

# Beyond Galled Trees - Decomposition and Computation of Galled Networks

Daniel H. Huson and Tobias H. Klöpper

Center for Bioinformatics (ZBIT), Tübingen University,  
Sand 14, 72076 Tübingen, Germany

**Abstract.** Reticulate networks are a type of phylogenetic network that are used to represent reticulate evolution involving hybridization, horizontal gene transfer or recombination. The simplest form of these networks are galled trees, in which all reticulations are independent of each other. This paper introduces a more general class of reticulate networks, that we call galled networks, in which reticulations are not necessarily independent, but may overlap in a tree-like manner. We prove a Decomposition Theorem for these networks that has important consequences for their computation, and present a fixed-parameter-tractable algorithm for computing such networks from trees or binary sequences. We provide a robust implementation of the algorithm and illustrate its use on two biological datasets, one based on a set of three gene-trees and the other based on a set of binary characters obtained from a restriction site map.

## 1 Introduction

Phylogenetic networks are graphs used for representing phylogenetic relationships between different taxa, and are usually employed when a tree representation does not suffice. There are many different types of phylogenetic networks and it is useful to distinguish between two main classes: *implicit* phylogenetic networks that provide tools to visualize and analyze incompatible phylogenetic signals, such as split networks [1, 2], and *explicit* phylogenetic networks that provide explicit scenarios of reticulate evolution, such as hybridization networks [3, 4, 5, 6, 7], HGT networks [8] and recombination networks [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

Although the latter three types of networks apply to quite different evolutionary scenarios, they share the common feature that they all contain *reticulation nodes* at which new sequences arise as a combination of sequences from two different predecessor sequences. Such networks are mathematically similar to each other and are collectively referred to as *reticulate networks*, which are the focus of this paper.

The different types of input to methods that compute these networks also share similarities and can be generally considered as *splits*, that is, bipartitionings of the underlying dataset, defined either by the edges of the input trees, in the case of hybridization or HGT networks [6], or by the non-constant columns of the alignments of binary sequences in the case of recombination networks [17].

Gusfield et al. [12, 16] introduce the term *galled tree*, that can be defined as a reticulate network in which no two reticulation nodes are contained in a common unoriented cycle, and provide an algorithm for computing this type of networks. To be precise, their definition requires all cycles to be node-disjoint. However, the combinatorial analysis of galled trees also works if we require cycles to be only edge-disjoint, and so we prefer to require only the latter property.

Interest in galled trees is based on the fact that they are computationally tractable. However, there is little reason to believe that reticulate networks have this simple structure in practice and so an understanding and treatment of reticulate networks of a more general nature is desirable. In [6] we provide an algorithm for computing reticulate networks that goes slightly beyond galled trees and accommodates reticulate networks in which reticulations may overlap along paths. The goal of this paper is to introduce a more general class of reticulate networks that go substantially beyond galled trees.

Given an input dataset, such as a collection of trees, or a multiple alignment of binary sequences, the computational goal is to determine a reticulate network that “explains” the given dataset, in terms of mutations, speciations and reticulate events, such as recombinations, HGT or hybridizations. For any given dataset, many different networks may exist that can explain it. Always, at least one exists, as shown in [20]. Considering reticulations to be expensive evolutionary events, one will prefer a most parsimonious solution. In the context of phylogenetic trees and hybridization networks, one may attempt to minimize the number of reticulation nodes. In the context of recombination networks, an alternative optimization goal is to minimize the number of recombination cross-overs, but we do not address this here.

The associated computational problem is NP-complete in full generality [21, 7]:

**Problem 1 (Most Parsimonious Reticulate Network).** *For a given input set  $\Delta$  (consisting of trees, binary characters or splits, depending on the concrete application), determine a reticulate network  $N$  that explains  $\Delta$  using a minimum number of reticulation nodes.*

Decomposition Theorems, which aim at dividing the task into independent sub-problems that can be identified by the pattern of incompatibilities in the input [16, 6], are an important tool for addressing Problem 1.

This paper makes five theoretical and practical contributions in the area of phylogenetic networks. Firstly, we introduce a natural generalization of galled trees, which we call *galled networks*, that go substantially beyond galled trees. This generalization allows for quite complicated configurations of reticulations, as present in real data. Secondly, we prove a Decomposition Theorem for galled networks that settles an open conjecture posed by Dan Gusfield at RECOMB 2005 [16] for this class of networks. Thirdly, we provide a fixed-parameter-tractable algorithm for computing galled networks from real data. Fourthly, we provide an implementation of our algorithm as a plug-in for SplitsTree4 [2], thus making the algorithm easily available to the community. Finally, we illustrate our results on two published datasets, one that uses three different gene trees to study the evolution of a set of fungal species [22] and the other that uses