On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo

Charles-Elie Rabier[1,2], Vincent Berry[3], Jean-Christophe Glaszmann[4], Fabio Pardi[3] and Céline Scornavacca[1]

[1] ISE-M, Univ.Montpellier, CNRS, EPHE, IRD, Montpellier, France
[2] IMAG, Univ.Montpellier, CNRS, Montpellier, France
[3] LIRMM, Univ.Montpellier, CNRS, Montpellier, France
[4] AGAP, CIRAD, Montpellier, France

Corresponding author: charles-elie.rabier@umontpellier.fr

Complete genomes for numerous species in various life domains (Denoeud et al. 2014, Badouin et al. 2017, Garsmeur et al. 2018), and even for several individuals for some species (Hapmap Consortium 2003, 3000 Rice Genome Project 2014) are nowadays available thanks to next generation sequencing. To process such a large amount of data, methods need not only to be accurate, but also time efficient. We present here an efficient method dedicated to phylogenetic network inference.

In phylogenetics, species tree inference has been studied extensively for many years, and the theory behind it is relatively well known. However, a species tree is unable to model complex biological events such as horizontal gene transfer (e.g. procaryotes, Koonin et al. 2001, but also among eucaryotes, Szollhosi et al. 2015), hybridization (plants and animals, Mallet 2007), introgression (e.g. citrus, Minamikawa et al. 2017) and recombination. In contrast, phylogenetic networks, that differ from species trees because of reticulate edges, are able to capture all those phenomenons.

We present here a novel way to compute the likelihood of biallelic markers given a phylogenetic network. This computation is at the heart of a Bayesian network inference method – called SNAPPNET, as it extends the SNAPP method (Bryant et al., 2012). SNAPPNET is available as a package of the well-known Beast 2 software (Bouckaert et al., 2014 and 2019). This package partly relies on code from SNAPP method (Bryant et al., 2012) to handle sequence evolution and on code from SPECIESNETWORK (Zhang et al., 2018) to modify the network during the MCMC as well as to compute network priors.

Our approach differs from that of Zhang et al. (2018) in that SNAPPNET takes a matrix of biallelic markers as input while SPECIESNETWORK expects a set of nucleotide alignments for which it samples possible gene trees as part of its process. Thus, the considered substitution models differ but more importantly, our method does not need to consider gene tree inference as an intermediary step. Following SNAPP, SNAPPNET's computations integrate over all possible tree histories for a locus, while SPECIESNETWORK considers only a sample of locus trees from the infinite number of possible topologies and branch lengths.

SNAPPNET is much closer to the MCMCBiMarkers method of Zhu et al. (2018), which also extends the SNAPP method (Bryant et al., 2012) to network inference. Both methods take biallelic markers as input, rely on the same model of evolution and also both sample networks in a Bayesian framework. However, SNAPPNET is exponentially more efficient in computing likelihoods for non-trivial networks. Also, the methods differ in the way the Bayesian inference is conducted.

In this poster, we will describe SNAPPNET and compare its performances with MCMCBiMarkers on simulated data. We will also give an illustration on rice data.

Acknowledgements

This work was supported by the Key Initiative Muse Data Science (I-SITE MUSE: ANR-16-IDEX-0006) and by the project Genome Harvest ref. ID1504-006 ("Investissements d'avenir", ANR-10-LABX-0001-01).

References

D. Bryant et al. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular biology and evolution*, 29(8), 1917-1932.

C. Zhang et al. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular biology and evolution*, 35(2), 504-517.

J. Zhu et al. (2018). Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLoS computational biology*, 14(1), e1005932.