Advances in Applied Mathematics 66 (2015) 22–45

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

A parsimony-based metric for phylogenetic trees

Vincent Moulton, Taoyang Wu [*]

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom

ARTICLE INFO

Article history:
Received 17 November 2012
Accepted 6 October 2014
Available online 6 March 2015

MSC:
05C05
05C99
92B10

Keywords:
Metric
Phylogenetic trees
Parsimony score
Tree operations
Unit neighborhood
Diameter

ABSTRACT

In evolutionary biology various metrics have been defined and studied for comparing phylogenetic trees. Such metrics are used, for example, to compare competing evolutionary hypotheses or to help organize algorithms that search for optimal trees. Here we introduce a new metric d_{P} on the collection of binary phylogenetic trees each labeled by the same set of species. The metric is based on the so-called parsimony score, an important concept in phylogenetics that is commonly used to construct phylogenetic trees. Our main results include a characterization of the unit neighborhood of a tree in the d_{P} metric, and an explicit formula for its diameter, that is, a formula for the maximum possible value of d_{P} over all possible pairs of trees labeled by the same set of species. We also show that d_{P} is closely related to the well-known tree bisection and reconnection (TBR) and subtree prune and regraft (SPR) distances, a connection which will hopefully provide a useful new approach to understanding properties of these and related metrics.

[*] Corresponding author. Fax: +44 1603 593345.
E-mail addresses: vincent.moulton@cmp.uea.ac.uk (V. Moulton), taoyang.wu@uea.ac.uk (T. Wu).

http://dx.doi.org/10.1016/j.aam.2015.02.002
0196-8858/© 2015 Elsevier Inc. All rights reserved.

1. Introduction

In evolutionary biology, researchers are often faced with the problem of comparing two evolutionary or phylogenetic trees on a given set of species. This problem commonly arises because there are various methods to construct such trees, and these often give different solutions which then need to be compared. In addition, some of the methods for constructing phylogenetic trees rely on searching through the set of all possible trees, and it can be useful to compare trees so as to efficiently organize such searches (see, e.g. [20]). For these reasons various metrics have been developed for comparing phylogenetic trees, see e.g. [1–3,9,15,16,18,19]. These metrics have different properties which can make them more (or less!) useful depending on the situation in which they are to be used. For example, the so-called Robinson–Foulds metric [16] can give a quick way to compare trees, but is somewhat coarse in identifying details, whereas other metrics, such as the quartet-distance [9], can pick out more fine detail but can be more difficult to work with computationally.

In this paper, we introduce a new tree metric which is based on the concept of parsimony. To define this metric we first need to recall some concepts in phylogenetics (cf. [17]). Let X be a finite set, corresponding to a set of species. A *character* on X is a surjective map χ from X into another finite set $\mathcal{C}$. In biology, characters are commonly morphological (e.g. a species in X either has fins or not) or genetic (e.g. the nucleotide in some position of the DNA for a species in X is A, T, C or G). Now, given such a character χ, and a phylogenetic tree $\mathcal{T} = (V, E)$ on X (i.e. a graph-theoretical tree with vertex set V, edge set E and leaf-set X, such that every interior vertex has degree three), an *extension* $\overline{\chi}$ of χ to $\mathcal{T}$ is a map $\overline{\chi} : V \to \mathcal{C}$ with $\chi(x) = \overline{\chi}(x)$ for all $x \in X$. The *changing number* of $\overline{\chi}$ is the cardinality of the set $\Delta(\overline{\chi})$ consisting of all edges $\{u, v\}$ in $\mathcal{T}$ with $\overline{\chi}(u) \neq \overline{\chi}(v)$. The extension χ is *optimal* if it has the minimum changing number over all possible extensions of χ to $\mathcal{T}$, and the *parsimony score* $l(\mathcal{T}, \chi)$ of χ on $\mathcal{T}$ is defined as the changing number of an optimal extension of χ to $\mathcal{T}$. Note that in phylogenetics it is common practice to look for a phylogenetic tree that minimizes the sum of the parsimony scores over a given set of characters (see, e.g. [17, Chapter 5]) as such a tree is considered to represent a simplest explanation for how present-day species might have evolved.

Now, given two phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on X, we define

$$d_{\mathrm{P}}(\mathcal{T}, \mathcal{T}') = \max_{\chi \in \Xi(X)} \Big| l(\mathcal{T}, \chi) - l(\mathcal{T}', \chi) \Big|,$$

where $\Xi(X)$ denotes the set of all characters on X. In other words, $d_{\mathrm{P}}(\mathcal{T}, \mathcal{T}')$ is the largest difference in the parsimony scores for the trees $\mathcal{T}$ and $\mathcal{T}'$ over all possible characters on X. Note that, by definition, $d_{\mathrm{P}}(\mathcal{T}, \mathcal{T}') = d_{\mathrm{P}}(\mathcal{T}', \mathcal{T})$ and that $d_{\mathrm{P}}(\mathcal{T}, \mathcal{T}') \geq 0$ with equality holding if and only if $\mathcal{T} = \mathcal{T}'$. Moreover, if $\mathcal{T}''$ is also a phylogenetic tree on X, and χ^* is a character on X with $d_{\mathrm{P}}(\mathcal{T}, \mathcal{T}') = |l(\mathcal{T}, \chi^*) - l(\mathcal{T}', \chi^*)|$, then we have