Discrete Applied Mathematics 213 (2016) 47–59

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

An algorithm for reconstructing ultrametric tree-child networks from inter-taxa distances

M. Bordewich, N. Tokac *

School of Engineering and Computing Sciences, Durham University, Lower Mountjoy, South Road, Durham, DH1 3LE, UK

ARTICLE INFO

Article history:
Received 4 September 2015
Received in revised form 25 April 2016
Accepted 11 May 2016
Available online 11 June 2016

Keywords:
UPGMA
Phylogenetic networks
Ultrametric networks

ABSTRACT

Traditional "distance based methods" reconstruct a phylogenetic tree from a matrix of pairwise distances between taxa. A phylogenetic network is a generalisation of a phylogenetic tree that can describe evolutionary events such as reticulation and hybridisation that are not tree-like. Although evolution has been known to be more accurately modelled by a network than a tree for some time, only recently have efforts been made to directly reconstruct a phylogenetic network from sequence data, as opposed to reconstructing several trees first and then trying to combine them into a single coherent network. In this work we present a generalisation of the UPGMA algorithm for ultrametric tree reconstruction which can accurately reconstruct ultrametric tree-child networks from the set of distinct distances between each pair of taxa.

1. Introduction

The evolutionary history of organisms is generally represented by a phylogenetic tree. One popular and fast method for reconstructing phylogenetic tree from DNA or protein sequence data is to first compute a matrix of pairwise distances between the taxa, and then infer the phylogenetic tree from this distance matrix. Such approaches are called *distance-based* methods, and they are very widely used due to their simplicity and speed. The two most well known and long standing approaches are UPGMA [13] and Neighbour Joining [11]. In recent years several variants of these and new approaches have been suggested, including Least Squares [8], BioNJ [9] and Balanced Minimum Evolution [7]. The properties and accuracy of distance based methods have been widely studied, see for example [2,3,10].

In this paper, we consider the task of reconstructing phylogenetic networks from distance data. A phylogenetic network is a generalisation of a phylogenetic tree, which can be used to describe the evolutionary history of a set of species that is non-tree like because of reticulation events such as hybridisation, horizontal gene transfer or recombination. The reconstruction of restricted classes of phylogenetic network from inter-taxa distances have been studied in a number of recent papers. A key feature of this problem is that in a network there is no longer a unique distance between a pair of taxa (as there is in a tree), so one must work with shortest distances, average distances or sets or subsets of distances. Chan et al. [6] take a matrix of inter-taxa distances and reconstruct an ultrametric galled network (more commonly called a galled tree or a level-1 network) such that there is a path between each pair of taxa having the weight given in the matrix, if such network exists. Willson [14] studied the problem of determining the network given the average distance between taxa, where each reticulation vertex assigns a probability to its two incoming arcs. He manages the reconstruction of phylogenetic networks which have a single reticulation cycle from such distances in polynomial time [15]. In a recent paper [4], Bordewich and Semple showed that

* Corresponding author.
E-mail addresses: m.j.r.bordewich@durham.ac.uk (M. Bordewich), nihan.tokac@durham.ac.uk (N. Tokac).

http://dx.doi.org/10.1016/j.dam.2016.05.011
0166-218X/

(unweighted) tree-child phylogenetic networks may be reconstructed from the multi-set of path lengths between taxa and that temporal, tree-child, phylogenetic networks may be reconstructed from the set of path lengths between taxa, each in polynomial time in the size of the input.

In this paper, which builds on and extends the approach of [4], we present a polynomial-time algorithm (which we have called NetworkUPGMA) that reconstructs an ultrametric tree-child network from the set of distances between each pair of taxa. Our algorithm offers an improvement over previous works in two ways. First ultrametric tree-child networks are a much wider class of networks than networks with only a single reticulation or ultrametric galled networks, which are a subclass of ultrametric tree-child networks. In particular note that: the total number of reticulations in a tree-child network on n taxa can be as large as $n-1$ [5], whereas a galled network has at most $n/2$ reticulations; and the interrelation of reticulations may be more complex, as each 2-connected component of our networks may contain many reticulations (again linear in the number of taxa), whereas in a galled network there can only be one reticulation in each 2-connected component. Second, the algorithm takes the *set of distances* between each pair of taxa as input, where Bordewich and Semple [4] required the *multi-set of path lengths* (for unweighted tree-child networks). This is an important distinction: the distance matrices come from estimating evolutionary distance based upon sequence data of some type. Real phylogenies are weighted: edge weights correspond to some measure of genetic difference. Furthermore, while it is quite conceivable that by sampling different genes or regions of the genome one might build up an accurate picture of the set of different evolutionary path weights between a given pair of taxa, it seems hard to imagine how one might manage to measure the number of distinct evolutionary paths of a given observed weight. Thus the set of distances seems a much more reasonable input for an algorithm in practice. (Note, however, that only through study of the multi-set problem did we gain the understanding needed to tackle this newer work.)

2. Definitions and statement of results

In this section we formally define the central concepts of phylogenetic networks and give further definitions which we shall require in order to present our algorithm and proof. Throughout the paper, standard notation and terminology follows Semple and Steel [12]. X denotes a non-empty finite set of taxa. A rooted phylogenetic X-tree $\mathcal{T}$ 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. An X-tree is binary if either $|X| = 1$ or the root has degree two and every other interior vertex has degree three.

2.1. Ultrametric tree-child networks

A phylogenetic network $\mathcal{N}$ on X is a rooted, connected, directed acyclic graph with the following properties:

(i) exactly one node (the root) has in-degree 0 and all other nodes have in-degree 1 or 2,
(ii) any node with in-degree 2 (called a reticulation) has out-degree 1 and all other nodes have out-degree 0 (called leaves) or 2 (called tree vertices), and
(iii) each node with out-degree 0 is labelled with a distinct element of X (taxon).

Note that, what we have called a phylogenetic network is sometimes referred to as a *binary* phylogenetic network.

A network $\mathcal{N}$ is *weighted* if there is a positive weighting (or length) associated with each arc, which is strictly positive for all tree arcs (those arcs whose head is a tree vertex or leaf). For arc $e = (u, v)$ the weight is denoted by l_e or $l(u, v)$. The weight of a path is the sum of the weights of arcs it contains. An *ultrametric network* is a weighted phylogenetic network such that every directed path from the root to any leaf has the same weight [1,6]. This implies that for any vertices u, v such that there is a directed path from u to v in $\mathcal{N}$, every path from u to v has the same weight, which we denote $d_{u,v}$.

Let $\mathcal{N}$ be a phylogenetic network on X. For any two vertices u and v in $\mathcal{N}$ that are joined by an arc (u, v), we say u is a parent of v and, conversely, v is a child of u. Cardona et al. [5] discussed "tree-child" networks, in which every vertex that is not a leaf has a child that is a tree vertex or leaf. We say an ultrametric network is *ultrametric tree-child network* if every non-leaf has a child which is either a tree vertex or a leaf. For vertices u, v such that there is a directed path from u to v in $\mathcal{N}$, we say the path is a *tree-path* if every vertex on the path, except possibly u, is a tree vertex or a leaf. Note that in a tree-child network every vertex has a tree-path to a leaf.

2.2. Distance matrices

Given a phylogenetic network $\mathcal{N}$ on X, we define the *set-distance matrix $\mathcal{D}$ of inter-taxa distances* as follows. For any two elements $x, y \in X$, an up–down path from x to y is an underlying path $x, v_1, v_2, \ldots, v_{k-1}, y$ in $\mathcal{N}$ such that, for some $i \leq k-1$, $\mathcal{N}$ contains the arcs

$$(v_i, v_{i-1}), (v_{i-1}, v_{i-2}), \ldots, (v_1, x)$$

and

$$(v_i, v_{i+1}), (v_{i+1}, v_{i+2}), \ldots, (v_{k-1}, y).$$

The weight of an up–down path is the sum of the weights of the two directed paths it contains. The vertex v_i is called the peak of the up–down path. In any rooted network $\mathcal{N}$, a *least common ancestor* of two vertices x and y is a vertex v such