Journal of Theoretical Biology 269 (2011) 245–255

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

When two trees go to war

Leo van Iersel[b,*,2], Steven Kelk[a,**,1]

[a] Centrum voor Wiskunde en Informatica (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[b] University of Canterbury, Department of Mathematics and Statistics, Private Bag 4800, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 9 June 2010
Received in revised form
26 October 2010
Accepted 26 October 2010
Available online 31 October 2010

Keywords:
Reticulation
Phylogenetic network
Cluster
Triplet
Character

ABSTRACT

Rooted phylogenetic networks are used to model non-treelike evolutionary histories. Such networks are often constructed by combining trees, clusters, triplets or characters into a single network that in some well-defined sense simultaneously represents them all. We review these four models and investigate how they are related. Motivated by the parsimony principle, one often aims to construct a network that contains as few *reticulations* (non-treelike evolutionary events) as possible. In general, the model chosen influences the minimum number of reticulation events required. However, when one obtains the input data from two binary (i.e. fully resolved) trees, we show that the minimum number of reticulations is independent of the model. The number of reticulations necessary to represent the trees, triplets, clusters (in the softwired sense) and characters (with unrestricted multiple crossover recombination) are all equal. Furthermore, we show that these results also hold when not the number of reticulations but the level of the constructed network is minimised. We use these unification results to settle several computational complexity questions that have been open in the field for some time. We also give explicit examples to show that already for data obtained from three binary trees the models begin to diverge.

1. Introduction

One of the main challenges in phylogenetics is to reconstruct evolutionary histories from biological data of currently living organisms. The traditional and most widely used model for representing evolutionary histories is the phylogenetic tree. However, recent years have seen more and more interest in the generalisation of phylogenetic trees to phylogenetic networks, which can model non-treelike evolution. These phylogenetic networks contain special nodes, called *reticulations*, in which previously diverged lineages recombine. These nodes represent "reticulate" evolutionary phenomena such as hybridisation, recombination or lateral (horizontal) gene transfer. For a full overview of theory and methods concerning phylogenetic networks, see Huson et al. (to appear), Nakhleh (2009), and Semple (2007).

Motivated by the parsimony principle, a phylogenetic network with fewer reticulations is often preferred over a network with more reticulations, when both networks represent the available data equally well. Alternatively, one can aim to minimise the "level" of the constructed network, i.e. the number of reticulations per tangled part of the network, see Fig. 1. Thus, it is interesting to compute the minimum number of reticulations, or alternatively the minimum level, necessary to represent certain data by a phylogenetic network.

How these minima depend on the chosen model is still very poorly understood. Many algorithms and software packages (see Huson et al., to appear; Nakhleh, 2009; Semple, 2007 and the overview we give in Section 2) are available for many different models, but how these models are related, and whether they are essentially different, often remains undiscussed. This article illuminates the relation between several such models. The special case of an input consisting of two phylogenetic trees has been discussed repeatedly in different contexts (Bordewich et al., 2007; Bordewich and Semple, 2007a, b; Collins et al., to appear; Huson et al., 2009; van Iersel et al., 2010; Wu and Jiayin, 2010). We take a closer look at this special case and show that it is indeed very special: three fundamentally different models turn out to be, in this special case, equivalent.

We focus on four models for the construction of phylogenetic networks. Probably the most natural one is the "tree-model" which aims at combining several phylogenetic trees into a single phylogenetic network that precisely displays each of the trees; e.g., see Baroni et al. (2005). This is especially interesting when certain parts of the genome (e.g. genes) are known to have evolved in a tree-like

[*] Corresponding author.
[**] Principal corresponding author. Tel.: +31 20 5924265; fax: +31 20 5924199.
E-mail addresses: l.j.j.v.iersel@gmail.com (L. van Iersel), S.M.Kelk@cwi.nl (S. Kelk).
[1] The author was funded by a Computational Life Sciences grant of The Netherlands Organisation for Scientific Research (NWO).
[2] The author was funded by the Allan Wilson Centre for Molecular Ecology and Evolution.

0022-5193 © 2010 Elsevier Ltd. Open access under the Elsevier OA license.
doi:10.1016/j.jtbi.2010.10.032

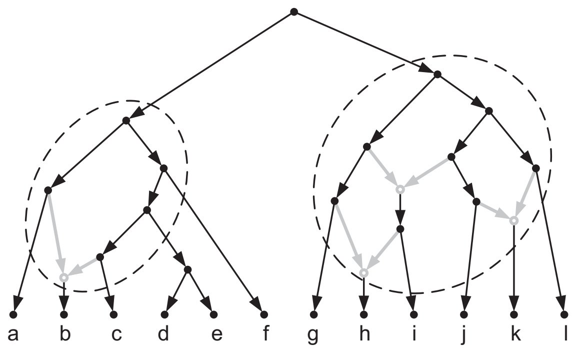

Fig. 1. A phylogenetic network with four reticulations (grey, unfilled vertices). This is a level-3 network, because the tangled parts (encircled) contain at most three reticulations each.

fashion. One can then generate a phylogenetic tree for each tree-like part of the genome separately, and combine them into a phylogenetic network that represents each of the trees.

Another model is to extract a set of *triplets* (phylogenetic trees with three taxa each) and to combine them into a phylogenetic network that represents each of the triplets; e.g., see van Iersel et al. (2008). Triplets can be constructed in two ways. Firstly, one can use any of the methods for constructing phylogenetic trees for some or all combinations of three taxa (using a fourth taxon as an outgroup in order to root the triplet). Alternatively, one can first construct one or more phylogenetic trees (on all taxa) and subsequently find the set of triplets that are contained in these trees. The main motivation for the latter approach is that representing all triplets might require fewer reticulations than representing the entire trees. In Section 3.3, we indeed give an explicit example of three trees for which the triplets in the trees can be represented with fewer reticulations than necessary to represent the trees themselves. On the other hand, this section also shows that, for two fully resolved trees, the number of reticulations needed to represent the trees or the triplets in the trees are always the same. Moreover, these results also hold when the level rather than the total number of reticulations is minimised.

A third model extracts a set of *clusters* and combines those into a phylogenetic network; e.g., see Huson et al. (2009). Clusters can be obtained from morphological data or from phylogenetic trees. The latter approach has a similar motivation as in the triplet-model. The clusters from the trees might be representable using fewer reticulations than that would be necessary to represent the trees themselves. In addition, the clusters described by a phylogenetic tree are biologically the most interesting features of the tree, because they describe putative monophyletic groups of species (also called clades). In Section 3.2, we show that clusters are in some sense 'between' triplets and trees. The number of reticulations required by the triplets is always less than or equal to the number of reticulations required by the clusters, and this latter number is in turn less than or equal to the number of reticulations required to represent the trees themselves. In Section 3.3, we give examples of sets of three trees for which these inequalities are strict. However, in this section we also show that, for two fully resolved trees, the number of reticulations needed to represent the clusters is always equal to the number of reticulations needed to represent the triplets or trees. We again show that all these results also hold when the level rather than the total number of reticulations is minimised.

The last model we consider in this article constructs phylogenetic networks from *binary characters*. This kind of data consists of a matrix of 0's and 1's and can for example be constructed from DNA, morphological data or phylogenetic trees. Binary characters have been well studied in the field of population genetics (Song et al., 2005). In Section 3.1, we clarify the relation between this model and the cluster model mentioned above, to put our main results in the correct context.

The structure of the remainder of this article is as follows. The next section describes the mathematical models in detail, gives an overview of known results for each model, and summarises our results. In Section 3 we prove our unification results and in Section 4 we use these results to prove several computational complexity results. We end the article in Section 5 with some concluding remarks.

2. Mathematical models and summary of results

2.1. Phylogenetic networks

Consider a set of taxa $\mathcal{X}$. A *rooted phylogenetic network* on $\mathcal{X}$ is a directed acyclic graph with exactly one vertex with indegree-zero (the *root*) in which the outdegree-zero nodes (the *leaves*) are bijectively labelled by $\mathcal{X}$. It is common to identify a leaf with the taxon it is labelled by and it is usually assumed that there are no nodes with indegree and outdegree one; we adopt both conventions. Nodes with indegree at least two are called *reticulations*. The edges entering a reticulation are called *reticulation edges*. Nodes that are not reticulations are called *tree nodes*. A phylogenetic network is called *binary* (or fully resolved) if all reticulations have indegree two and outdegree one and all other nodes have outdegree zero or two. In this article we only consider rooted (as opposed to unrooted) phylogenetic networks and for this reason we henceforth omit the prefix "rooted".

As mentioned before, we are interested in minimising either the number of reticulation events or the level of the constructed network. The following subtlety has to be taken into account when reticulations with indegree higher than two are considered. When counting such reticulations, indegree-d reticulations are counted $d-1$ times, because such reticulations represent $d-1$ reticulate evolutionary events (of which the order is not specified). Hence, using $\delta^-(v)$ to denote the indegree of a node v, we formally define the *number of reticulations* in a phylogenetic network $N=(V,E)$ as

$$\sum_{v \in V:\delta^-(v)>0} (\delta^-(v)-1) = |E|-|V|+1.$$

Thus, we define the following fundamental problem MinRet. Given some data describing some taxa, find a phylogenetic network that "represents" the given data and contains a minimum number of reticulations over all phylogenetic networks that represent the given data. We consider three specific variants of this problem: MinRetTrees, MinRetTriplets and MinRetClusters, for data consisting of trees, triplets and clusters respectively.

Let us now formally define the level of a phylogenetic network. A *biconnected component* is a maximal subgraph that cannot be disconnected by removing a single node. A biconnected component is *trivial* if it is equal to a single edge and *nontrivial* otherwise. For $k \in \mathbb{N}$, a phylogenetic network is called a *level-k* network if each nontrivial biconnected component contains at most k reticulations. See Fig. 1 for an example of a phylogenetic network with four reticulations. The grey, unfilled vertices are reticulations and the grey edges are reticulation-edges. This is a level-3 network, because the nontrivial biconnected components (encircled by dashed lines) contain at most three reticulations each.

We are now ready to define the following MinLev variant of the fundamental problem. Given some data describing some taxa, find a level-k phylogenetic network that "represents" the given data such that k is as small as possible. There are again three versions: MinLevTrees, MinLevTriplets and MinLevClusters, for data consisting of trees, triplets and clusters respectively.