A phylogeny of apes based on the mitochondrial genome

Bioinformatics Community of Practice, BecA-ILRI Hub, September 2018
Trainer: Jean-Baka Domelevo Entfellner

The aim of this practical is to work on a phylogenetic study of seven species of apes and discuss the results by comparing different methods of phylogenetic inference.

The data we are going to work on is a subset of the mammalian sequences analysed by Cao et al. (*Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders*, J. Mol. Evol 47, 307–322). It is made of the 12 proteins encoded by the heavy strand of the mitochondrial genome. They are concatenated into one long sequence because they seem to have similar substitution patterns.

(1) Get from GenBank the 7 sequences corresponding to that dataset (accession numbers D38112 to D31116, X97707 and X99256).

(2) Align them by yourself (using the HPC, your own computer or https://ngphylogeny.fr/) using 2 different aligners of your choice among ClustalO, Mafft and Muscle. Are the alignments you get significantly different? What is the proportion of gaps? With Seaview or Jalview, try and get more advanced statistics about the observed differences on the various sites.

(3) Calculate how many binary unrooted trees exist for these 7 taxa. Is this number tractable, i.e. would it be feasible to test them all?

(4) Which substitution model would you use to calculate pairwise distances between taxa on this dataset, or to use in ML inferences on the same dataset? Why?

(5) Exploring as many tree topologies as possible, we would like to get scatterplots of the log-likelihood as a function of the parsimony tree length, and of the likelihood tree length as a function of the parsimony tree length. Define what are these values. See if with highly customizable software (e.g. PhyML or RAxML) it is possible to get the data to build such plots.

(6) More to come...