Lexi Xia

Autumn 2020

Phylogenetic Tree DRP

During this quarter, I studied the application of statistics in evolutionary biology that scientists could utilize statistical models such as maximum parsimony, maximum likelihood, or Bayesian inference to construct a Phylogenetic Tree. In this project, I read the book Tree Thinking and learned what is the phylogenetic tree, as well as some new statistical models that can be applied in the evolutionary biology field. Despite reading the textbook, we also explore some other resources for better understanding the statistical methods.

A phylogenetic, or evolutionary tree, represents the evolutionary relationships among a set or a group of organisms, generally called taxa. The Phylogenetic trees are just hypotheses, not facts; The "believed" tree often depends on how biologists view the importance of some characters. And the topology of trees can vary a lot when using different methods and models to construct the tree. The statistical methods chosen in constructing trees decide the pattern of branching in a phylogenetic tree, which reflects how species or other groups evolved from common ancestors.

The two most important methods we explored in this project were the maximum likelihood and Bayesian inference. The maximum likelihood (ML) method evaluates the phylogenetic hypothesis by Computing the likelihood of seeing the data we observe given the tree we are assuming, pick the tree with the highest probability. The second approach is Bayesian Inference. The Bayesian method is relatively new to the evolutionary biology field but has been widely accepted and used recently. Bayesian inference (BI) programs use a Markov chain Monte-Carlo (MCMC) algorithm to estimate the posterior distribution of phylogenies by sampling trees in proportion to their likelihood. The resulting chain can be summarized in a consensus tree made of compatible splits found among trees in the chain. When applying Baye's theorem in the phylogenetic tree, the bayesian approach combines the prior probability of a tree P(A) with the likelihood of the data (B) to produce a posterior probability distribution on trees P(A|B). The posterior probability of a tree will indicate the probability of getting the tree we assume given the data. Thus, the tree with the highest probability will be the tree best represents the data set.

There are also many other methods and algorithms being used to construct and compare trees. But the Phylogenetic trees are just hypotheses, not facts; The "believed" tree often depends on how biologists view the importance of some characters. And the topology of trees can vary a lot when using different methods and models to construct the tree. None of the algorithms chosen in constructing trees are perfect. And a lot of factors should be taken into account when choosing a model.

My takeaway from the project is that I found phylogenetic tree is very interesting and I want to explore more on how these statistical methods can be applied to construct trees. I haven’t got a chance to construct a tree by hand, so I would like to do more research and find a dataset to practice it.