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

Charles-Elie Rabier

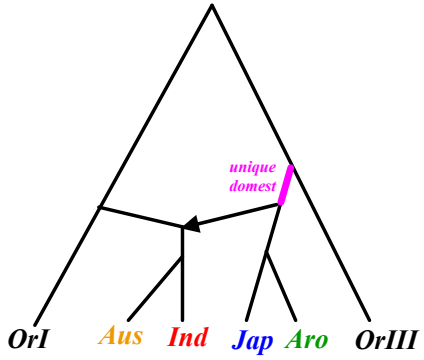
Vincent Berry, Jean-Christophe Glaszmann
Fabio Pardi et Céline Scornavacca

Genome Harvest / KIM Data & Life Sciences
ISEM, Institut des Sciences de l'Evolution de Montpellier
IMAG, Institut Montpelliérain Alexander Grothendieck
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique
UMR AGAP, Amélioration Génétique et adaptation des plantes, CIRAD

Phylogenetic networks

Phylogenetic networks are Directed Acyclic Graphs (DAG) that allow to detect reticulation events, such as :

- Admixture (e.g. humans)
- Introgressions (e.g. plants and animals)
- Horizontal gene transfers (e.g. bacteria)



Some key points on networks :

- Edge length = evolutionary time
- Reticulation nodes have 2 parents and represent reticulation events
- We want to obtain a probability distribution of networks (uncertainty on clades)
- The more data we have, the more precisely we can infer the network

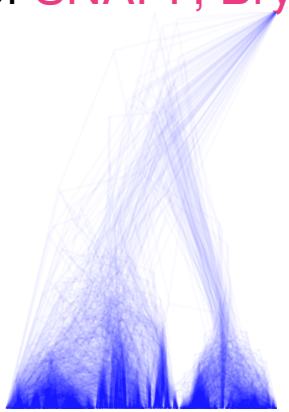
Data in our study :

- A Data matrix containing an alignment of m biallelic markers
- Markers can be SNPs or random sites including invariant sites
- A certain distance must be preserved between genomic sites

Our new Bayesian method, SNAPPNet

- N : phylogenetic network (topology, branch lengths, population sizes)
- X_i : data for marker i
- G_i : locus tree for marker i
- m markers, $Data = (X_1, \ldots, X_m)$
- $\mathbb{P}(Data \mid N)$: likelihood of the data given the network
- P(N) : network prior
  birth hybridization process of Zhang et al. MBE 2018
- Posterior probability distribution (extension of SNAPP, Bryant et al. MBE 2012)

$$\mathbb{P}(N|Data) \propto \left(\prod_{i=1}^{m} \int_{\psi} \mathbb{P}(X_i|G_i)\mathbb{P}(G_i|S)dG_i \right) P(N)$$
$$\propto \mathbb{P}(Data \mid N)\, P(N)$$



$\Rightarrow$ Markov Chain Monte Carlo (MCMC) in order to sample the posterior probability distribution $\mathbb{P}(N|X_1, \ldots, X_m)$ using new algorithms dedicated to networks

SNAPPNet implemented in BEAST

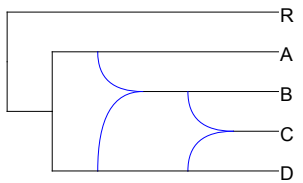
How SNAPPNet computes the likelihood $\mathbb{P}(Data \mid N)$

Contrary to trees, network edges can not be treated independently

- We compute joint probability distributions
- We minimize computing time by minimizing the conditional part in computation

Dataset ID	CPU time	
	SNAPPNet (in minutes)	MCMCBiMarker (in hours)
1	5.559	35.9354
2	5.6763	34.2433
3	5.7351	32.6519
4	5.446	34.2011
5	5.5996	33.2354

SNAPPNet (Us) vs MCMCBiMarker
(Zhu et al., Plos Comp Biol 2018)

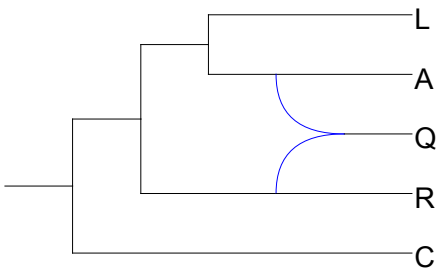


a level-2 network

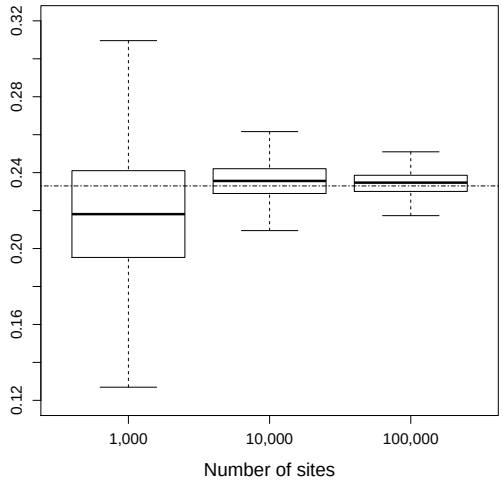
Illustration of SNAPPNet on simulated and real data

- Simulation study

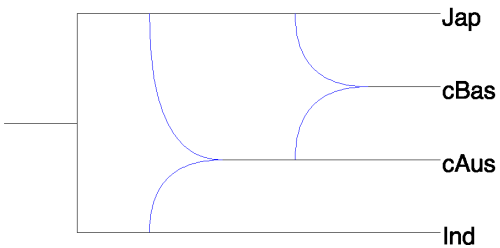
True network



Estimated network length



- Study on rice genome (4 subpopulations ***Ind***ica, ***Jap***onica, circum***Aus*** and circum**Bas**mati using data from *Wang et al., Nature 2018*)



Rice evolution scenario inferred by SNAPPNet

c*Aus* : result of an early(old) combination between the *Ind* and the *Jap* lineages

c*Bas* : a later combination between the c*Aus* and the *Jap* lineages

New scheme that appears compatible within the simpler schemes that have been proposed so far (e.g. *Choi et al, Genome Biology 2020*)