Journal of Theoretical Biology 417 (2017) 100–108

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

On the quirks of maximum parsimony and likelihood on phylogenetic networks

Christopher Bryant[a], Mareike Fischer[b], Simone Linz[c,*], Charles Semple[d]

[a] Statistics New Zealand, Wellington, New Zealand
[b] Department for Mathematics and Computer Science, Ernst Moritz Arndt University, Greifswald, Germany
[c] Department of Computer Science, University of Auckland, New Zealand
[d] School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Keywords:
Hardwired
Likelihood
Parsimony
Phylogenetic networks
Softwired

ABSTRACT

Maximum parsimony is one of the most frequently-discussed tree reconstruction methods in phylogenetic estimation. However, in recent years it has become more and more apparent that phylogenetic trees are often not sufficient to describe evolution accurately. For instance, processes like hybridization or lateral gene transfer that are commonplace in many groups of organisms and result in mosaic patterns of relationships cannot be represented by a single phylogenetic tree. This is why phylogenetic networks, which can display such events, are becoming of more and more interest in phylogenetic research. It is therefore necessary to extend concepts like maximum parsimony from phylogenetic trees to networks. Several suggestions for possible extensions can be found in recent literature, for instance the softwired and the hardwired parsimony concepts. In this paper, we analyze the so-called big parsimony problem under these two concepts, i.e. we investigate maximum parsimonious networks and analyze their properties. In particular, we show that finding a softwired maximum parsimony network is possible in polynomial time. We also show that the set of maximum parsimony networks for the hardwired definition always contains at least one phylogenetic tree. Lastly, we investigate some parallels of parsimony to different likelihood concepts on phylogenetic networks.

1. Introduction

Traditionally, phylogenetic trees are used to represent ancestral relationships between species. However, more recently, investigations into hybridization and lateral gene transfer challenge the model of a phylogenetic tree. While lateral gene transfer, which is commonly accepted in prokaryotes, continues to be an area of discussion in multicellular organisms (Soucy et al., 2015), hybridization is more broadly accepted as a commonplace process in many groups of eukaryotes (Mallet, 2005; Mallet et al., 2016). For example, the impact of hybridization on New Zealand's flora and fauna was analyzed by Morgan-Richards et al. (2009), who confirmed hybridization between at least 19 pairs of endemic species ranging from plants and insects to fish and birds. Likewise, Marcussen et al. (2014) report that the present-day bread wheat genome has resulted from several ancient hybridization events. Indeed, it is now increasingly acknowledged that phylogenetic networks are better suited to represent the evolutionary history of species. To accurately describe complex evolutionary histories, it is therefore essential to provide biologists with tools that allow for investigations into relationships that do not always fit a strict tree model.

In the light of the popularity of maximum parsimony (MP) as a tool to reconstruct phylogenetic trees from a sequence of morphological or molecular characters, consideration is currently being given to extending parsimony to phylogenetic networks. Similar to parsimony on phylogenetic trees (reviewed in Felsenstein (2004)), one distinguishes the small and big parsimony problem. In terms of phylogenetic networks, the small parsimony problem asks for the parsimony score of a sequence of characters on a (given) phylogenetic network, while the big parsimony problem asks to find a phylogenetic network for a sequence of characters that minimizes the score amongst all phylogenetic networks. It is the latter problem that evolutionary biologists usually want to solve for a given data set, and it is this problem that is the focus of this paper.

Recently, two different approaches for parsimony on phylogenetic networks have been proposed, referred to as *hardwired* and *softwired* parsimony. The hardwired framework, introduced by Kannan and Wheeler (2012), calculates the parsimony score of a phylogenetic network by considering character-state transitions along every edge of the network. A slightly different approach was taken by Nakhleh

[*] Corresponding author.
E-mail addresses: chris.bryant@stats.govt.nz (C. Bryant), email@mareikefischer.de (M. Fischer), s.linz@auckland.ac.nz (S. Linz), charles.semple@canterbury.ac.nz (C. Semple).

http://dx.doi.org/10.1016/j.jtbi.2017.01.013
Received 29 August 2016; Received in revised form 28 October 2016; Accepted 6 January 2017
Available online 11 January 2017

et al. (2005), who defined the softwired parsimony score of a phylogenetic network to be the smallest (ordinary) parsimony score of any phylogenetic tree that is displayed by the network under consideration. Although one can compute the hardwired parsimony score of a set of binary characters on a phylogenetic network in polynomial time (Semple and Steel, 2003), solving the small parsimony problem is in general NP-hard under both notions (Fischer et al., 2015; Jin et al., 2009; Nguyen et al., 2007). In contrast, the small parsimony problem on phylogenetic trees is solvable in polynomial time by applying Fitch-Hartigan's (Fitch, 1971; Hartigan, 1973) or Sankoff's (Sankoff, 1975) algorithm.

Given that it is in general computationally expensive to solve the small parsimony problem on networks, effort has been put into the development of heuristics (Kannan and Wheeler, 2012), and algorithms that are exact and have a reasonable running time despite the complexity of the underlying problem (Fischer et al., 2015; Kannan and Wheeler, 2014). However, in finding ever quicker and more advanced algorithms to solve the small parsimony problem, an analysis of MP networks under the hardwired or softwired notion, and their biological relevance has fallen short. The only exceptions are two practical studies (Jin et al., 2006, 2007) that aim at the reconstruction of a particular type of a softwired MP network for which the input does not only consist of a sequence of characters, but also of a given phylogenetic tree $\mathcal{T}$ (e.g. a species tree) and a positive integer k. More precisely, this version of softwired parsimony adds k reticulation edges to $\mathcal{T}$ such that the softwired parsimony score of the resulting phylogenetic network is minimized over all possible solutions.

In this paper, we present the first analysis of MP networks by highlighting a number of flaws that underlie the previously introduced methods of hardwired and softwired parsimony. We reveal fundamental properties of phylogenetic networks reconstructed under these two methods that are simultaneously surprising and undesirable. For example, we show that an MP network under the hardwired definition tends to have a small number of reticulations, while an MP network under the softwired definition tends to have many reticulations. Even stronger, we show that, for any sequence of characters, there always exists a phylogenetic tree that is an MP network under the hardwired definition. While some of our findings have independently been stated in Wheeler (2015), we remark that the author does not give any formal proofs. In conclusion, the properties we find question the biological meaningfulness of MP networks and emphasize a fundamental difference between the hardwired and softwired parsimony framework on phylogenetic networks. We then shift towards maximum likelihood concepts on phylogenetic networks and analyze whether or not the Tuffley-Steel equivalence result for phylogenetic trees also holds for networks. It is well known that under a simple substitution model, parsimony and likelihood on phylogenetic trees are equivalent (Tuffley and Steel, 1997). However, as we shall show, parsimony on networks is not equivalent to one of the most frequently-used likelihood concepts on networks. Nevertheless, the equivalence can be recovered using functions that resemble likelihoods, but are not true likelihoods in a probability theoretical sense. We call these functions pseudo-likelihoods. In this sense, the equivalence of the different parsimony concepts to pseudo-likelihoods rather than likelihoods can be viewed as another drawback of the existing notions of parsimony.

The remainder of the paper is organized as follows. The next section contains notation and terminology that is used throughout the paper. We then analyze properties of MP networks under the hardwired and softwired definition in Section 3. Additionally, this section also considers the computational complexity of the big parsimony problem under both definitions. Then, in Section 4, we re-visit the Tuffley-Steel equivalence result for parsimony and likelihood, and investigate in how far it can be extended from trees to networks. We end the paper with a brief conclusion in Section 5.

Lastly, it is worth noting that our results are presented as general as possible. For example, we do not bound the number of character states of any character that is considered in this paper. Furthermore, the only restriction in the definition of a phylogenetic network (see next section for details) is that the out-degree of a reticulation is exactly one. As a reticulation and speciation event are unlikely to happen simultaneously, this restriction is biologically sensible and, in fact, only needed to establish Theorem 5.

2. Preliminaries

2.1. Trees and networks

A *rooted phylogenetic tree on* X is a rooted tree with no degree-two vertices (except possibly the root which has degree at least two) and whose leaf set is X. Furthermore, a rooted phylogenetic tree on X is *binary* if each internal vertex, except for the root, has degree three. A natural extension of a rooted phylogenetic tree on X that allows for vertices whose in-degree is greater than one is a *rooted phylogenetic network* $\mathcal{N}$ *on* X which is a rooted acyclic digraph that satisfies the following three properties:

(i) X is the set of vertices of in-degree one and out-degree zero,
(ii) the out-degree of the root is at least two, and
(iii) every other vertex has either in-degree one and out-degree at least two, or in-degree at least two and out-degree one.

Similar to rooted phylogenetic trees, we call X the *leaf set* of $\mathcal{N}$. Furthermore, each vertex of $\mathcal{N}$ whose in-degree is at least two is called a *reticulation* and represents a species whose genome is a mosaic of at least two distinct parental genomes, while each edge directed into a reticulation is called a *reticulation edge*. To illustrate, a rooted phylogenetic network on $X = \{1, 2, 3, 4\}$ and with one reticulation is shown on the left-hand side of Fig. 1. Moreover, for two vertices u and v in $\mathcal{N}$, we say that u is a *parent* of v or, equivalently, v is a *child* of u if (u,v) is an edge in $\mathcal{N}$. Lastly, note that a rooted phylogenetic tree on X is a rooted phylogenetic network on X with no reticulation.

Let $\mathcal{N}$ be a rooted phylogenetic network on X and let $\mathcal{T}$ be a rooted phylogenetic tree on X. We say that $\mathcal{T}$ is *displayed* by $\mathcal{N}$ if, up to contracting vertices with in-degree one and out-degree one, $\mathcal{T}$ can be obtained from $\mathcal{N}$ by deleting edges and non-root vertices, in which case the resulting acyclic digraph is an *embedding* of $\mathcal{T}$ in $\mathcal{N}$. Intuitively, if $\mathcal{T}$ is displayed by $\mathcal{N}$, then all ancestral information inferred by $\mathcal{T}$ is also inferred by $\mathcal{N}$. The two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ that are displayed by the network shown on the left-hand side of Fig. 1 are presented on the right-hand side of the same figure. Lastly, we use $\mathcal{D}(\mathcal{N})$ to denote the set of all rooted phylogenetic trees that are displayed by $\mathcal{N}$.

2.2. Characters

Let G be an acyclic digraph. We denote the vertex set of G by $V(G)$ and the edge set of G by $E(G)$. Furthermore, we call X a *distinguished set* of G if it is a subset of the vertices of G whose out-degree is zero such that, if G is a rooted phylogenetic network $\mathcal{N}$ (resp. a rooted

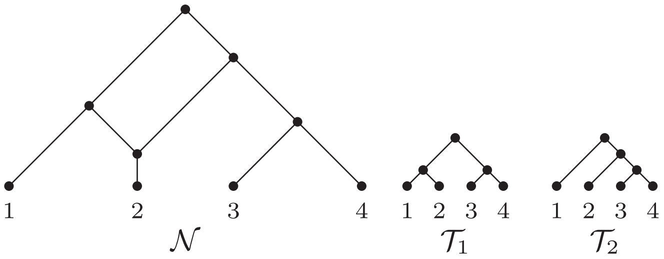

Fig. 1. Left: A rooted phylogenetic network $\mathcal{N}$ on leaf set $X = \{1, 2, 3, 4\}$. Right: The two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ on X displayed by $\mathcal{N}$.