Constructing the simplest possible phylogenetic network from triplets

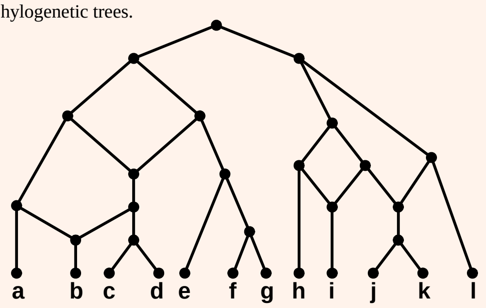
(Two efficient algorithms for combining multiple smaller evolutionary hypotheses into one overall hypothesis)

Leo van Iersel[2] and Steven Kelk[1]

[1]Centrum voor Wiskunde en Informatica (CWI) and [2]Technische Universiteit Eindhoven (TU/e)

1) Phylogenetic networks generalise phylogenetic (evolutionary) trees

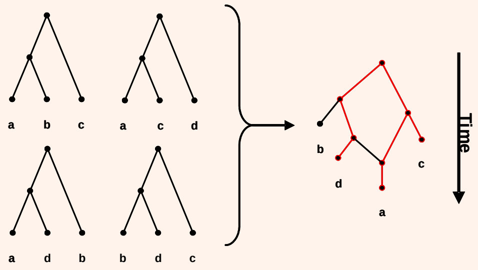
A phylogenetic network is a directed acyclic graph that visualises a rooted evolutionary history containing so-called reticulations such as recombinations, hybridisations or lateral gene transfers. Phylogenetic networks are thus **powerful generalisations** of phylogenetic trees.



An example of a (level-2) phylogenetic network. Arcs are assumed to be directed downwards, away from the root.

Here we consider the construction of a **simplest possible** phylogenetic network consistent with an input set T, where T contains at least one phylogenetic tree on three leaves (a triplet) for each combination of three taxa. Each triplet represents a hypothesis about the evolutionary history of those three taxa. In essence, we thus wish to find a single phylogenetic network containing all the taxa and **which does not contradict any of the hypotheses represented by the input triplets**.

To quantify the complexity of a network we consider both the total number of reticulations (i.e. vertices with indegree 2) and the number of reticulations per biconnected component, called the **level** of the network. The lower the level of a network, the more 'tree-like' it is, with level-0 networks being phylogenetic trees.

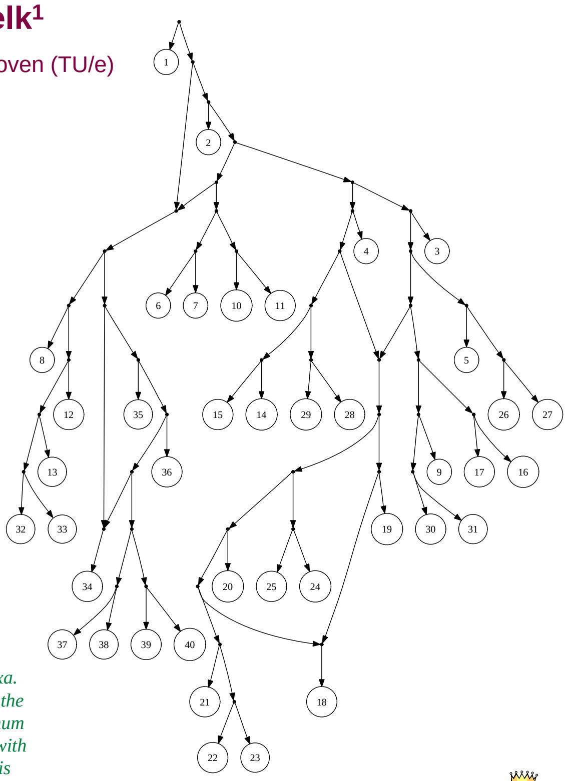


*In this example we are given as input the triplets ab|c, ac|d, ad|b and bd|c. We say a network is consistent with a triplet if there is a subdivision of the triplet embedded in the network. There is no rooted binary tree with leaf set {a,b,c,d} that is consistent with all four triplets. But, as shown here, there does exist a rooted **network** consistent with all four of them. By way of illustration an embedding of ac|d within the network is shown in red. The network is level-1.*

3) From theory to practice: MARLON (Minimum Amount of Reticulation Level One Network)

We have implemented the level-1 algorithm in Java and, using the Maximum Likelihood (ML) package PHYML to generate the input triplets, have tested it on simulated data with promising results. The resulting package **MARLON** has been made publicly available alongside the accompanying test data.

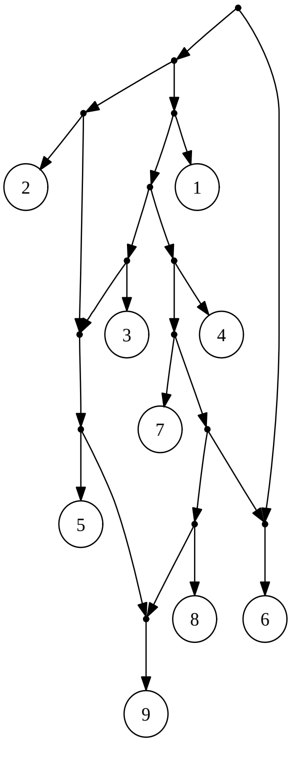
The figure on the right shows the output of MARLON on simulated triplet data for 40 taxa. Amongst all level-1 networks consistent with the triplet set, this is guaranteed to have a minimum number of reticulation vertices (i.e. vertices with indegree 2). Simulations in which MARLON is compared to the package T-REX suggest MARLON is competitive in terms of its capacity to recreate the original network.



4) From theory to practice: SIMPLISTIC (SIMPLe network heurISTIC)

Our theoretical results also form the basis for a second package, **SIMPLISTIC**, that we have also made publicly available. The chief advantage of SIMPLISTIC is that it **always** returns some phylogenetic network consistent with all the input triplets as a solution (whilst striving to minimise the level of that network) even if there are many errors in the input triplets and/or the underlying evolution is of a high level.

Tests of SIMPLISTIC on yeast data of Fraser et al. (Same-sex mating and the origin of the Vancouver Island Cryptococcus gattii outbreak, Fraser et al, Nature 437:7063, 1360-1364 (2005)) have produced interesting insights into the possible location of reticulate evolutionary events within that group of yeasts.



The figure on the left shows a simple example of the type of networks that SIMPLISTIC can produce. (This is a level-3 network: there are three reticulation vertices nested together.) Existing algorithms such as MARLON (and the related algorithm LEVEL2 [1]) perform well if the input data is consistent with level-1 or level-2 networks. SIMPLISTIC complements this approach by always returning a network even if the input data is more complex than level-2.

2) Theoretical results

We give polynomial-time (i.e. efficient) algorithms for constructing a level-1 respectively a level-2 network that contains a minimum number of reticulations and is consistent with T (if such a network exists) [2]. These algorithms extend the results given in [3] and [1]. In addition, we show that if T is precisely equal to the set of triplets consistent with some network, then we can construct such a network with smallest possible level in polynomial time, if k is a fixed upper bound on the level of the network [2].

5) Next steps

MARLON and SIMPLISTIC are freely available on-line [4]. A basic, but important step is to add a graphical user interface to SIMPLISTIC. A crucial extension to SIMPLISTIC which will further extend its relevance to real-world data is to relax the constraint that it only returns networks consistent with 100% of the input triplets. This is important because, although SIMPLISTIC is theoretically efficient, it can take a long time to terminate if the input data is heavily 'knotted' i.e. is only consistent with networks of very high level. This is work in progress!

Part of this research has been funded by the Dutch BSIK/BRICKS project AFM2.

For further information please contact S.M.Kelk@cwi.nl

Literature

[1] Leo. van Iersel , Judith Keijsper, Steven Kelk, Leen Stougie, Ferry Hagen and Teun Boekhout, Constructing level-2 phylogenetic networks from triplets, *Proceedings of RECOMB2008 (Research in Computational Molecular Biology 2008)*, LNCS 4955, 450-462 (2008)

[2] Leo van Iersel and Steven Kelk, Constructing the simplest possible phylogenetic network from triplets, arXiv:0805.1859v2 [q-bio.PE] (2008)

[3] Jesper Jansson, Nguyen Bao Nguyen and Wing-Kin Sung, Algorithms for combining rooted triplets into a galled phylogenetic network, *SIAM Journal on Computing* 35(5), 1098–1121 (2006)

[4] http://homepages.cwi.nl/~kelk