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

Celine Scornavacca

February 2nd, 2017

1 Phylogenetic networks from distances

Question 1 *Using `SplitsTree`, compute the split decomposition network for the following matrix:*

$$\begin{array}{ccccc}
\mathtt{a} & 0 & 4 & 5 & 4 \\
\mathtt{b} & 4 & 0 & 3 & 4 \\
\mathtt{c} & 5 & 3 & 0 & 3 \\
\mathtt{d} & 4 & 4 & 3 & 0
\end{array}$$

Refer to the tab "Help" → "Network Syntax" → "Distances" to learn how to specify a distance matrix in Nexus format, or look for this information in the manual.

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 chicken populations (see file `chickens.nexus`). Additionally, reconstruct the split decomposition network for the same data. In what way do the two networks differ?*

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 trees

Question 5 *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 6 *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 5? Does the network change if we modify the number of runs or apply the refined heuristic?*

Question 7 *For the data set of Question 6, 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 8 *Now open the software* `Dendroscope`*. Reroot the trees of Question 5 at the taxon S. cerevisiae (Hint: draw all the trees in the same window and select them all). Compute rooted consensus networks with the same thresholds used in Question 5 (use the Cluster Network Consensus method). Can you spot the same incongruences? Do the same for the trees of Question 6, rooted at A._thaliana.*

Question 9 *Still in* `Dendroscope` *and with the trees of Question 5 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? Can you spot some well-supported clades?*

Question 10 *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 11 *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? Compute also the hardwired distance, the rSPR distance, the hybridization number and the DTL reconciliation cost (with unitary costs for the events). Can you explain what those distances mean (Otherwise, I will ☺)? Apply the triplet consensus method* `simplistic` *to the trees. Align them using the tanglegram algorithm and save the image.*

3 Phylogenetic networks from sequences

Question 12 *Using* `SplitsTree`*, compute the median network for this binary matrix.*

a	0000000
b	0110000
c	1101100
d	1110110
e	0110101

Refer to the tab "Help" → "Network Syntax" → "Characters" to learn how to specify a sequence matrix in Nexus format, or look in the manual.

Question 13 *Using* `SplitsTree`*, compute the median network for the matrix contained in the file* `lugens-1.txt`*.*

Question 14 *Using SplitsTree, open the file example-quasi-median.nexus containing the quasi median network for the following multi-state characters matrix:*

a	A	A	A	A	A
b	T	T	A	A	A
c	A	T	A	T	T
d	A	A	T	T	C
e	A	A	C	T	C

How many vertices are in this network? For how many sequences? Can you guess the main problem when reconstructing (quasi) median networks?

Question 15 *Recently, 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... you will not find it in the manual.)*

Question 16 *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?*

4 Phylogenetic networks + ILS 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.

In commandPhyloNET_Triticeae.txt you will find an example input for PhyloNet to perform some analyses on the data set of Question 11 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. 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.

Modify these files to have a glimpse of the potential (and limits) of PhyloNet.
If time, think about how to control for Model Complexity, e.g. via K-fold cross-validation (see the InferNetwork_ML_CV method), AIC, BIC ...

5 Bonus section

I did not talk about reconciliations in this workshop but Matthew and Toni did, and I work on this topic a lot. So, I provided you with a nice program (ecceTERA) to reconcile gene trees with a species tree/network, correct gene trees, give the list of orthologous genes, etc. Try it if you want ☺