Christian R Revels-Robinson
DREU Final Report
Summer 2021

ABSTRACT

This paper details the scholarly research conducted under the supervision of Dr. Kevin Liu of Michigan State University. The Distributed Research Experience for Undergraduates allowed a multitude of young intellectuals to study under leading computer scientists across the nation. Dr. Kevin Liu heads a Phylogenetics Lab at Michigan State University in which he explores genetic and genomic sequencing using specialized computer programs. One of the most vital aspects of this research is the statistical approximations component as there are various statistical distributions used to calculate the history of certain species and organisms. This report details my first undergraduate computer science research experience.

PAPER

The undergraduate students studying under Dr. Liu were largely first time researchers. Further, a significant amount of these students did not have a formal collegiate level-biology background. For example, I have not taken a biology course since I was in high school. Thus, one of the first requirements was to read an immense amount of background information. Dr. Liu first recommended I read through Computational Phylogenetics: An introduction to designing methods for phylogeny estimation by Tandy Warnow. This textbook provided an in-depth introduction to phylogenetic estimation. One of the biggest tenets of phylogenetics is the explanation of evolution and this can be in the field of speech and language or biological characteristics. Warnow explained the Cavender-Farris-Neyman(CFN) model and evolutionary tree which is one of the oldest models, specifically rooted and unrooted examples. Warnow also detailed some of the more practical evolutionary models. During this time, it was immensely helpful to speak with my PhD candidate mentor who spoke to me about her original scholarship and just how useful CFN is not. Warnow's readings also took a deep dive into the significance of Trees within the framework of phylogenetics. Warnow took the description of trees to the next level in detailing the construction of trees from true subtrees. I learned immensely about tree construction algorithms including: the ASSU algorithm which allows one to construct rooted trees from rooted triples, the all quartets method, constructing trees from a subset of irs quartet trees, as well as inferring quartet trees from other quartet trees. Moreover, one of the more difficult concepts to understand came as Warnow embarked on the ways in which one can construct trees from qualitative characters. Warnow described a few methods of accomplishing this which included tree construction for Maximum Parsimony via the Fitch algorithm and the Sankoff algorithm. Warnow then described tree construction from compatibility and tree

construction based on Maximum Compatibility. Finally, the Warnow set of readings concluded with a walk through the distance-based tree estimation methods: UPGMA, Additive Matrices, The Four Point Method, The Naive Quartet Method, The Q* Method, Neighbor Joining, Distance-Based Methods as Functions, Optimization Methods, and Minimum Evolution. During the second half of the Warnow readings, Warnow introduced Molecular Phylogenetics and the statistical gene tree estimation methods. There are levels to the models of site evolution. There can be site evolution amongst the nucleotides or the amino acids. The methods used to survey this type of evolution are maximum likelihood, bayesian methods, as well as the statistical properties of maximum parsimony and maximum compatibility. One is also able to learn more about gene tree estimation by estimating branch support, sample complexity, heterotachy and the No Common Mechanism models.

For the second set of readings Dr. Liu tasked me with Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory by Jotun Hein, Mikkel H. Schierup and Carsten Wiuf. This reading began by explaining the "basic coalescent" using the Y chromosome data set as well as the Wright-Fisher Model. Gene Genealogies, similar to Warnow's readings, began with a lot of explanation surrounding relevant distributions such as that of the geometric and exponential. The reading then described the discrete time coalescent in terms of two genes as well as *n* genes. Gene Genealogies then detailed the mathematical models of alleles: the infinite alleles model, the infinite sites model and the finite sites model which helped understand the algorithms for simulating sequence evolution. For example, the reading used the Wright-Fisher Model with mutation to survey some of the algorithms for simulating sequence evolution. Gene Genealogy took a much simpler approach to coalescent and phylogenetic trees as well as analyzing nested subsamples and hanging subtreets(unbalanced trees). The final section of readings assigned from this source detailed the coalescence with population structure(finite island model, coalescent tree in the finite island model, general models of subdivision and non-equilibrium models.

The final set of readings came from Joseph Felsenstein's work Inferring Phylogenies. Felsenstein began his work by detailing Parsimony methods with a particular emphasis on rootedness and unrootdness, branch lengths and methods used of rooting the tree. Similar to some of the previous readings Felsenstein then took a deep dive into ways to count evolutionary changes which included the previously mentioned Fitch and Sankoff algorithms. Felsenstein ruminated on the connection between the two algorithms and how each respective algorithm is used to modify trees. Dissimilarly from some of the other authors, Felsenstein spoke about bifurcating trees and multifurcating trees as well as their tree shapes. In the second half of this reading Felsenstein went through distance matrix methods and their statistical rationale. Some of the methods Felsenstein discussed included the least squares method, generalized least squares, the Jukes-Cantor model, minimum evolution, clustering algorithms,

UPGMA and least squares, Neighbor-joining, and Distance Wagner. The next section centered on Models of DNA evolution such as Kimura's two-parameter model, f84, HKY, the general time-reversible model, LogDet distances and the Tamura-Nei model. The aforementioned section also covered rate variation between sites or loci. Felsenstein then went on to describe likelihood methods and how to compute the likelihood of a tree while considering ambiguity and error rates. Some of the topics reviewed in this section included rates varying among sites such as Hidden Markov models, autocorrelation of rates, HMM's for other aspects of models as well as how to estimate the states. One of the most interesting portions of this reading was the bit on whether Machine LEarning estimates are consistent which included a length proof. Felsenstein then brought up the Bayesian inference of phylogeny and Markov chain and Monte Carlo methods which appears to be quite consistent across the field. Overall, the set of readings allowed me to grow immensely as a student of evolution. Before embarking on this textbook journey I was not familiar with many of the statistical methods or software for phylogenetic study. My PhD mentor waas pivotal throughout this series of readings as I felt bombarded by the amount of content I was expected to retain. She truly helped me filter the information and make the learning process more concise. I am forever grateful for her patience and wisdom.

TUTORIALS

After I finished this series of readings I was tasked with working directly with the phylogenetic software to simulate and build evolutionary trees. This task proved quite challenging as I struggled to work with my newer Macbook device which was not compatible with a decent amount of the industry standard software. I tirelessly tinkered with the INDELible software. INDELible is an application for biological sequence simulation. INDELible is able to simulate the evolution of nucleotides, amino acids, or codon data sets through insertion, deletion and substitution in continuous time. While I was stuck on the problem of using the software on my laptop as well as the fact that my environment did not allow me access to any other feasible methods to access the application, I persisted. Ultimately, as a result of the meetings with my PhD mentor I opted to use a virtual environment. I soon transformed the harddrive of my computer by installing VirtualBox to make use of the Ubuntu operating system since my operating system did not allow for use of the software. My mentor helped me to discover that newer Mac models do not operate 64 bit programs which was very surprising. Learning about virtual environments as a result of this hiccup was one of the bigger and more unexpected takeaways of the research process. Ultimately, I was able to create the non-ultrametric birth-death tree. The program allowed me to use a designated amount of taxa as well as to set the tree depth. I then used this tree as the "model species tree" to use the ms software to perform a single simulation under the multi-species coalescent model. MS is a program for generating samples under neutral models. The program is aimed at investigating the statistical properties of samples. When using MS, I set the number of samples and

replicates. I was also able to use certain flags on the command line to output the gene tree corresponding to the simulated local coalescent history as well as the time for divergence event and the specified populations. I was able to get one local gene tree as output. After getting this output I then simulated the DNA sequence evolution down the gene tree using Seq-Gen. Seq-Gen is a software program that simulates the evolution of nucleotide or amino acid sequences along a phylogeny, primarily using different models of the substitution process. I used the Jukes-Cantor model with the Seq-Gen platform. After I received the data from this iteration I had officially simulated a full synthetic dataset for the experiment. The next step in the process was to use FastTree to estimate a maximum likelihood gene tree under the Jukes-Cantor model. FastTree is a software that infers approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences. While the estimated gene tree was not topologically identical to the true gene tree, you are able to quantify the topological differences between the two trees using the Robinson-Foulds distance. I used the symmetric difference function in the dendropy method to calculate the Robinson Foulds distance. The aforementioned series of experiments did not go as intended as the process took various weeks. The biggest challenge came with navigating the respective software while making sure the input and output were as expected.

Overall, I would like to thank Dr. Liu as well as the DREU team for dedicating resources to individuals like myself who are interested in conducting research. As Dr. Liu has said numerous times, research is about the journey as the results are not always ideal nor is the intended process.