Kyle D’Souza

Week 1 – 6/25-6/29/2018

Statistics Research 2018

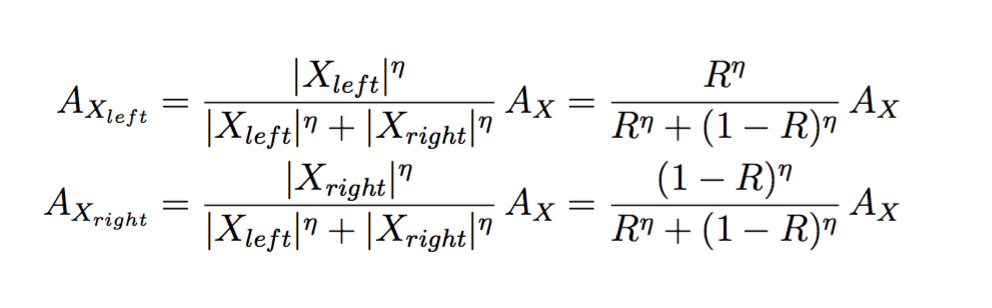
Proposal

**A. Motivation & Research Question**

Phylogenetic, or evolutionary, trees, have been used to represent the evolutionary relationships among organisms. All life on Earth can be modeled as part of a phylogenetic tree, indicating common ancestry. Correspondingly, phylogenetic trees have become instrumental to understanding biodiversity, evolution, ecology, and genomes.

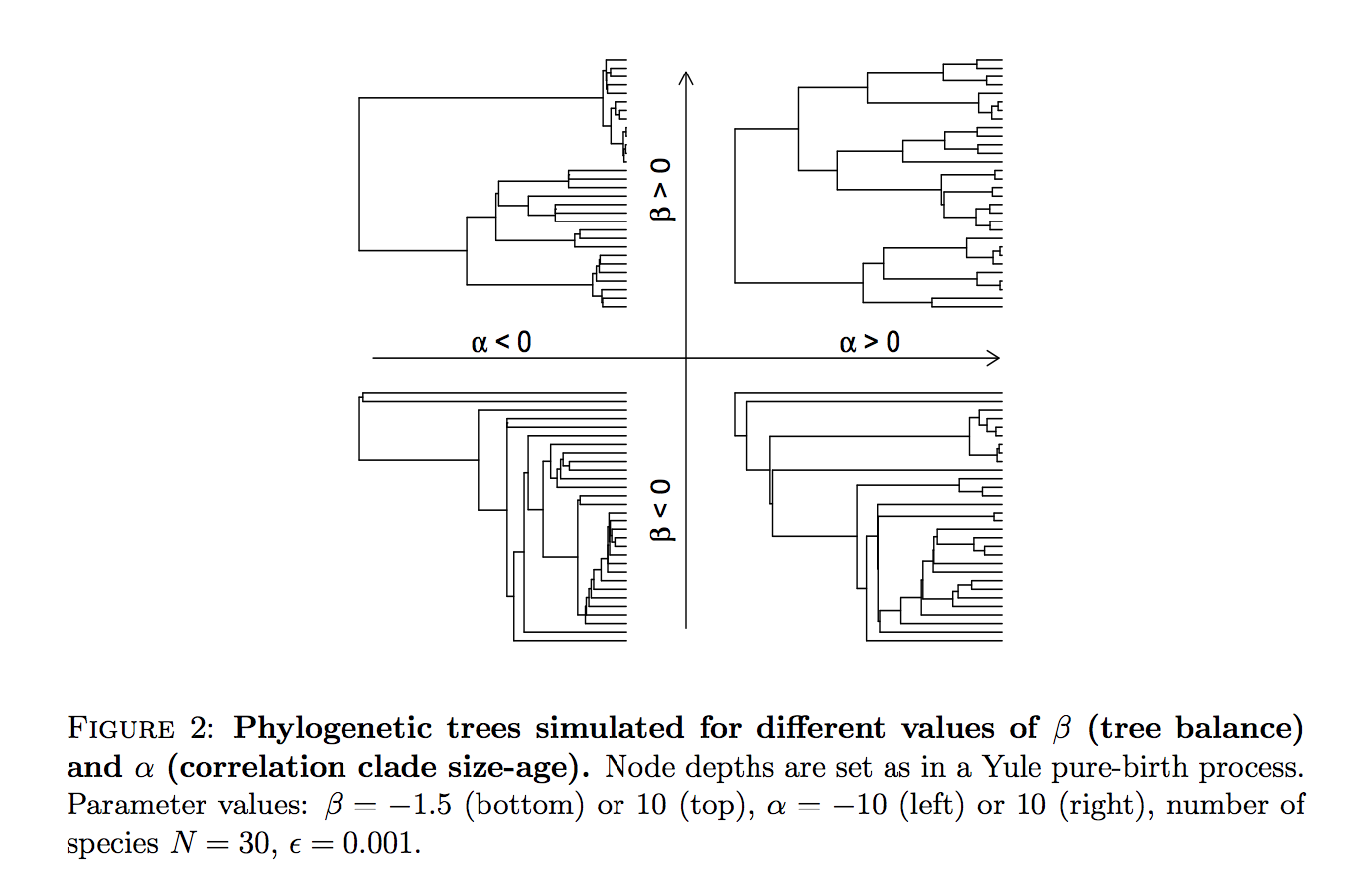
In 1925, mathematician Udny Yule published the paper: "A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S.". In it, Yule proposed a stochastic process that leads to a distribution with a power-law tail – in this case, the distribution of species and genera in what was termed the Yule process, or preferential attachment. The paper was insightful, as it showed that trees of evolutionary processes were unusually unbalanced. Since then, over the past 80 years, researchers have been puzzled trying to predict the observed pattern of imbalance. Researchers have tried to understand this through stochastic models, scatter diagrams as well as balance statistics, but these still do not allow us to compare two trees of different sizes without some form of standardization. Additionally, the need for a quantitative tool to measure differences between tree structures has been apparent for over 40 years (cf. Dobson (1975), Farris (1973), Sokal and Rohlf (1962), Bobisud and Bobisud (1972)) and researchers have analyzed this work using a variety of different distance axioms.

In recent years, there has been new research on using new tree structures to map evolutionary legacy and phylogenetic diversity. In particular, Amaury et al. (2017) introduced the concept of ranked tree shapes, a new, 23 sampling-consistent, three-parameter model generating random trees with covarying 24 topology, clade relative depths and clade relative extinction risks. These trees are an extension of the Beta-splitting model described by Aldous (2001), except with two additional parameters: parameter α, which quantifies the correlation between the richness of a clade and its relative depth, and a parameter η, which quantifies the correlation between the clade and its frequency. The abundance is described by the following equation:



Additionally, the tree simulations for different alpha and beta values in this example are shown here:

*Figure A: Phylogenetic Trees simulated for different alpha and beta values (tree balance)*

**

The primary goal of my research over the next 7 weeks is to develop metrics behind ranked tree shapes, namely calculating distance between trees. To do so, I will assemble a significant literature review on relevant topics around phylogenetic trees and simulate different datasets on ranked tree shapes to see if the axioms on measuring distance between trees are accurate. This process will use ancestral inference based off of data from UK Biobank, which will be explored in more detail in the following section.

**B. Potential Data Sources**

For this work, I will use data from UK Biobank (http://www.ukbiobank.ac.uk/), which holds a large amount of data from half a million participants aged 40-69 recruited between 2006-2010 in the United Kingdom. While I would eventually like to look at a specific gene (i.e. obesity, height, risk of certain illness) to focus on examples of random and non-random (human influenced) evolutionary process, this will be slightly exogenous to the goals of our current research, which is namely to focus purely at the sequence data.

**C. Tentative Methodology**

In order to build a metric that can work for a variety of different types of data, I would like to simulate these trees and build the trees with data that can test different parameters of the model. More specifically, I plan to run the model on both data that results from a large amount of selection as well as data that results from purely random selection (no specific biological force).

In terms of methodology, I will genome sequences of individuals from the UK Biobank, and construct a maximum likelihood tree of the genetic data using ancestral inference and a particular coalescent process. For each tree, I will compute a series of statistics – parent/child statistics, distance, clades, etc. Lastly, I will apply a series of axioms to analyze the distance between trees.

**D. Sources**

Aldous, David J. "Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today." *Statistical Science* (2001): 23-34.

Colijn, G. Plazzotta; A Metric on Phylogenetic Tree Shapes, *Systematic Biology*, Volume 67, Issue 1, 1 January 2018, Pages 113–126, <https://doi.org/10.1093/sysbio/syx046>

Holmes, Susan. “Modern Statistics for Modern Biology”. Chapters 2, Chapters 8. (2018). <http://web.stanford.edu/class/bios221/book/>

Luikart, Gordon, et al. "The power and promise of population genomics: from genotyping to genome typing." *Nature reviews genetics* 4.12 (2003): 981.

Maliet, Odile, Fanny Gascuel, and Amaury Lambert. "Ranked tree shapes, non-random extinctions and the loss of phylogenetic diversity." *bioRxiv* (2017): 224295.

doi: <https://doi.org/10.1101/224295>

Margush, T. “Distances Between Trees.” *Discrete Applied Mathematics* 4 (1982) p. 281-290. <https://www.sciencedirect.com/science/article/pii/0166218X82900506>

Tamura, Koichiro, et al. "MEGA6: molecular evolutionary genetics analysis version 6.0." *Molecular biology and evolution* 30.12 (2013): 2725-2729.

Uricchio, Lawrence H., Tandy Warnow, and Noah A. Rosenberg. "An analytical upper bound on the number of loci required for all splits of a species tree to appear in a set of gene trees." *BMC bioinformatics* 17.14 (2016): 417.

**E. Timeline**

Week 1:

Review Literature et al., Write Proposal, Aldous Reading.

Week 2:

Submit Proposal, Chapter 2 of Susan Holmes textbook, complete Maliet

Week 3:

Develop Metrics, Chapter on Trees in Susan Holmes textbook, assemble all setup necessary for simulations.

Week 4:

Simulate and annotate – have a rough outline of paper.

Week 5:

Present project at midterm point for feedback and iteration.

Week 6:

Finish first writeup of paper.

Week 7:

Edit and Review – iterate on past metrics.

Week 8:

Submit Final Draft.

**F. Notes**