

Reducing Distortion in Phylogenetic Networks

Daniel H. Huson¹, Mike A. Steel², and Jim Whitfield³

¹ Center for Bioinformatics (ZBIT), Tübingen University, Germany
`huson@informatik.uni-tuebingen.de`

² Allan Wilson Centre, University of Canterbury, Christchurch, New Zealand
`m.steel@math.canterbury.ac.nz`

³ Department of Entomology, University of Illinois at Urbana-Champaign, USA
`jwhitfie@life.uiuc.edu`

Abstract. When multiple genes are used in a phylogenetic study, the result is often a collection of incompatible trees. Phylogenetic networks and super-networks can be employed to analyze and visualize the incompatible signals in such a data set. In many situations, it is important to have control over the amount of incompatibility that is represented in a phylogenetic network, for example reducing noise by removing splits that do not recur among the source trees. Current algorithms for computing hybridization networks from trees are based on a combinatorial analysis of the arising set of splits, and are thus sensitive to false positive splits. Here, a filter is desirable that can identify and remove splits that are not compatible with a hybridization scenario. To address these issues, the concept of the distortion of a tree relative to a split is defined as a measure of how much the tree needs to be modified in order to accommodate the split, and some of its properties are investigated. We demonstrate the usefulness of the approach by recovering a plausible hybridization scenario for buttercups from a pair of gene trees that cannot be obtained by existing methods. In a second example, a set of seven gene trees from microgastrine braconid wasps is investigated using filtered networks. A user-friendly implementation of the method is provided as a plug-in for the program SplitsTree4.

1 Introduction

In systematics, the evolution of different species is of interest, however, phylogenetic inference is often based on the DNA or protein sequence of homologous genes and the resulting *gene trees* are usually interpreted as estimations of an underlying *species tree*. A common observation is that different genes give rise to different trees, even in the absence of tree-reconstruction errors, and this fact can usually be explained by mechanisms such as incomplete lineage sorting, duplication-and-loss, horizontal gene transfer (e.g. in bacteria) or hybridization (e.g. in plants).

Although phylogenies based on single gene analysis [32] continue to play a central role in phylogenetics, biologists interested in the evolution of specific groups of taxa often sequence and use more than one gene to infer the phylogeny

of the taxa [23], the hope being that as more data is brought into the analysis, a better “species-signal” to “gene-noise” ratio will be obtained and that deviating signals from individual genes can be filtered out.

If the goal is simply to obtain a good estimation of the species tree and if there is evidence that a majority of the genes under study have evolved in a similar way along the same species tree, then one approach is to concatenate the alignments given for each of the genes to produce one large dataset, to which tree-building methods are then applied [23,25]. If each of the genes is long enough to contain strong phylogenetic signals for the group of taxa under investigation, then a second approach is to compute individual gene trees, to summarize them using a (usually somewhat unresolved) consensus tree and then to interpret the consensus as a representation of the well-supported parts of the species tree [30,10,26].

In both cases, the final result suppresses all incompatible signals. However, if the actual incongruencies of the individual gene trees are themselves of interest, then a representation of the data set that maintains (some of) the incompatible signals may be useful. Such a representation is given by the concept of a “split network” [1] and methods for computing such networks are presented in [8] and are implemented in the program SplitsTree4 [15].

To obtain an explicit model of reticulate evolution, reticulate networks are used [15] that explain a given set of trees in terms of hybridization, horizontal gene transfer or recombination events [13,7,19,17,18]. Current methods for determining a hybridization scenario that explains a given set of trees operate by performing a combinatorial analysis of the total set of splits of the trees to identify a hybridization network that generates the trees [22,17]. By definition, combinatorial methods are very sensitive to false positive splits, that is, splits that are incompatible to other splits in the input due to reasons such as homoplasy, tree-estimation error, incomplete lineage sorting etc.

Given a collection (or *profile*) P of k gene trees all inferred on the same set of taxa X , one approach to constructing a set of splits that summarize the set of trees, without eliminating all incompatibilities, is given by the consensus network method [2,14]. This method consists of returning all splits that occur in at least αk of the given input trees, for a given threshold $\alpha \in [0, 1]$.

A main drawback of the consensus network approach is that in practice typical data sets often consist of *partial trees*, that is, gene trees that each only mention some subset X' of the total taxon set X . Partial trees arise because the sequence data for some gene has not yet been sequenced, or because the gene is not present in the genome, for some taxon.

Given a profile of partial gene trees, the Z-closure method [16] computes a *super network* on the full taxon X that summarizes all the input trees. This approach first uses an inference rule to construct a set of splits on the full taxon set and then, as above, a network construction algorithm [8] is employed to obtain a split network. A practical weakness of this method is that it does not provide a natural parameter (such as α above) with which one can control the amount of incompatibility that is represented in the resulting network.