ViroInf Workshop

October 5, 2022, Tutorial by Sophie Kersting

Some methods for tree reconstruction

We will go through all four exercises together, step by step.

1.) Introduction to R:
Examples and exercises can be found in the R script `1IntroductionToR.R` in the viroinf-hiddensee GitHub Repository → tutorials → wednesday_tree_reconstruction. Here are several useful shortcuts when working with R:

Usage	**Windows/Linux**	**Mac**
Execute next command:	Ctrl + Enter	Cmd + Return
Create „<-“ symbol:	Alt + -	Option + -
Show help for marked function:	F1	F1
Show code of marked function:	F2	F2

2.) Tree reconstruction methods:
Examples and exercises on NJ, MP and ML as well as the RF distance can be found in the R-script `2TreeReconstructionMethods.R`.

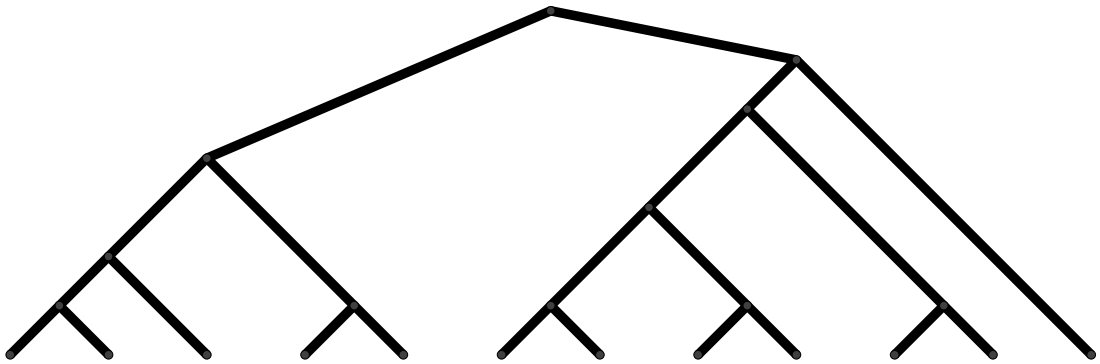
3.) Evidence for a criminal case:
„A gastroenterologist was convicted of attempted second-degree murder by injecting his former girlfriend with blood or blood-products obtained from an HIV type 1 (HIV-1)-infected patient under his care. Phylogenetic analyses of HIV-1 sequences were admitted and used as evidence in this case, representing the first use of phylogenetic analyses in a criminal court case in the United States.“[1]
Analyze the HIV-1 env and pol data sets from 1991 and find the evidence (more information on the structure and labels of the data set can be found in the `README.md` in the datasets directory of the viroinf-hiddensee GitHub Repository). The R-script `3ACriminalCase.R` contains a few hints which guide you through the analysis.

[1] From *Molecular evidence of HIV-1 transmission in a criminal case* by Metzger et al., 2002.

4.) A brief introduction to tree balance indices:
Tree (im)balance indices measure how balanced or imbalanced the topology of a phylogenetic tree is (ignoring edge lengths). There is a vast range of such indices. Today, we will have a look at the Colless index:
The Colless index $C(T)$ of a binary tree $T \in \mathcal{BT}_n^*$ is defined as

$$C(T) \coloneqq \sum_{v \in \mathring{V}(T)} bal_T(v) = \sum_{v \in \mathring{V}(T)} |n_{v_1} - n_{v_2}|$$

with $\mathring{V}(T)$ denoting the set of interior vertices of T, v_1 and v_2 denoting the children of v and n_u being the number of descendant leaves of vertex u.



Tree imbalance indices like the Colless index are commonly used to create hypothesis tests to see whether an evolutionary model fits as an explanation for the evolutionary history of a given tree. A small exercise on this can be found in the R-script `4TreeBalance.R`.

Further information and tutorials:
Phylogenies in R: `http://www.phytools.org/Cordoba2017/ex/2/Intro-to-phylogenies.html`
Estimating phylogenetic trees with R: `https://cran.r-project.org/web/packages/phangorn/vignettes/Trees.html`
More on `phangorn` package: `https://mran.microsoft.com/snapshot/2018-03-30/web/packages/phangorn/vignettes/phangorn-specials.pdf`
Clean R code: `https://style.tidyverse.org/syntax.html`