Practical session – Phylogenetic Network analysis using SplitsTree, Dendroscope and PhyloNet

Celine Scornavacca

January 25th, 2019

1 Phylogenetic networks from distances

Question 1 *Using* ***SplitsTree****, compute the BioNJ tree, the UPGMA tree and the split decomposition network for the following matrix:*

a	0	4	5	4
b	4	0	3	4
c	5	3	0	3
d	4	4	3	0

Refer to the tab "Help" → "Network Syntax" → "Distances" to learn how to specify a distance matrix in Nexus format, or use the file ***north.nex*** *in the example folder of* ***SplitsTree****. How these trees and networks are related? Explore the leftmost tab to see the computed set of splits for each method.*

Question 2 *Using* ***SplitsTree****, compute the split decomposition network for the dataset of primate lentiviruses in the file* ***PLV-split-network.nex****. What does the network tell us? Which are the splits in the split decomposition and what are their weights?*

Question 3 *Using* ***SplitsTree****, compute the neighbor-net network for the a data set of 44 Vietnamese wild and domestic chicken populations (see file* ***chickens.nexus****). Additionally, reconstruct the split decomposition network for the same data. In what way do the two networks differ? Assumed that domestic populations are labeled by HGXX, what the neighbor-net network suggests ?*

Question 4 *Are the above-cited methods the only distance-based methods available in* ***SplitsTree****? If not, what are the others? Play around with them, using the data sets of the previous questions.*

2 Phylogenetic networks from sequences

Question 5 *Using* ***SplitsTree****, compute the median network for the following binary matrix.*

Refer to the tab "Help" → "Network Syntax" → "Characters" to learn how to specify a sequence matrix in Nexus format, or use the file ***dolphins_binary.nex*** *in the example folder of* ***SplitsTree****.*

a	0000000
b	0110000
c	1101100
d	1110110
e	0110101

Question 6 *Using* `SplitsTree`*, compute a pruned quasi median network for the matrix contained in the file* `lugens-1.txt`*. (why a quasi median one?) choosing the option* `Label Edges` *and the option* `Show Haplotypes`*. Can you interpret the network? How many vertices are in this network? For how many sequences? Can you guess the main problem when reconstructing (quasi) median networks?*

Question 7 *While studying the phylogeographic structure of lineages of the fungus Fusarium graminearum, scientists discovered that the nuclear 3-O-acetyltransferase gene (TRI101) has undergone intragenic recombination in one of the strains. Reconstruct the recombination network from the data set contained in the file* `recombNet.txt` *using* `SplitsTree`*. Can you validate this theory? The taxon O13393 is the outgroup. (The Recombination Network method is well hidden in* `SplitsTree`*... use Rectangular Phylogram as Layout method).*

Question 8 *Are the sequence-based methods we discussed applicable in phylogenomic studies? Are the above-cited methods the only sequence-based methods available in* `SplitsTree`*? If not, which are the others? Can you translate sequence matrices into distance matrices?*

3 Phylogenetic networks from trees

Question 9 *For the set of phylogenetic trees on eight yeast species contained in* `rokas.nex`*, compute three split networks that represent all splits that occur in at least one input tree, in more than* 5% *of all trees, and in more than* 30% *of all trees, respectively. (Use the option edge weights to obtain readable networks). Additionally, compute the majority consensus tree for these trees. Please describe the networks and the relationships among them.*

Question 10 *Compute the supernetwork for the set* $\mathcal{T}$ *of 5 phylogenetic trees contained in* `kim1.nex`*. Can you interpret this network? What is the Z-closure? Why do we need to use it for* this *data set but not for the trees of Question 9? Does the network change if we modify the number of runs or apply the refined heuristic?*

Question 11 *For the data set of Question 10, compute the filtered Z-closure super network based only on splits found in strictly more than two of the gene trees. How does the network change?*

Question 12 *Now open the software* `Dendroscope`*. Reroot the trees of Question 9 at the taxon S. cerevisiae (Hint: draw all the trees in the same window, select them all, search for the taxon name and reroot). Compute rooted consensus networks with the same thresholds used in Question 9 (use the Cluster Network Consensus method). Can you spot the same incongruences? Do the same for the trees of Question 10, rooted at A._thaliana.*

Question 13 *Still in* `Dendroscope` *and with the trees of Question 9 rooted at S. cerevisiae, run both the Cluster Network Consensus and the Galled Network Consensus methods with*

threshold 5%. What do you notice? Can you say why a network is more complicated than the other?

Question 14 *The file* `Triticeae.txt` *contains 225 trees describing the evolutionary relationship among the Triticeae, a tribe of grasses. (Try to) construct a consensus with threshold= 0 with the Cluster Network Consensus or the Galled Network Consensus methods. What do you notice? Play with the threshold to get a reasonable network. Do the same for the file* `Triticeae_collapsedAt80.txt`*, containing the same trees as the previous file with all branches of less than 80% support have been collapsed. What do you notice?*

Question 15 *For the two trees in the file* `phyBwaxy-trees.txt`*, compute their hybridisation network(s). Why do you get several networks and how are they related? Hint: use the tanglegram algorithm to compare them.*

4 Phylogenetic networks + ILS

4.1 PhyloNet from trees

A general overview of `PhyloNet` can be found at `https://wiki.rice.edu/confluence/display/PHYLONET/PhyloNet+3+General+Overview` and a list of all commands can be found at `https://wiki.rice.edu/confluence/display/PHYLONET/List+of+PhyloNet+Commands` (we will mainly focus on the first section of the list).

Question 16 *In* `commandPhyloNET_Triticeae.txt` *you will find an example input for* `PhyloNet` *to perform some analyses on the data set of Question 14 such as computing the likelihood of a network given the branch lengths of the gene trees or inferring the best network under the parsimony framework. Run it and have a look to the output to understand it. (The output can be found in* `out_PhyloNet_0.txt` *if needed). Modify the file to run a ML analysis with the maximum number of reticulations equals to 1. Then, modify the file to run a maximum pseudo-likelihood analysis with maximum number of reticulations equals to 1.What do you notice? (The outputs can be found in* `out_PhyloNet_1.txt` *and* `outPhyloNet_2.txt` *if needed). What the option* `-bl` *does? What happens if you remove it from the command to compute the probability of sn0 ? (If you cannot find how to modify the commands using the PhyloNet webpage, see the commands in the file* `commandPhyloNET_Triticeae-h.txt`*)*

Question 17 *In* `commandPhyloNET_TriticeaeSmall.txt`*, you will find a smaller example on which to run the more computationally expensive methods such as computing the likelihood of a network without using the information of branch lengths of the gene trees, inferring the best network under the ML framework or performing a Bayesian estimation of the posterior distribution of phylogenetic networks. (The output can be found in* `out_PhyloNet_3.txt` *if needed). Use the the Summarization function of* `MCMC_GT` *on the file* `out_PhyloNet_3.txt_part` *to evaluate convergence. In the ML analysis, can you asses the topology robustness via parametric bootstrap? (The outputs can be found in* `out_PhyloNet_4.txt` *if needed). In the Bayesian analysis, can you modify the chain length, the burn-in length and the sample frequency, and use pseudo likelihood instead of full likelihood to get some results in the limited amount of time we have?* ☺

FYI: Scalability (from a talk of Yun Yu):

- Maximum Parsimony: 30 taxa, 3 or 4 reticulations
- Maximum Likelihood: 10 taxa, 2 or 3 reticulations
- Maximum Pseudo-Likelihood: 30 taxa
- Bayesian: 10 taxa, 2 or 3 reticulations

4.2 PhyloNet from sequences

Question 18 *`PhyloNet` also permits to reconstruct networks from bi-allelic genetic markers (SNPs, AFLPs, etc). Run `commandPhyloNET_dolphins_binary.txt` to perform a maximum likelihood estimation of phylogenetic networks from this kind of data. Understand all the parameters in the command line and look at the output file `out_PhyloNet_5.txt` to learn how to read the output.*

Note that, if we had the time (a lot of time ☺), we could have run `MCMC_BiMarkers` to do a Bayesian estimation of networks for biallelic markers.

Unfortunately, we are running out of time and we will not be able to test the command `MCMC_SEQ` which permits the co-estimation of reticulate phylogenies (ILS and hybridization), gene trees, divergence times and population sizes on sequences from multiple independent loci. You can find all the information you need here: `https://wiki.rice.edu/confluence/display/PHYLONET/MCMC\_SEQ`. Note all that a competitor method as been implemented in BEAST and can be run via BEAST2 (with SpeciesNetwork module preinstalled). Your favorite workshops even provide you with a tutorial for it: `http://evomicsorg.wpengine.netdna-cdn.com/wp-content/uploads/2018/01/tutorial.pdf`

Still, these methods are **very** time consuming and they are applicable only to very few taxa.

4.3 Bonus section: Snaq from trees

Question 19 *Use the file `runSnaq.jl` to run a pseudo-likelihood analysis based on quartets on the data set of Question 14 (from Julia, do include("runSnaq.jl"). Output available in the file `out_Snaq.txt`*

A tutorial for `Snaq` can be found here: `https://github.com/crsl4/PhyloNetworks.jl/wiki`.