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
1055-7903/

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 parsimonous 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