6-months internship proposal (first half of 2020)

Computing the scanwidth of a DAG efficiently

Main laboratory: ISEM, Montpellier, France. Céline Scornavacca (Molecular Phylogeny and Evolution team) is a specialist in phylogenetic networks, combinatorial algorithms, modelling in phylogenomics.

Partner laboratories: LIGM, Paris, France. Mathias Weller (Algorithmics for Bioinformatics team) is a specialist in parameterized algorithms, structural parameterization, preprocessing and graph theory.

Skills required: strong background in algorithms, analytical skills, C++/Java programming and an interest in evolutionary models will be a plus.

How to apply: Cover letter and CV (with academic transcript) to be sent to celine.scornavacca@umontpellier.fr and mathias.weller@u-pem.fr.

Background: Phylogenetic networks are rooted and leaf-labelled directed acyclic graphs (DAGs) used to depict the evolution of a set of species in the presence of reticulate events such as hybridizations, where two species combine their genetic material to create a new species. Reconstructing these networks from molecular data is challenging and current algorithms fail to scale up to genome-wide data. Aiming at designing faster parameterized algorithms for this task, we recently stumbled [2] on a new width parameter for DAGs, which we call "scanwidth". To get an intuition, imagine a scanner line traversing a network from the leaves to the root; at any moment, its width is the number of arcs it cuts. As the line moves up, it traverses nodes, changing the set of arcs it cuts and, hence its width. The *cutwidth* of the network is the largest width achieved by such a traversing line. Now, consider multiple independent scanner lines, each one scanning an arc incoming to a different leaf of the network. Whenever a node could be passed by two different lines, they are merged to form a single one. This naturally generalizes the cutwidth to a smaller width measure that we call *scanwidth*. As with the cutwidth, different orders in which the nodes are passed imply different values of the final width and the goal is to minimize it. The scanwidth broadens the arsenal of width measures that can be used to attack hard problems in phylogenetics and permitted us to design a faster parameterized algorithm for network reconstruction, which allowed us to handle several real-world datasets within minutes instead of weeks [2]. Still, since deciding the scanwidth is NP-complete even for very restricted classes of networks [1], in our implementation we actually used a simple heuristics to compute it, leaving a large potential for improvement.

Task: The intern will study how the scanwidth relates to other width measures and graph-theoretic problems as well as design algorithms to compute and approximate this parameter efficiently. It is also desirable to develop an implementation of the resulting algorithms.

Practical Information: The internship is a full-time position for 4-6 months, during which the intern is expected to be present in Montpellier, France. Monthly salary is about 600€.

Note: If the internship will come to fruition, the intern will have the possibility to continue her/his work by way of a PhD scholarship (funding already acquired).

References

[1] Berry V, Scornavacca C, Weller M, Proceedings of the 46th International Conference on Current Trends in Theory and Practice of Computer Science, 2020.

[2] Rabier C-E , Berry V, F. Pardi F, and Scornavacca C. On the inference of complicated phylogenetic networks by Markov chain monte-carlo. submitted