Summary of the PhD Thesis
Combinatorial Methods for Phylogenetic Network Reconstruction

Philippe Gambette

Preamble

We first introduce the context of the manuscript by recalling the origin of phylogenetic trees as ways or describing evolution of species or classifying them. We mention some applications of phylogeny and explain how a phylogenetic tree is a simplified view of a tokogeny, or ancestral recombination graph, which describes genetic relationships between individuals. Hence, a more accurate but more complex model to describe this tokogeny, given hybridation and other biological processes of exchange of genetic material between species, is the phylogenetic network. As well as phylogenetic networks generalize trees to model evolution (then they are called "explicit"), other kinds of phylogenetic networks, known as "abstract" or "data-display networks" also generalize trees as a way to classify species and visualize relationships between them.

Then we explain the choice of this thesis to focus on combinatorial methods for phylogenetic network reconstruction (instead of geometric or statistical methods): the abundance of data, and in particular the development of tree databases is a motivation to reconstruct phylogenetic networks from trees instead of processing huge quantities of sequence data. In this context, interesting problematics are:

- finding links between existing methods and properties of the mathematical objects which have been introduced in existing combinatorial methods.
- developing new combinatorial methods, both with a theoretical interest to better understand the properties of the objects we are studying, or with a practical interest to provide fast algorithms.
- discussing the relevance and reliability of these methods by studying their limits and conditions of use, and confronting them to real data.

Then the outline of the thesis, which we detail below, is given, as well as a list of publications which were prepared or published from this work:

- Published articles:
  - [**ISMB2009**] Daniel Huson, Regula Rupp, Vincent Berry, Philippe Gambette & Christophe Paul: Computing Galled Networks from Real Data, *Bioinformatics* 25(12), *Proceedings of the seventeenth Annual Conference on Intelligent Systems for Molecular Biology & eighth European Conference on Computational Biology* (ISMB'09), p. i85-i93, 2009.
  - [**CPM2009**] Philippe Gambette, Vincent Berry & Christophe Paul: The Structure of Level-*k* Phylogenetic Networks, *Proceedings of the twentieth Annual Symposium on Combinatorial Pattern Matching* (CPM'09), *LNCS* 5577, p. 289-300, 2009.
  - [**IFCS2009**] Philippe Gambette & Jean Véronis: Visualising a Text with a Tree Cloud, *Proceedings of the International Federation of Classification Societies 2009 Conference* (IFCS'09), *Studies in Classification, Data Analysis, and Knowledge Organization* 40, p. 561-570, 2010.
- Submitted or in preparation:
  - [**Encodings2010**] Philippe Gambette & Katharina T. Huber : A Note on Encodings of Phylogenetic Networks of Bounded Level, *Journal of Mathematical Biology*, submitted, 2010.

- [**Quartets2010**] Philippe Gambette, Vincent Berry & Christophe Paul: Quartets and Unrooted Phylogenetic Networks, *in preparation*, 2010.

Part I) A combinatorial approach of phylogenetic networks

Chapter 1) Trees and networks as combinatorial objects

In sections 1.1 to 1.3, we first give basic definitions about graphs, directed graphs, phylogenetic trees (rooted or unrooted) and networks (abstract or explicit). We also define decompositions of trees as subsets of leaves: triplets, quartets, clusters (hardwired or softwired) and splits.

We then discuss the interpretation of multifurcations and multireticulations (vertices of degree >3) in the network, in section 1.3.3. We show why they are difficult to handle in the context of reconstruction from triplets, which explains why in the context of triplets or quartets we will focus on binary phylogenetic networks.

In section 1.4 we give some definitions of restricted classes of phylogenetic networks:

- abstract rooted phylogenetic networks, or in fact the restricted cluster system which is associated to each of them: weak hiearchies, prepyramids, *k*-weak hierarchy.
- abstract unrooted phylogenetic networks (median networks and split networks), or in fact the restricted split system which is associated to each of them: weakly compatible, circular, *k*-compatible.
- rooted explicit phylogenetic networks studied in this thesis: galled networks, and level-*k* networks. For the latter we give structure properties: they can be seen as a tree of blobs (bridgeless components) which can be summed up as simple graph motifs called "generators" of level at most *k* (results published in sections 2.1, 2.2 and 3 of attached file **CPM2009**).
- unrooted explicit phylogenetic networks: unrooted level-*k* networks. We also explain and prove the link with rooted networks (see section 3 of attached file **Quartets2010**)
- other rooted explicit phylogenetic networks: regular, tree-child, tree-sibling, normal.

We conclude this chapter with links between restricted classes of phylogenetic networks, in section 1.5:

- subsets of hardwired and softwired clusters of level-1 networks are prepyramids (a result a bit stronger than proposition 1 of attached file **Encodings2010**)
- subsets of splits of unrooted level-1 networks are exactly circular split systems (see section 4 of attached file **Quartets2010**)
- diagrams which provide a summary of the inclusion relationships among restricted classes of phylogenetic networks, as well as references to the publications where these inclusions are proved.

Chapter 2) Combinatorial reconstruction algorithms

In section 2.1.1, we first give an overview of existing results on approaches to reconstruct phylogenetic networks from trees, especially about the hybridization number, and about consensus networks. Then we explain the motivations to reconstruct a network consistent with the triplets, quartets, clusters or splits of the input trees, instead of the trees themselves (computational complexity, reliability of the input trees), and give an overview of existing methods to reconstruct networks from such data. We give a more detailed description of

existing triplet methods in section 2.1.2, as we generalize some of these results to quartets later.

In section 2.2 we present results on unrooted level-k network reconstruction from quartets (we follow sections 5 to 7 of attached file **Quartets2010**).

In section 2.3 we present a new practical method to reconstruct galled network from softwired clusters, which come from a set of gene trees for example (we follow sections 3 to 5 of attached file **ISMB2009**).

Part II) Practical use of combinatorial methods

Chapter 3) Limits of combinatorial methods

In section 3.1 we consider the problem of noise and silence in the data, as combinatorial methods usually work on exact and complete data. After citing existing approaches to correct the noise in triplet data, we give an algorithm in time $O(6^t n+n^4)$ to edit the minimum number t of triplets in a dense triplet set on n leaves, to make it consistent with a tree. Then we cite existing approaches to handle silence in the data: supertrees to infer missing triplets, Z-closure, Q-imputation, M-closure and Y-closure to infer splits if there are missing taxa in the input trees.

In section 3.2 we describe the explosion of complexity when the level of a rooted phylogenetic network increases. This explosion is expressed by the number of generators which is exponential in the level (see sections 2.3 and 2.4 of attached file **CPM2009**). Also, we observe that the level of networks simulated by the coalescent model with recombination is high even with small recombination rates.

In section 3.3 we describe the problem of ambiguous phylogenetic networks given complete and correct data by sudying which networks are encoded by their triplet set (or softwired cluster set, or their set of contained trees). We characterize level-1 networks which are encoded by their set of triplets, clusters, or contained trees, i.e. such that no other level-1 network has exactly the same set of triplets, clusters, or contained trees: level-1 networks whose non-trivial blobs have strictly more than four vertices (see sections 3 to 5 of attached file **Encodings2010**). To conclude this chapter we also give examples of level-2 networks which are not encoded by their triplet or cluster systems. Hence solutions of phylogenetic network reconstruction methods from triplets or clusters should not be trusted blindly, as others solutions, with the same level, number of edges and reticulations, may also exist.

Chapter 4) Combinatorial methods on real data

In section 4.1 we discuss data selection and preprocessing. We first explain in section 4.1.1 that real data may in fact be more complex than the trees we have considered so far (for example gene trees with multiple taxa arising due to duplication or a lot of missing taxa due to deletions). It may also contain more information (gene trees with confidence scores on the edges or with estimated dates on the vertices). We cite existing methods which deal with these biological constraints on the data.

In section 4.1.2, we present a tool to choose among the many possible methods depending on the data, an interactive online bibliography, called "Who is who in Phylogenetic Networks",

available at *http://phylnet.info*.

In section 4.1.3 we present a data selection problem which arises when we try to use the methods presented in chapter 2 on real data: as these methods require input trees on the same taxa set, or dense set of triplets, we explain how to express the problem of finding a big set of trees with a big common set of taxa as finding a maximal edge biclique in a bipartite graph. We design the concept of an interface, HeurisTree, currently being implemented, which helps the user finding such a set of trees and taxa. The basis of this program is a new visualization tool called the tree cloud (see the attached file **IFCS2009** and *www.treecloud.org*). It displays the names of gene trees around a tree which reflects how similar their taxon sets are, and where the font size reflects how many taxa the tree contains. Hence, focusing on gene trees written in big fonts, in the same area of the tree cloud, helps finding an appropriate set of gene trees and taxa.

Finally, in section 4.2, we illustrate combinatorial methods for phylogenetic network reconstruction on data extracted from the Hogenom database (more than 900 genomes, i.e. taxa, and 200 000 gene families, i.e. input trees).

Conclusion and perspectives

After giving some open problems, we conclude on the interest of combinatorial methods for phylogenetic network reconstruction, which should not be considered as giving a reliable result, but more as exploratory tools for giving candidate solutions. Mixing these approaches with a statistical evaluation of the results seems to be the key for a fast reconstruction of reliable phylogenetic networks.