Information Processing Letters 123 (2017) 36–41

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

A cubic-time algorithm for computing the trinet distance between level-1 networks

Vincent Moulton, James Oldman, Taoyang Wu *

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

ARTICLE INFO

Article history:
Received 4 July 2016
Received in revised form 17 February 2017
Accepted 10 March 2017
Available online 16 March 2017
Communicated by Ł. Kowalik

Keywords:
Phylogenetic tree
Phylogenetic network
Algorithms
Trinet
Robinson–Foulds metric

ABSTRACT

In evolutionary biology, phylogenetic networks are constructed to represent the evolution of species in which reticulate events are thought to have occurred, such as recombination and hybridization. It is therefore useful to have efficiently computable metrics with which to systematically compare such networks. Through developing an optimal algorithm to enumerate all trinets displayed by a level-1 network (a type of network that is slightly more general than an evolutionary tree), here we propose a cubic-time algorithm to compute the trinet distance between two level-1 networks. Employing simulations, we also present a comparison between the trinet metric and the so-called Robinson–Foulds phylogenetic network metric restricted to level-1 networks. The algorithms described in this paper have been implemented in JAVA and are freely available at https://www.uea.ac.uk/computing/TriLoNet.

1. Introduction

Various types of phylogenetic networks have been introduced to explicitly represent the reticulate evolutionary history of organisms such as viruses and bacteria in which processes such as recombination and lateral gene transfer occur [1]. Essentially, such networks are binary, directed acyclic graphs with a single root, whose leaves correspond to the organisms or species in question. Here we focus on level-1 networks, a type of phylogenetic network that is slightly more general than an evolutionary tree, and closely related to so-called galled-trees (see, e.g. [2]). Level-1 networks are characterized by the property that any two cycles within them are disjoint (see the next section for a formal definition and Fig. 1 for an example). Due to the availability of practical algorithms for their construction [3,4], level-1 networks have attracted much attention in recent years (see, e.g. [2,5–7]) and they have been used to, for example, represent the evolution of the fungus *Fusarium graminearum* [1], and that of HBV [4].

A key challenge for phylogenetic networks is to quantify the incongruence between two networks which represent competing evolutionary histories for a given dataset. Such pairs can arise, for example, when different networks are inferred using different methods or construction (see e.g. [8] for an overview of network building methods). In consequence, various metrics have been developed for comparing phylogenetic networks (cf. Chapter 6 in [1] for an overview). Ideally, such a metric should be efficient to compute since it may need to be repeatedly computed (for example, in simulations such as the ones that we present later in this paper). Moreover, it is useful if the diameter can be derived for the metric (i.e. the maximum value for the metric taken over all pairs of all possible networks) so that distances can be normalized.

Here we develop an efficient cubic-time algorithm to compute the trinet distance between two level-1 networks, that is, the number of trinets (i.e., networks on three taxa)

* Corresponding author.
E-mail addresses: vincent.moulton@cmp.uea.ac.uk (V. Moulton), James.Oldman@gmail.com (J. Oldman), taoyang.wu@gmail.com (T. Wu).

http://dx.doi.org/10.1016/j.ipl.2017.03.002
0020-0190/

displayed by one but not both networks. We also give the diameter of this metric. The trinet metric was introduced in [9] and used in [4] to compare the performance of network inference algorithms. Note that the trinet distance is closely related to the triplet distance, which is the number of 3-leaved trees exhibited by one but not both networks (see, e.g. [10]). However, in contrast to the trinet metric, the triplet metric is not proper in that there exist pairs of distinct level-1 networks whose triplet distance is zero. In addition to the trinet metric, other proper metrics that can be used for comparing level-1 networks include the tripartition metric [11], the path-multiplicity metric [12], the NNI metric [13], and the Robinson–Foulds metric [2]. Among these metrics, only the NNI metric was specifically defined for level-1 networks, while the others were introduced for more general classes of networks and can be restricted to level-1 networks to give proper level-1 metrics. However, establishing the diameters for these other metrics on level-1 metrics appears to be a challenging problem, although in this paper we shall derive the diameter for the restricted Robinson–Foulds metric.

In the next section we introduce some basic notation and state the main result: an optimal algorithm to enumerate the trinets displayed by a level-1 network and a cubic-time algorithm to compute the trinet distance between two level-1 networks (Theorem 1). In Section 3 we present some structural results concerning level-1 networks which we then use to prove the main result in Section 4. In Section 5 we present a comparative study between the trinet and the Robinson–Foulds metrics, in which we compute some empirical distributions for randomly generated level-1 networks. We conclude in Section 6 with a discussion of some future directions.

2. Preliminaries

Let X be a finite set of taxa with cardinality n. A *rooted phylogenetic network* (or simply a network) N on a finite set X is a simple, acyclic digraph with a unique root, no degenerate vertices (i.e., vertices with indegree one and outdegree one), whose leaves are bijectively labelled by the taxa in X. A network is *binary* if all non-leaf vertices have indegree and outdegree at most two, and all vertices with indegree two have outdegree one. A vertex is a *tree vertex* if it has outdegree two, and a *reticulation* if it has indegree two. A network is *level-k* if the maximum number of reticulations contained in any of its biconnected components is at most k. Note that a network is *level-1* if it is binary and all of its cycles (in its underlying graph) are disjoint [1] (see Fig. 1 for an example). All networks mentioned in this paper, unless stated otherwise, are level-1.

Given a network, an arc whose removal disconnects the network is a *cut arc*. If a vertex v is on a dipath from the root to a vertex u, then we say u is *below* v and v is *above* u, and write this as $u \preceq v$ (or $u \prec v$ when $u \neq v$ holds). The set $\mathcal{C}(v)$ of all taxa below a vertex v is called the *cluster* of v. A *common ancestor* of a taxon subset Y is a vertex v with $Y \subseteq \mathcal{C}(v)$. A *lowest common ancestor* (LCA) of Y is a common ancestor of Y that is not above any other common ancestors of Y. A *stable ancestor* of Y is a vertex contained in every dipath from the root

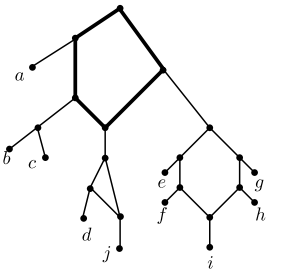
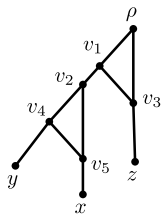


Fig. 1. A level-1 phylogenetic network with leaf set $X = \{a, b, \ldots, j\}$ containing a cycle of length five, highlighted in bold. Here we use the convention that all arcs are directed away from the root vertex which is at the top of the network.



(i)

$\mathrm{LSA}(x, y) = v_2$

$\mathrm{LSA}(x, z) = \rho$

$\mathrm{LSA}(y, z) = \rho$

(ii)

Fig. 2. An example of an LSA table: (i): A level-1 phylogenetic network N; (ii) The LSA table of N. Note that v_4 is the LCA of $\{x, y\}$ while we have $\mathrm{LSA}(x, y) = v_2$.

to some taxon in Y. The *lowest stable ancestor* (LSA) of Y is the unique vertex $\mathrm{LSA}(Y)$ such that $\mathrm{LSA}(Y)$ is below every stable ancestor of Y. Note that a LCA of Y is necessary below $\mathrm{LSA}(Y)$ (cf. [14]). Finally, the LSA *table* θ *of* N is the data structure that maps each pair of district taxa x, y to $\mathrm{LSA}(x, y) = \mathrm{LSA}(\{x, y\})$ (see Fig. 2 for an illustration).

A *binet* is a network on two taxa and a *trinet* is a network on three taxa. Up to relabelling, there exist two types of binets and eight types of trinets [9], all presented in Fig. 3. In the following, we will use the notation in that figure to refer to specific trinets and binets. Binets $T_0(x, y)$ and $S_0(x; y)$ are referred to as a *cherry* and a *reticulate cherry*, respectively. Note that a reticulate cherry is not symmetric, that is, $S_0(x; y)$ is distinct from $S_0(y; x)$.

Given a network N and a taxon subset $Y = \{y_1, \ldots, y_k\}$ of X, the network $N[Y] = N[y_1, \ldots, y_k]$ is the network obtained from N by deleting all vertices and arcs that are not on a dipath from $\mathrm{LSA}(Y)$ to some leaf in Y, and repeatedly suppressing degree 2 vertices and replacing parallel arcs by single arcs until neither operation is applicable. Let $\mathcal{B}(N)$ and $\mathcal{T}(N)$ be the set of all binets and trinets displayed by N, respectively. It is known that a level-1 network N is determined by its set $\mathcal{T}(N)$ of trinets [9].

The trinet distance $d_t(N, N')$ between two networks N and N' on the set X is the number of trinets contained in the symmetric difference $\mathcal{T}(N) \triangle \mathcal{T}(N')$ of the sets $\mathcal{T}(N)$ and $\mathcal{T}(N')$. The distance d_t is a metric on the set of level-1 networks [9]. Moreover,

$$d_t(N, N') \leq 2\binom{n}{3}, \tag{1}$$