1 Phylogenetic Networks I: Combinatorial Aspects

A fundamental problem in evolutionary biology is to represent the ancestral history of a collection of present-day species with a phylogenetic (evolutionary) tree. However, as the result of evolutionary processes such as lateral gene transfer and hybridisation, phylogenetic trees are insufficient for many collections:

> *... molecular phylogeneticists will have failed to find the 'true tree', not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot be properly represented as a tree.*
>
> Dolittle (1999)

Evolutionary events that include lateral gene transfer and hybridisation cause non-tree-like evolution. Collectively, referred to as reticulation events, these types of processes result in species being a composite of DNA regions derived from different ancestors.

The effect of reticulation in evolution has long been recognised. But the mathematical study of phylogenetic networks, the natural analogue of phylogenetic trees for non-tree-like evolution, is very recent. These notes provide a mathematical introduction to phylogenetic networks.

A word of warning. The notes are not meant to cover all aspects of phylogenetic networks. Their intended purpose is to give a taste of current problems and to highlight the complexities when working with phylogenetic networks.

Throughout these notes, X always denotes a non-empty finite set. A *rooted phylogenetic X-tree* $\mathcal{T}$ is a rooted tree in which the root has degree at least two and all other interior vertices have degree at least three, and whose leaf set is X. If $|X| = 1$, then $\mathcal{T}$ consists of the single vertex in X. In addition, $\mathcal{T}$ is *binary* if either $|X| = 1$ or the root has degree two and every other interior vertex has degree three.

For the most part, we are only interested in rooted binary phylogenetic trees and so, unless stated otherwise, we will always refer to a 'rooted binary phylogenetic tree' as a 'phylogenetic tree'.

Note. Pictorially, X represents a collection of present-day species and the root represents a hypothetical ancestor common to each of the species in X.

A *phylogenetic network* $\mathcal{N}$ *on* X is a rooted acyclic digraph with the following properties:

(i) the root has out-degree two;

(ii) a vertex with out-degree zero has in-degree one, and the set of vertices with out-degree zero is X;

(iii) all other vertices either have in-degree one and out-degree two, or in-degree two and out-degree one.

The vertices of $\mathcal{N}$ with in-degree two and out-degree one are called *reticulation vertices* (or simply *reticulations*), while the vertices of in-degree one and out-degree two are called *tree vertices*. The vertices of X, that is, the vertices with in-degree one and out-degree zero are called *leaves*. The edges directed into a reticulation are called *reticulation edges*. All other edges are called *tree edges*.

Note. If $|X| = 1$, then $\mathcal{N}$ consists of the single vertex in X. Furthermore, we do not allow parallel edges. Also, phylogenetic networks as described here are commonly referred to as binary phylogenetic networks in the literature.

Observe that a phylogenetic tree is a phylogenetic network with no reticulations.

Example 1.1.

Not surprisingly, phylogenetic networks provide a much richer, and thus more complicated, class of mathematical objects than phylogenetic trees. For example, the size

(number of vertices) of a phylogenetic tree is bounded by its number of leaves. In general, the size of a phylogenetic network is not bounded by its number of leaves. Why? (Can you think of an example?). Also, for a given pair of vertices u and v, it is possible for a phylogenetic network to have numerous (*underlying*) paths between u and v, whereas, a phylogenetic tree has exactly one (underlying) path.

Question. Find a phylogenetic network whose total number of vertices is not bounded by its number of leaves.

1.1 Tree-child networks

Let $\mathcal{N}$ be a phylogenetic network. If u and v are vertices of $\mathcal{N}$, and (u, v) is an edge of $\mathcal{N}$, we say that u is a *parent* of v and v is a *child* of u. A phylogenetic network $\mathcal{N}$ is a *tree-child network* if every vertex is either a leaf or has a child that is a tree vertex.

Example 1.2.

Question. Draw a phylogenetic network that is not a tree-child network.

Tree-child networks are an increasingly popular class of phylogenetic networks in the literature. They are general enough to be interesting, but also constrained enough to make many problems tractable.

Let $\mathcal{N}$ be a phylogenetic network on X with root ρ. Let v be a vertex of $\mathcal{N}$. We say that v has the *tree-path property* if there is a leaf ℓ in X such that there is a (directed) path from v to ℓ containing no reticulations except possibly v. Furthermore, v is *visible* if there is a leaf ℓ in X such that every (directed) path from ρ to ℓ traverses v.

Example 1.3.

Question. Draw a phylogenetic network with a vertex that is not visible.

Theorem 1.1. Let $\mathcal{N}$ be a phylogenetic network. Then the following statements are equivalent.

(i) $\mathcal{N}$ is a tree-child network.

(ii) Every vertex of $\mathcal{N}$ has the tree-path property.

(iii) Every vertex of $\mathcal{N}$ is visible.

Proof.

1.2 The Tree Containment problem

In this section and the next, we consider two combinatorial problems. Both problems centre around the concept of phylogenetic network 'displaying' a phylogenetic tree.

Although the evolution of certain species needs to be described with reticulation events, the evolution of a particular gene can generally be described without reticulation events. As a result, analysing the tree-like information in a phylogenetic network has become a common task.

Let $\mathcal{T}$ be a phylogenetic X-tree and let $\mathcal{N}$ be a phylogenetic network on X. We say that $\mathcal{N}$ *displays* $\mathcal{T}$ if $\mathcal{T}$ can be obtained from $\mathcal{N}$ by deleting edges and vertices, and contracting degree-2 vertices.

Note. Intuitively, $\mathcal{N}$ displays $\mathcal{T}$ if the ancestral history of the elements in X inferred by $\mathcal{T}$ is also inferred by $\mathcal{N}$.

Example 1.4.

Question. Let $\mathcal{N}$ be a phylogenetic network with k reticulations. How many phylogenetic X-trees are displayed by $\mathcal{N}$?

The first combinatorial problem we are interested in is the following:

Tree Containment
Instance: A phylogenetic network on X and a phylogenetic X-tree.
Question: Does $\mathcal{N}$ display $\mathcal{T}$?

In general, Tree Containment is NP-hard, that is, there is unlikely to exist a method for solving it that doesn't involve an exponential number of 'trial and errors'. However, for the class of tree-child networks, Tree Containment is 'polynomial-time' in the size of $\mathcal{N}$. To prove this, we make use of the next lemma.

Let $\mathcal{N}$ be a phylogenetic network on X and let $\{a, b\}$ be a two element subset of X. We say that $\{a, b\}$ is a *cherry* in $\mathcal{N}$ if x and y have the same parent. Furthermore, let p_a and p_b be the parents of a and b, respectively. Then $\{a, b\}$ is a *reticulated cherry* if either p_a is a reticulation and p_a is a child of p_b, or p_b is a reticulation and p_b is a child of p_a.

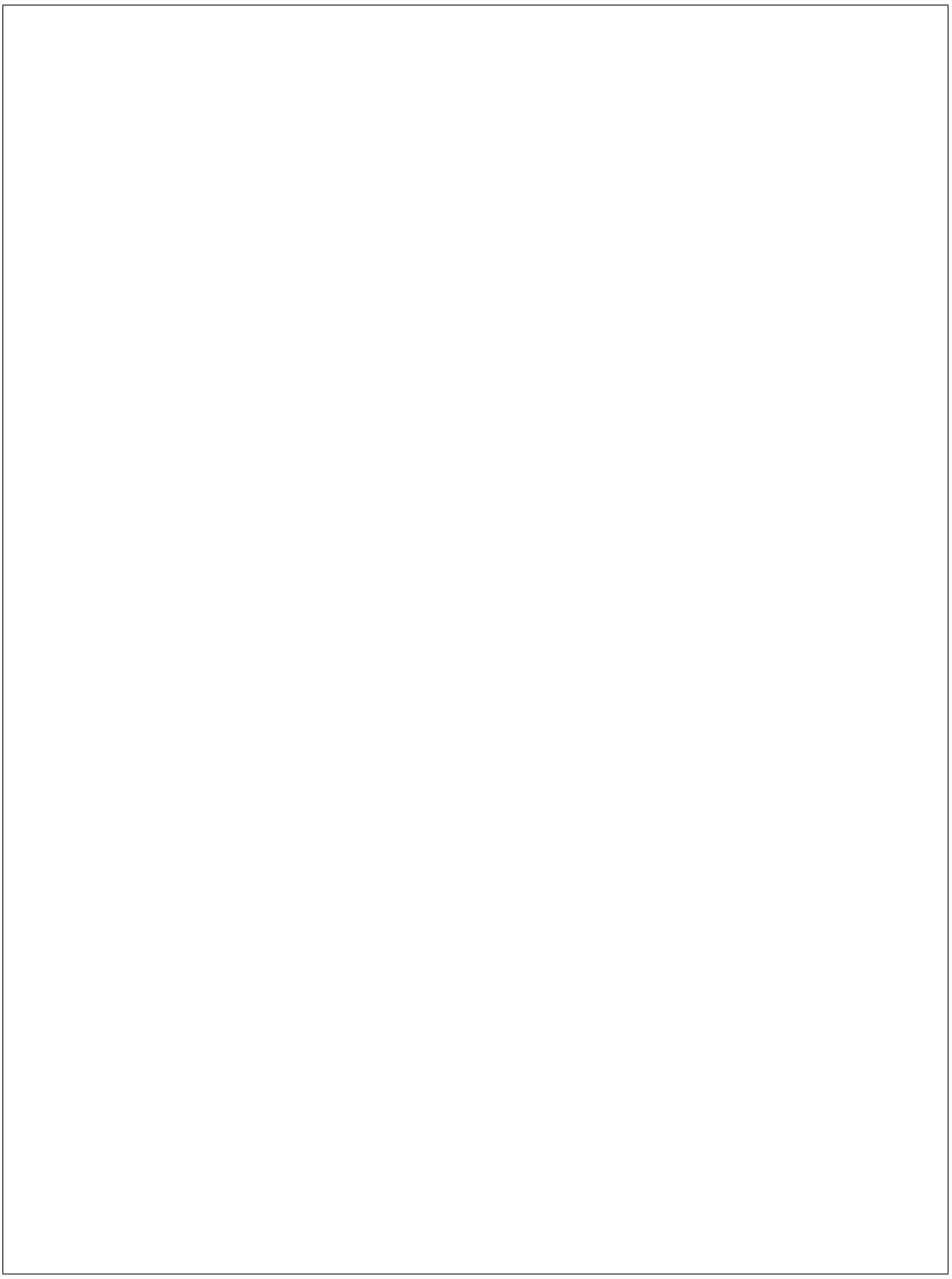
Example 1.5.

Lemma 1.2. Let $\mathcal{N}$ be a tree-child network on X, and suppose that $|X| \geq 2$. Then there is a 2-element subset $\{a, b\}$ of X such that either $\{a, b\}$ is a cherry in $\mathcal{N}$ or $\{a, b\}$ is a reticulated cherry in $\mathcal{N}$.

Proof. Left as an exercise. □

Theorem 1.3. Let $\mathcal{N}$ be a tree-child network on X and let $\mathcal{T}$ be a phylogenetic X-tree. Then Tree Containment for $\mathcal{N}$ and $\mathcal{T}$ can be decided in polynomial time.

Proof.



1.3 The Tree Based problem

In this section, we consider the second of our two combinatorial problems.

Let $\mathcal{N}$ be a phylogenetic network on X. We say that $\mathcal{N}$ is a *tree-based network* if it can be obtained from a phylogenetic X-tree by adjoining new edges whose end-vertices 'subdivide' edges of $\mathcal{T}$, in which case, $\mathcal{T}$ is a *base tree* of $\mathcal{N}$.

Example 1.6.

Note. The order in which the new edges are adjoined is not important. If $\mathcal{T}$ is a base tree of a tree-based network $\mathcal{N}$, then it is not necessarily the only such tree.

Question. If $\mathcal{T}$ is a base tree of a tree-based network $\mathcal{N}$, then $\mathcal{N}$ displays $\mathcal{T}$. But is every phylogenetic tree displayed by a tree-based network a base tree of $\mathcal{N}$?

Question. Is every phylogenetic network a tree-based network?

The concept of a tree-based network was introduced to quantify the notion that a phylogenetic network is simply just a phylogenetic tree with additional edges. This is particularly relevant to the longstanding debate of whether the evolution of certain species is tree-like with reticulation or whether the possibility of an underlying phylo-

genetic tree should be dispensed with altogether. The problem we are interested in is the following:

Tree Based
Instance: A phylogenetic network $\mathcal{N}$.
Question: Is $\mathcal{N}$ a tree-based network?

Much less is known about this problem than that of Tree Containment. One can show directly that every tree-child network is a tree-based network. However, it is also a corollary of the next theorem. (Why?)

Theorem 1.4. Let $\mathcal{N}$ be a phylogenetic network and suppose that $\mathcal{N}$ has no reticulations u and v such that u is a parent of v. Then $\mathcal{N}$ is a tree-based network.

Proof. Left as an exercise. This is more difficult that the previous exercises and questions. Nevertheless, have a go at proving it! □

Open problem. Find a condition that is both necessary and sufficient for a phylogenetic network to be a tree-based network.