From phylogenetic trees to networks

Encadrant : Fabio PARDI (LIRMM, pardi@lirmm.fr)

Co-encadrants : Vincent BERRY (LIRMM, vberry@lirmm.fr), Céline SCORNAVACCA (Institut des Sciences de l'Evolution de Montpellier, celine.scornavacca@umontpellier.fr)

Location: LIRMM, Montpellier

Overview

In biology, phylogenetic trees are used to represent the evolutionary history of genes, populations, species. They are typically reconstructed with a wide range of algorithms from the comparison of very long strings representing the molecular (DNA or protein) sequences found across different organisms. Phylogenetic reconstruction is part of the daily routine of many biologists, from epidemiologists to molecular geneticists. Currently, a novel and exciting subject of research in this field is emerging, where evolution is represented not by trees, but by special types of DAGs (directed acyclic graphs) called *phylogenetic networks* [1]. This is relevant to display *reticulate events*, where an organism inherits its genetic material from more than one ancestor, and which are common for example in viruses, bacteria and plants (see Fig. 1).

While phylogenetic tree reconstruction is a relatively well-studied research subject, which has led to the development of programs that are currently used by many thousands of users across the world, the same cannot be said for the reconstruction of phylogenetic networks. Much research, both basic and applied, is still needed for this field to reach maturity. As part of this stage, the student will be introduced to this field, and to a number of open problems relevant to the analysis of biological data made available by recent projects such as the 3,000 rice genomes project [2]. Aside from the necessary bibliographical study, the research work will be adapted to the student's inclination and taste (see more below).

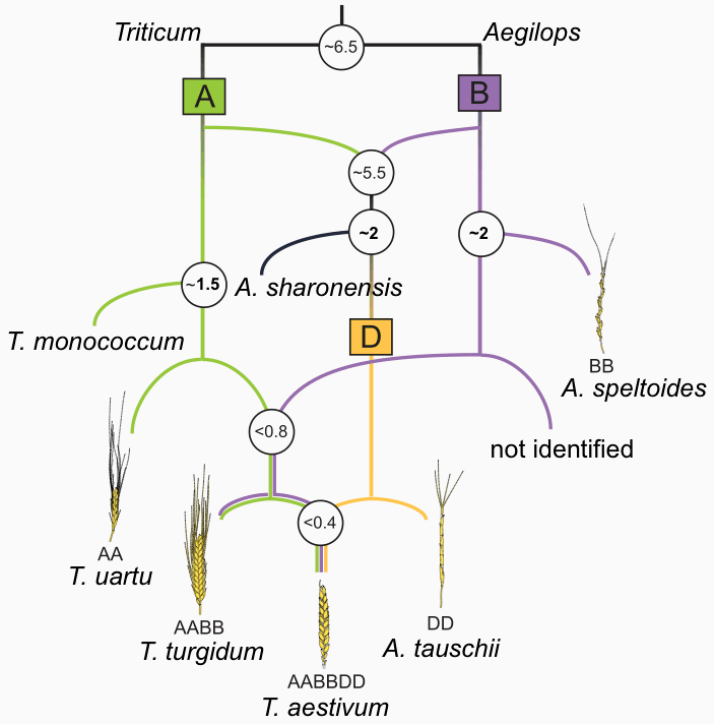

Figure 1. Phylogenetic network representing the reconstructed evolution of bread wheats (Marcussen et al. Science 345, 2014). The circles indicate putative dates (in million years ago) for speciation and reticulate events.

Research work

In the last decade, phylogenetic networks have been intensively studied from a mathematical and computational perspective. The bibliographic part of the stage will focus on a number of results published since the appearance of the textbook [1]. This will offer the opportunity for the student to identify problems that interest him/her and that can exploit his/her strengths and inclinations. The research work can range from mathematical – working on conjectures on the fundamental limitations of network reconstruction – to algorithmic – studying computational problems raised by modern data and devising efficient algorithms to solve them – to entirely applied – involving the implementation of programs and the analysis of real data.

From a theoretical perspective, a very interesting aspect of phylogenetic networks is that whereas in theory it is always possible to reconstruct a phylogenetic tree, given sufficient data for this task, the same does not hold for phylogenetic networks: most notably, the relative order of consecutive reticulate events is very difficult, if not impossible, to recover from the data [3]. The extent of this fundamental limitation for network reconstruction has yet to be fully understood and characterized. Another interesting research subject is the adaptation of distance-based methods to network reconstruction. Distance-based methods are a standard family of algorithms whose goal is to reconstruct a weighted tree from estimates of the distances

between its leaves, i.e. the lengths of the paths between them. While the matrix D^T containing all the path-length distances between the leaves of a positively weighted tree T can originate from a unique positively weighted tree T [4], and therefore determines T — and many algorithms can be used to recover T from D^T — asking similar questions for phylogenetic networks instead of trees leads to a number of open questions.

Moreover, like everywhere else in bioinformatics, the availability of ever faster and cheaper sequencing techniques is causing the need of computationally efficient techniques to analyze the new data. Of particular relevance for this project are phylogenetic methods adapted to genome-size plants data, which have probably undergone reticulate events and therefore require a network representation (see Fig. 1). This is a two-fold challenge: First, we need to speed up the computationally-hard task of reconstructing phylogenetic networks to be applicable to full-size genotypic data for plants (e.g., up to 20 million SNPs across 3000 individuals are available in the case of rice *Oryza sativa* [2]). Second, a new model has to be designed, as several founders are known (e.g., three to four for *Citrus*, and six to seven for *Oryza* and *Banana*) whereas standard phylogenetic networks only have one root. The small number of founders may lead to consider fixed-parameter tractable algorithms exploiting this and other constraints to run in reasonable time.

Bibliography

[1] DH Huson, R Rupp, C Scornavacca. *Phylogenetic Networks.* Cambridge University Press (2011)

[2] JY Li, J Wang, RS Zeigler. The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience 3(2014), no. 1, pp. 1-3 (2014)

[3] F Pardi, C Scornavacca. Reconstructible phylogenetic networks: Do not distinguish the indistinguishable PLoS Computational Biology 11(4): e1004135 (2015)

[4] Peter Buneman. The recovery of trees from measures of dissimilarity. In D.G. Kendall and P. Tautu, editors, *Mathematics the the Archeological and Historical Sciences*, pp. 387–395. Edinburgh University Press (1971)