	T
B	-	A	C	A	C	A	C
C	T	-	T	-	-	-	-
D	-	-	-	-	T	-	-
E	-	-	-	-	-	-	C
F	-	-	-	-	T	A	T
G	T	-	-	-	-	-	-
H	-	A	-	G	-	-	-
I	C	-	T	A	-	G	-
J	-	G	-	-	-	-	-

Fig. 2. A case where 7 compatible characters (i.e. defining compatible subtrees), producing a unique most parsimonious tree (of 7 steps and no homoplasy), produce a different tree (of 8 steps) under maximum likelihood (with the JC69 model). See text for additional discussion.

relationship between T and P follows a pattern similar to that for binary characters as t and homoplasy (i.e. extra steps, or number of steps beyond the minimum possible) grow. In the case of binary characters, the homoplastic steps always equal the number of steps minus 1; such is not the case for 4-state characters, because different characters may have the same value of s steps, but a different minimum, thus a different value of homoplastic steps e. For matrices including all characters with e steps of homoplasy, the cases where $T = P$ for all tree-shapes continue being those in which $e \leq 0.25t$. But a character with 4 states will have a minimum possible number of 3 steps at the most, so that $s \leq e + 3$, implying that the total number of steps s for $T = P$ in 4-state characters must be $s \leq 0.25t + 3$ (note that the case of $t = 12$, shown in Table 3, for which $T = P$ for all tree-shapes when the number of steps $s = 6$, obeys this inequality). For a large enough t, the inequality for 4-state characters $s \leq 0.25t + 3$ is in practice the same as that for binary characters, $s \leq 0.25t$, which is to say that both the maximum possible numbers of steps, and the maximum possible homoplasy, for T to be unambiguously recovered from the corresponding matrix, are similar for binary and for 4-state characters.

As in the case of binary characters, ML produces results that are similar (overall) to those of MP, but not exactly identical. Testing cases with significant numbers of taxa using PAUP* (with the JC69 model) becomes difficult, for tree searches proceed extremely slowly (due to the large numbers of characters). The largest case we could analyze exhaustively is for 8 taxa. ML on $M_{T,e=1}$ produces $T = L$ for 7 of the 11 possible tree-shapes, but $T \neq L$ for the remaining 4 (when MP, see Table 3, produces $T = P$ for each of the tree-shapes). Thus, using 4-state instead of binary data makes parsimony *more* able to recover T from $M_{T,e=1}$ (with all shapes instead of only 10 producing $T = P$), while it makes ML *less* able to recover T from $M_{T,e=1}$ (with 7 shapes instead of 8 producing $T = L$). Similarly, all 11 tree-shapes produce $T = P$ for $M_{T,s=4}$ in the case of MP, while all tree-shapes produce $T \subset L$ in the case of ML.

For binary data, the matrix of all possible combinations of 0s and 1s is perfectly undecisive for both parsimony and a 2-state Mkv model (as discussed in Section 3). However, in the case of 4-state characters, the matrix with all possible combinations of states 0–3 (or ACGT) is, surprisingly, not fully undecisive under MP: for such matrix, some tree-shapes have fewer steps than others. This seems to be so because the

Table 3
Results of maximum parsimony analysis for matrices containing all the possible 4-state characters with different numbers of steps (s) and different numbers of extra steps (e), for each of the tree-shapes, for different numbers of taxa. Cases where the matrix for every tree-shape produces the same tree T as single most parsimonous tree P (i.e. $T = P$) are indicated with (+); cases where every one of the shapes produces most parsimoniou(s) tree(s) different from T are indicated with (−). In the rest of cases, the number of shapes for which $T = P$ is indicated first, followed by the number of shapes where T is one among multiple equally parsimonious trees, ending with the number of shapes where T is not most parsimonious for the corresponding matrix. The Ø symbol indicates that no possible distribution of 4 states can have (for the given number of taxa) that number of extra steps.

Taxa	s = 2	e = 2	s = 3	e = 3	s = 4	e = 4	s = 5	e = 5	s = 6	e = 6	s = 7	e = 7
5	+	Ø	−									
6	+	−	0,1,2	Ø	−							
7	+	−	+	Ø	−	Ø						
8	+	–	+	−	8,0,3	Ø	−	Ø				
9	+	−	+	−	+	−	−	Ø				
10	+	+	+	–	+	−	+	Ø	−	Ø		
11	+	+	+	−	+	−	+	−	−	Ø	−	Ø
12	+	+	+	+	+	−	+	−	+	−	−	Ø

number of characters that appear as synapomorphies depends on the size of the group. For example, for 2-taxon groups the two taxa must have an identical state for the character to be mapped as a synapomorphy; instead, for groups with more taxa, one of the taxa (nested within a subgroup) may have a different state, but as long as its successive sister groups have the same state (different from its outgroup), the character will be mapped as an unambiguous synapomorphy nonetheless. Because of this, the numbers of characters supporting each possible taxon partition are not the same –some partitions are supported by more characters than others.

The matrix with all combinations of 4 states is perfectly undecisive when analyzed under JC69 or 4-state Mkv: all tree-shapes have the same likelihood, as calculated by PAUP*. The frequencies of the different state patterns in these matrices is the same one expected when large numbers of characters are generated, assigning any of the 4 states at random to each of the terminals, or when characters are evolved on a tree with infinite branch lengths. Those data carry no phylogenetic signal, and they are reassuringly perceived by current ML methods as completely lacking any information to discriminate trees.

8. Conclusions and open questions

The relationships between confidence, ease of finding P or L, and equivalence between MP and ML, do not simply depend on amounts of homoplasy, but instead on more complex aspects of how the homoplasy is distributed. The present analysis determined the maximum amount of homoplasy in a set of binary characters over a tree T to be $0.25t$ for T to be the same as P (and L), by empirically enumerating all possible combinations for low numbers of taxa, and explained this number with a general argumentation. It would be interesting to determine that number from a more formal analysis, and if possible, to identify the conditions that in the boundary of $s = 0.25t$ cause P to be the same as T for some tree-shapes but not for others.

Acknowledgments

We thank Mike Steel, Salva Arias, and an anonymous reviewer for comments on the ms., and Clau Szumik and Santi Catalano for discussion and encouragement. The research for this study was supported by deeply appreciated grants from CONICET (PIP 0687, PUE 0070, to PAG) as well as NSF, NASA and FAPESP (via a grant to J. Cracraft and L. Lohman, entitled “Assembly and evolution of the Amazonian biota and its environment: an integrated approach”, with participation from PAG).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ympev.2018.01.020.

References

Baum, B., 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41, 3–10.

Böcker, S., Bryant, D., Dress, A., Steel, M., 2000. Algorithmic aspects of tree amalgamation. J. Algorithms 37, 522–537.

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

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

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts, pp. 664.

Ford, E., St. John, K., Wheeler, W., 2015. Towards improving searches for optimal phylogenies. Syst. Biol. 64, 56–65.

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

Goloboff, P., 1991. Homoplasy and the choice among cladograms. Cladistics 7, 215–232.

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

Goloboff, P., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

Goloboff, P., 2014. Hide and vanish: data sets where the most parsimonious tree is known but hard to find, and their implications for tree search methods. Mol. Phylogenet. Evol. 79, 118–131.

Goloboff, P., 2015. Computer science and parsimony: a reappraisal, with discussion of methods for poorly structured data sets. Cladistics 31, 210–225.

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

Goloboff, P., Pol, D., 2007. On divide-and-conquer strategies for parsimony analysis of large data sets: Rec-I-DCM3 versus TNT. System. Biol. 56, 485–495.

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

Goloboff, P., Simmons, M., 2014. Bias in tree searches and its consequences for measuring groups supports. Syst. Biol. 63, 851–861.

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

Goloboff, P., Arias, J., Szumik, C., 2017. Comparing tree-shapes: beyond symmetry. Zool. Scripta 46. http://dx.doi.org/10.1111/zsc.12231.

Huber, K., Moulton, V., Steel, M., 2005. Four characters suffice to convexly define a phylogenetic tree. SIAM J. Discrete Math. 18, 835–843.

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

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

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.

Ragan, M., 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1, 51–58.

Roch, S., 2006. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. IEEE/ACM Trans. Comput. Biol. Bioinform. 3, 92–94.

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

Roshan, U., Moret, B., Williams, T., Warnow, T., 2004. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In: Proceedings 3rd IEEE Computational Systems Bioinformatics Conference (CSB 2004), pp. 98–109.

Schmidt, H., Strimmer, K., Vingron, M., von Haeseler, A., 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18, 502–504.

Steel, M., 1992. The complexity of reconstructing trees from qualitative characters and subtrees. J. Classif. 9, 91–116.

Steel, M., Charleston, M., 1995. Five surprising properties of parsimoniously colored trees. Bull. Math. Biol. 57, 367–375.

Steel, M., Penny, D., 2005. Maximum parsimony and the phylogenetic information in multistate characters. In: Albert, Victor (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, pp. 163–178.

St. John, K., 2016. The shape of phylogenetic treespace. Syst. Biol. http://dx.doi.org/10.1093/sysbio/syw025.

Swofford, D., 2002. PAUP*: Phylogenetic Analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA.

Yang, J., Grünewald, S., Xu, Y., Wan, X., 2014. Quartet-based methods to reconstruct phylogenetic networks. BMC Syst. Biol. 8 (21), 1–12. http://dx.doi.org/10.1186/1752-0509-8-21.