Kyle D'Souza
Week 1 – 6/25-6/29/2018
Statistics Research 2018

Proposal

A. Motivation & Research Question

Phylogenetic, or evolutionary trees, are commonly used to represent the evolutionary relationships among organisms. All life on Earth can be modeled as part of a phylogenetic tree, reflecting shared 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 tried to describe the asymmetry found in natural evolutionary processes. 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 across different species and genera. 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. Slowinski 1990; Guyer and Slowinski 1993; Kirkpatrick and Slatkin 1993; Mooers and Heard 1997; Stam 2002; Blum and François 2006; Purvis et al. 2011; Wu and Choi 2015) 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, sampling-consistent, three-parameter model generating random trees with covarying 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:

$$A_{X_{left}} = \frac{|X_{left}|^{\eta}}{|X_{left}|^{\eta} + |X_{right}|^{\eta}} A_X = \frac{R^{\eta}}{R^{\eta} + (1-R)^{\eta}} A_X$$
$$A_{X_{right}} = \frac{|X_{right}|^{\eta}}{|X_{left}|^{\eta} + |X_{right}|^{\eta}} A_X = \frac{(1-R)^{\eta}}{R^{\eta} + (1-R)^{\eta}} A_X$$

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

Figure A: Phylogenetic Trees simulated for different α and β values (tree balance)

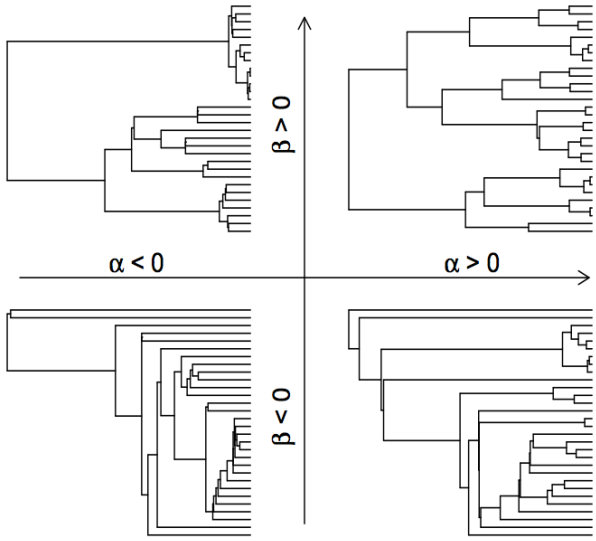

FIGURE 2: **Phylogenetic trees simulated for different values of β (tree balance) and α (correlation clade size-age).** Node depths are set as in a Yule pure-birth process. Parameter values: $\beta = -1.5$ (bottom) or 10 (top), $\alpha = -10$ (left) or 10 (right), number of species $N = 30$, $\epsilon = 0.001$.

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 ascertain the most appropriate distance between ranked tree shapes. 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 telomeres to study aging) 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 on ranked tree shapes under a variety of models, I would like to simulate these models and use estimated trees from previous studies that can test different

parameters of the model. More specifically, I plan to simulate from models with different selective pressures and different α and β values.

In terms of methodology, I will use sequence of individuals from the UK Biobank, and construct a maximum likelihood tree of the genome 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 analyze the properties of different metrics 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