Mathematical Biosciences 259 (2015) 12–19

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Tree-like reticulation networks—When do tree-like distances also support reticulate evolution?

Andrew R. Francis [a,*], Mike Steel [b]

[a] Centre for Research in Mathematics, School of Computing, Engineering and Mathematics, University of Western Sydney, Sydney, New South Wales, Australia
[b] Biomathematics Research Centre, University of Canterbury, Christchurch, Canterbury, New Zealand

ARTICLE INFO

Article history:
Received 6 August 2014
Revised 22 September 2014
Accepted 31 October 2014
Available online 11 November 2014

Keywords:
Phylogeny
Reticulation network
Hybridisation
Horizontal gene transfer
Distance measures

ABSTRACT

Hybrid evolution and horizontal gene transfer (HGT) are processes where evolutionary relationships may more accurately be described by a reticulated network than by a tree. In such a network, there will often be several paths between any two extant species, reflecting the possible pathways that genetic material may have been passed down from a common ancestor to these species. These paths will typically have different lengths but an 'average distance' can still be calculated between any two taxa. In this article, we ask whether this average distance is able to distinguish reticulate evolution from pure tree-like evolution. We consider two types of reticulation networks: hybridisation networks and HGT networks. For the former, we establish a general result which shows that average distances between extant taxa can appear tree-like, but only under a single hybridisation event near the root; in all other cases, the two forms of evolution can be distinguished by average distances. For HGT networks, we demonstrate some analogous but more intricate results.

1. Introduction

Evolutionary relationships between present-day taxa (species, genera etc.) are usually represented by a phylogenetic tree, which shows a branching pattern of speciation from some ancestral taxon to the taxa we observe today [1]. However, reticulate evolution is known to complicate this simple 'tree model' due to processes such as the formation of hybrid species [2], and other mechanisms where genetic material is exchanged between species (such as horizontal gene transfer (HGT)) or within a species (recombination, a process we do not consider further in this paper). Consequently, phylogenetic networks that allow 'vertical' branching through time as well as 'horizontal' reticulation events have increasingly been recognised as providing a more complete picture of much of the evolutionary history of life [3–5].

This transition has brought with it a number of mathematical and computational problems—in particular, how to reconstruct and analyse such networks, and how to distinguish different types of reticulation from tree-like evolution [6,7]. In this note we consider one aspect of the latter topic, namely the question of whether or not, if we knew the average evolutionary distance between each pair of species, we could determine whether the species network could have been a tree, or whether some more complicated reticulate history is required.

In a phylogenetic tree, the evolutionary distance between two present-day species is simply the path length from each species to the other via its most recent common ancestor (here, 'evolutionary distance' typically refers to the actual or expected amount of genetic change). However, for networks, there may be many paths linking two present-day species, and the evolutionary distance will be some average of these path lengths. Nevertheless, it is conceivable that in some cases, these distances might still appear to fit a tree exactly. We explore this question for two classes of networks: those relevant to hybrid evolution; and those relevant to HGT. Both are special cases of a more general description of (binary) 'reticulation' networks, which we now define.

1.1. Definitions: reticulation networks

Following Ref. [8], a *reticulation network* N on a finite set X is a rooted acyclic digraph (V, A) with the following properties:

(i) the *root* vertex has in-degree 0 and out-degree 2;
(ii) X is the set of vertices with out-degree 0 and in-degree 1 ('leaves');
(iii) all remaining vertices are *interior vertices*, and each such vertex either has in-degree 1 and out-degree 2 (a *tree vertex*) or in-degree 2 and out-degree 1 (a *reticulation vertex*);
(iv) the arc set A of N is the disjoint union of two subsets, the set of 'reticulation arcs' A_R and the set of 'tree arcs' A_T; moreover each reticulation arc ends at a reticulation vertex, and each reticulation vertex has at least one incoming reticulation arc;

* Corresponding author. Tel.: +61 2 9685 9236.
E-mail addresses: a.francis@uws.edu.au (A.R. Francis), mike.steel@canterbury.ac.nz (M. Steel).

http://dx.doi.org/10.1016/j.mbs.2014.10.008
0025-5564/

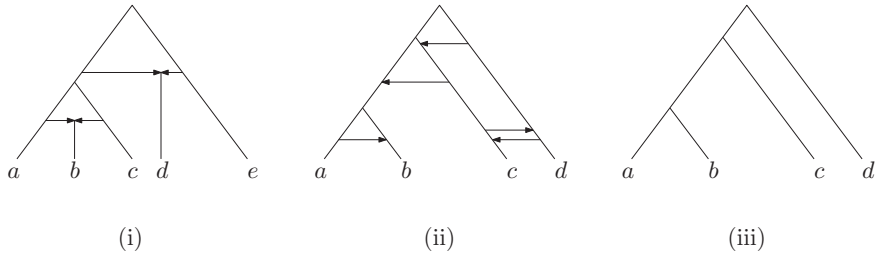

Fig. 1. (i) A hybridisation network on $\{a, b, c, d, e\}$ (usually extant species); (ii) an HGT network on $\{a, b, c, d\}$; and (iii) the tree T_N obtained from the HGT network N in (ii) by deleting all reticulation arcs. Reticulate arcs in (i) and (ii) are drawn as arrows; in each case the reticulate vertices are at the endpoints of the reticulate arcs. Note that (i) has four reticulation arcs and two reticulation vertices, while (ii) has five reticulation arcs and five reticulation vertices.

(v) every interior vertex has at least one outgoing tree arc; and
(vi) there is a function $t : V \to \mathbb{R}$ so that (a) if (u, v) is a tree arc then $t(u) < t(v)$, and (b) if (u, v) is a reticulation arc, then $t(u) = t(v)$.

Condition (vi) embodies the biological requirement that the network has a temporal representation that reflects the order of speciation events, and for which reticulation events involve two species that co-exist at some point in time.

In applications, X typically denotes a set of extant (present day) species. Two types of reticulation networks are particularly relevant in evolutionary biology (for different reasons, as we explain shortly) and these will be the main classes we will consider in this paper. The distinction is in the pair of arcs ending at a reticulation vertex in property (iii). Namely,

- in a *hybridisation network*, *both* arcs ending in a reticulation vertex are reticulation arcs, and
- in a *horizontal gene transfer network*, exactly *one* of the arcs ending in a reticulation vertex is a reticulation arc.

A simple example of each type is shown in Fig. 1.

Hybridisation networks model settings where a new species arises from members of two lineages, a process that occurs in plants, fish, and some animals [2,9], while HGT models the situation where a gene (or genes) is transferred from one species to another (a process that is common in bacteria) [10].

2. Reticulation networks and average distances

2.1. Basic properties of reticulation networks

Firstly, observe that a reticulation network N on X has no reticulation vertices if and only if N is a rooted binary phylogenetic X-tree (as defined, for example, in Ref. [11]).

Moreover, any hybridisation network is necessarily a *tree-child network*; that is, from any interior vertex in N, there is a path to a leaf that avoids any reticulation vertex. Tree-child networks have a number of desirable combinatorial and computational properties (see e.g. Refs. [12,13]).

Hybridisation networks have bounded size once $n = |X|$ is specified, since such a network can have at most $n - 2$ reticulation vertices [14]. To see this, note that in any digraph, the sum of the out-degrees equals the sum of the in-degrees so we obtain:

$$2 + 2t + r = \sum_{v \in V} \deg_{\text{out}}(v) = \sum_{v \in V} \deg_{\text{in}}(v) = n + t + 2r, \tag{1}$$

where t and r refer to the number of tree vertices and hybridisation vertices, respectively. Note that each hybridisation vertex corresponds to two parent tree vertices, and hence $t \geq 2r$ in a hybridisation network. Eq. (1) gives $n = t + 2 - r$, and using $t \geq 2r$ we obtain:

$$r \leq n - 2. \tag{2}$$

A consequence of this bound is that, up to isomorphism, there are only finitely many hybridisation networks for any given n (the enumeration of hybridisation networks has recently been investigated by McDiarmid et al. [14]).

By contrast, an HGT network with a given leaf set X can have arbitrarily many reticulation vertices, and so there are infinitely many HGT networks for a given X. However, an HGT network N has a useful property that is absent in a hybridisation network: an HGT network always has an associated canonical rooted binary phylogenetic X-tree T that is obtained from N by deleting all the reticulation arcs (and suppressing any resulting vertices that have both in-degree 1 and out-degree 1). We denote this tree with the notation T_N (an example is shown in Fig. 1).

Given any reticulation network N on X, suppose that for each reticulation vertex, we delete exactly one of the in-coming arcs. The resulting graph is a rooted tree with leaf set X and a root that coincides with the root of N. Moreover, if we suppress any resulting vertices that have both in-degree 1 and out-degree 1 we obtain a rooted binary phylogenetic X-tree, T. We say that T is *displayed* by N and we let $\mathcal{T}(N)$ denote the set of all the (at most) 2^r such trees that are displayed by N.

2.2. Tree metrics

Consider any unrooted phylogenetic X-tree $T = (V, E)$ together with a weight function $w : E \to \mathbb{R}^{>0}$ that assigns strictly positive weights to each edge of the tree. Then (T, w) induces a distance function on X as follows: For each pair of leaves x, y on a tree T, the *tree distance* between them is defined as the sum of the weights of the edges that lie on the (unique) path in T connecting x and y. That is:

$$d_{(T,w)}(x, y) := \sum_{e \in P(T;x,y)} w(e),$$

where if $x = y$ we set $d_{(T,w)}(x, y) = 0$ (the empty path has length zero). The resulting function $d_{(T,w)} : X \times X \to \mathbb{R}^{\geq 0}$ is a metric on X.

A metric on X that can be represented in this way on some phylogenetic X-tree is said to be a *tree metric*. This holds if and only if the metric satisfies the 'four-point condition'. This states that for any four (not necessarily distinct) points u, v, w, y from X, two of the three sums $d(u, v) + d(w, y)$; $d(u, w) + d(v, y)$; $d(u, y) + d(v, w)$ are equal, and are greater than or equal to the other one. This classic characterisation of tree metrics dates back to the 1960s (for more recent treatments, see Refs. [11,15]). Moreover, if d is a tree metric on X, then d can be written $d = d_{(T,w)}$ for precisely one choice of the pair (T, w), where T is a phylogenetic X-tree, and w a strictly positive edge weight function. In the case where T is binary, we will say that d is a *binary tree metric*.

2.3. Average distances on networks

A reticulation network can be thought of as a 'weighted union' of the trees displayed by N. We formalise this idea, and extend it to bring in distances, as follows:

For each vertex v in the set V_R of reticulation vertices of N, let $R(v)$ denote the two arcs that end at v. Suppose we are given a reticulation network $N = (V, A)$ on X along with: