Molecular Phylogenetics and Evolution 80 (2014) 165–168

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Short Communication

# The most parsimonious tree for random data

Mareike Fischer [a], Michelle Galla [a], Lina Herbst [a], Mike Steel [b,*]

[a] Department for Mathematics and Computer Science, Ernst-Moritz-Arndt University, Greifswald, Germany
[b] Allan Wilson Centre, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

*Article history:*
Received 2 June 2014
Revised 18 July 2014
Accepted 19 July 2014
Available online 29 July 2014

*Keywords:*
Tree
Maximum parsimony
Random data
Central limit theorem

ABSTRACT

Applying a method to reconstruct a phylogenetic tree from random data provides a way to detect whether that method has an inherent bias towards certain tree 'shapes'. For maximum parsimony, applied to a sequence of random 2-state data, each possible binary phylogenetic tree has exactly the same distribution for its parsimony score. Despite this pleasing and slightly surprising symmetry, some binary phylogenetic trees are more likely than others to be a most parsimonious (MP) tree for a sequence of $k$ such characters, as we show. For $k = 2$, and unrooted binary trees on six taxa, any tree with a caterpillar shape has a higher chance of being an MP tree than any tree with a symmetric shape. On the other hand, if we take any two binary trees, on any number of taxa, we prove that this bias between the two trees vanishes as the number of characters $k$ grows. However, again there is a twist: MP trees on six taxa for $k = 2$ random binary characters are more likely to have certain shapes than a uniform distribution on binary phylogenetic trees predicts. Moreover, this shape bias appears, from simulations, to be more pronounced for larger values of $k$.



## 1. Introduction

The 'shape' of reconstructed evolutionary trees is of interest to evolutionary biologists, as it should provide some insight into the processes of speciation and extinction (Aldous, 2001; Aldous et al., 2011; Hey, 1992; Holton et al., 2014; Lambert et al., 2013; Stadler, 2013). In this paper, 'shape' refers just to the discrete shape of the tree (i.e. we ignore the branch lengths); the advantages of this are that it simplifies the analysis, and it also confers a certain robustness (i.e. the resulting probability distribution on discrete shapes is often independent of the fine details of an underlying speciation/extinction model (Aldous, 1995; Lambert et al., 2013)). For example, if all speciation (and extinction) events affect all taxa at any given epoch in the same way, then we should expect the shape of a reconstructed tree to be that predicted by the discrete 'Yule–Harding' model (Aldous, 2001; Harding, 1971; Lambert et al., 2013). In fact, a general trend (see e.g. Aldous, 2001) is that the shape of phylogenetic trees reconstructed from biological data tends to be a little less balanced than this model predicts, but is more balanced than what would be obtained under a uniform model in which each binary phylogenetic tree has the same probability (this model is sometimes also called the 'Proportional-to-Distinguishable-Arrangements' (PDA) model (Rosen, 1978)).

There are, however, other factors which can lead to biases in tree shape. One is non-random sampling of the taxa on which to construct a tree (influenced, for example, by the particular interests of the biologists or the application of a certain strategy to sample taxa). Another cause of possible bias is that a tree reconstruction method may itself have an inherent preference towards certain tree shapes. A way to test this latter possibility is to apply the tree reconstruction method to data that contain no phylogenetic signal at all, in particular, purely random data, where each character is generated independently by a process that assigns states to the taxa uniformly (e.g. by the toss of a fair coin in the case of two states). For some methods, such as 'TreePuzzle', such data leads to very balanced trees (similar to the Yule–Harding model (Vinh et al., 2010; Zhu et al., 2013)). However, other methods, such as maximum likelihood and maximum parsimony, lead to less balanced trees, that are closer in shape to the uniform model, as recently reported in Holton et al. (2014). In the case of maximum parsimony, the two-state symmetric model has the even-handed property that every binary tree has exactly the same distribution of its parsimony score on $k$ randomly generated characters. Thus, it might be supposed that the maximum parsimony (MP) tree for such a sequence of characters would also follow a uniform distribution. However, while this holds in special cases, it does not hold in general, as we show below.

* Corresponding author. Fax: +64 33642587.
*E-mail address:* mike.steel@canterbury.ac.nz (M. Steel).

http://dx.doi.org/10.1016/j.ympev.2014.07.010
1055-7903/

### 1.1. Trees and parsimony: definitions and basic properties

In phylogenetics, graphs, especially trees, are used to describe the ancestral relationships among different species. A main goal of phylogenetics is to infer an evolutionary tree from data available from present-day species. In graph theory, a tree $T = (V, E)$ consists of a connected graph with no cycles. Certain leaf-labelled trees ('phylogenetic trees') are widely used where the set of extant species label the leaves and the remaining vertices represent ancestral speciation events (Felsenstein, 2004). There are different methods of reconstructing a phylogenetic tree. One of the most famous tree reconstruction methods is maximum parsimony. For a given tree and discrete character data, the parsimony score can be found in polynomial time by using the Fitch–Hartigan algorithm (Fitch, 1971; Hartigan, 1973). The parsimony score counts the number of changes (mutations) required on the tree to describe the data. This problem of finding the optimal parsimony score for a given tree is often called the 'small parsimony' problem. The 'big parsimony' problem aims at finding the most parsimonious tree ('MP tree') amongst all possible trees. This problem has been proven to be NP-hard (Foulds and Graham, 1982).

In this paper, we assume that each taxon from the leaf set $X$ of the tree is assigned a binary state (0 or 1) independently, and with equal probability (the case where the two states have different probabilities is less interesting, since then the distribution of the parsimony score of a fixed binary tree is easily seen to depend on the shape of the tree, even for a single character). This process is then repeated (also independently) to generate a sequence of characters (defined formally below). For binary trees with random data, we are interested in the probability that a tree is an MP tree, and also what happens when the length of the sequences or the number of leaves gets larger. In particular, we wish to determine whether each tree is equally likely to be selected as an MP tree.

**Definition 1** (*Binary phylogenetic trees*). An *(unrooted) binary phylogenetic X-tree* is a tree $T$ with leaf set $X$ and with every interior (i.e. non-leaf) vertex of degree exactly three. We will let $UB(X)$ be the set of unrooted binary phylogenetic $X$-trees. When $X = [n] = \{1, \ldots, n\}$, we will write $UB(n)$.

**Definition 2.** [Character, extension, parsimony score]

- A *character on* $X$ over a finite set $R$ of character states is any function $f$ from $X$ into $R$; $f : X \to R$. In this paper we will consider two-state characters; $f : X \to \{0, 1\}$.
- A function $\bar{f} : V \to R$ such that $\bar{f}|_X = f$ is said to be an *extension* of $f$ since it describes an assignment of states to all vertices of $T$ that agrees with the states that $f$ stipulates at the leaves.
- Let $ch(\bar{f}, T) := \left|\{e = \{u, v\} \in E : \bar{f}(u) \neq \bar{f}(v)\}\right|$ be the *changing number* of $\bar{f}$. Given a character $f : X \to R$, the *parsimony score of* $f$ on $T$, denoted $ps(f, T)$, is the smallest changing number of any extension of $f$, i.e.:

$$ps(f, T) := \min_{\bar{f}: V \to R, \bar{f}|_X = f} \{ch(\bar{f}, T)\}.$$

An extension $\bar{f}$ of $f$ for which $ch(\bar{f}, T) = ps(f, T)$ is said to be a *minimal extension*.

Let $\mathcal{C} = (f_1, \ldots, f_k)$ be a sequence of characters on $X$. The *parsimony score of* $\mathcal{C}$ on $T$, denoted $ps(\mathcal{C}, T)$, is defined by $ps(\mathcal{C}, T) := \sum_{i=1}^{k} ps(f_i, T)$.

## 2. Comparing given trees

Let $X_k(T)$ be the parsimony score of $k$ random two-state characters on $T \in UB(n)$. We will see shortly (Proposition 1) that the distribution of $X_k(T)$ does not depend on the shape of $T$; it just depends on $n$. Notice that $X_k(T) = X_1 + X_2 + \cdots + X_k$, where $X_i$ (for $i = 1, \ldots, k$) form a sequence of independent and identically distributed random variables (with common distribution $X_1(T)$). If $\mathbb{P}(X_k(T) = l)$ denotes the probability that $T$ has parsimony score $l$ then, from Steel (1993), we have, for each $l \in [1, \lfloor n/2 \rfloor]$:

$$\mathbb{P}(X_1(T) = l) = \frac{2n - 3l}{l} \cdot \binom{n - l - 1}{l - 1} \cdot 2^{l-n}, \tag{1}$$

with $\mathbb{P}(X_1(T) = 0) = 2^{1-n}$ and $\mathbb{P}(X_1(T) = l) = 0$ for $l > \lfloor n/2 \rfloor$. Furthermore, $\mathbb{E}[X_1(T)] = \frac{3n-2-\left(\frac{-1}{2}\right)^{n-1}}{9} \sim \frac{n}{3}$ is the expected parsimony score of $T$, and $\mathbb{E}[X_k(T)] = k \cdot \mathbb{E}[X_1(T)]$. An immediate consequence of (1) is the following.

**Proposition 1.** *For every* $k \geqslant 1$ and $n \geqslant 2$, *the distribution of the parsimony score of* $k$ *independent random binary characters (i.e.* $X_k(T)$*) is the same for all* $T \in UB(n)$.

### 2.1. Comparing two trees by their parsimony score

We begin this section by describing a tree rearrangement operation on binary phylogenetic trees (Semple and Steel, 2003, Chapter 2.6), namely tree bisection and reconnection (TBR). Let $T$ be a binary phylogenetic $X$-tree and let $e = \{u, v\}$ be an edge of $T$. A TBR operation is described as follows. Let $T'$ be the binary tree obtained from $T$ by deleting $e$, adding an edge between a vertex that subdivides an edge of one component of $T \setminus e$ and a vertex that subdivides an edge of the other component of $T \setminus e$, and then suppressing any resulting degree-two vertices. In the case that a component of $T \setminus e$ consists of a single vertex, then the added edge is attached to this vertex. $T'$ is said to be obtained from $T$ by a single TBR operation.

**Proposition 2.** *Let* $T,\ T' \in UB(n)$.

- (i) *If* $T$ *and* $T'$ *are one TBR apart, then* $\mathbb{P}(X_k(T) < X_k(T')) = \mathbb{P}(X_k(T') < X_k(T))$ *holds for all* $k \geqslant 1$.
- (ii) *If* $T$ *and* $T'$ *are more than one TBR apart, then the equality* $\mathbb{P}(X_k(T) < X_k(T')) = \mathbb{P}(X_k(T') < X_k(T))$ *can fail, even for* $k = 1$ *and* $n = 6$.

**Proof.**

(i) From Bryant (2004, Lemma 5.1), if $T$ and $T'$ are one TBR apart then for any character $f$, $\left|ps(f, T) - ps(f, T')\right| \leqslant 1$. In particular,

$$\left|X_1(T) - X_1(T')\right| \leqslant 1. \tag{2}$$

For $k \geqslant 1$, let $\Delta_k = X_k(T) - X_k(T')$. Then if $T,\ T' \in UB(n)$ are one TBR apart, then $\Delta_1 = X_1(T) - X_1(T')$ is either 0, 1 or $-1$, by (2). Moreover, $\mathbb{P}(\Delta_1 = m) = \mathbb{P}(\Delta_1 = -m)$ for all $m \in \{0, 1 - 1\}$, since $\mathbb{E}[\Delta_1] = 0$, by Proposition 1. Furthermore, $\Delta_k = D_1 + \cdots + D_k$, where $D_1, \ldots, D_k$ are independent and identically distributed as $\Delta_1$, so we have:

$$\begin{aligned}
\mathbb{P}(\Delta_k = m) &= \sum_{\substack{m_1, \ldots, m_k \in \{-1,0,1\}: \\ m_1 + \cdots + m_k = m}} \mathbb{P}(D_1 = m_1 \wedge D_2 = m_2 \wedge \cdots \wedge D_k = m_k) \\
&= \sum_{\substack{m_1, \ldots, m_k \in \{-1,0,1\}: \\ m_1 + \cdots + m_k = m}} \prod_{j=1}^{k} \mathbb{P}(D_j = m_j) = \sum_{\substack{m_1, \ldots, m_k \in \{-1,0,1\}: \\ m_1 + \cdots + m_k = m}} \prod_{j=1}^{k} \mathbb{P}(D_j = -m_j) \\
&= \sum_{\substack{m'_1, \ldots, m'_k \in \{-1,0,1\}: \\ m'_1 + \cdots + m'_k = -m}} \mathbb{P}(D_1 = m'_1 \wedge D_2 = m'_2 \wedge \cdots \wedge D_k = m'_k) = \mathbb{P}(\Delta_k = -m).
\end{aligned}$$

This provides the equality $\mathbb{P}(X_k(T) < X_k(T')) = \mathbb{P}(X_k(T') < X_k(T))$ for all $k \geqslant 1$.