1 Primary Goal: Review of Bayesian Methods for Phylogenetic Reconstruction

The primary goal of this project is to produce a paper which explores Bayesian approaches to phylogenetic tree reconstruction. What follows is a rough outline of the sections of that paper.

1.1 Motivation: The Flaws of MP and ML Methods

In our first section, we will explore the problems posed by ML (maximum likelihood) and MP (maximum parsimony) methods. Based on our initial readings, these problems include the statistical inconsistencies that MP methods can produce (long branch attraction) and the computational intensity of ML methods. A good resource for this section will be a paper by Holder and Lewis which explores the pros and cons of various phylogenetic algorithms.

1.2 The Theory of Bayesian Phylogenetic Reconstruction

1.2.1 Bayesian Probability

Before introducing the phylogenetic algorithms, or even MCMC, we will need to cover some basic Bayesian probability theory, including the motivation for treating parameters as random variables and some theorem's/formula's for calculating posterior distributions.

Reference: Robert W. Keener: Bayesian Estimation

1.2.2 MCMC

Now we will briefly explain how MCMC, specifically Hastings-Metropolis MCMC, allows us to efficiently estimate the posterior distribution. We will probably not include any proofs of the correctness of the algorithm, but we will explain it's inputs, outputs, and general logic.

References: another section in the statistics textbook cited above, or this seminal paper by Hastings

1.2.3 Bayesian Reconstruction with MCMC

This will probably be the meatiest section of the paper. Here we will illustrate how the theory in the previous two sections can be applied to reconstructing phylogenetic trees. We will introduce the concept of evolutionary models as a method for calculating the probability of the observed DNA data given a tree. We will then explain different approaches to actually using MCMC to calculate the probability of a given phylogenic tree given the data. The information from this section will be based in large part on the following primary literature:

- Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods
- Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method
- Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees

1.3 Survey State of the Art Methods

In our final section, we will survey the current state of the art software to understand how Bayesian methods can be improved and made more efficient. A tentative list of software to be reviewed includes, MrBayes, Bali-Phy, Beast.

2 Stretch Goal: Implement a Bayesian Reconstruction Algorithm

If we have time, we will implement a basic algorithm MCMC for recreating phylogenetic trees. We would only attempt this after the paper had been written. Writing a program that accurately performs MCMC seems like it could be a bit of a debugging nightmare, so we would want to be sure that we had plenty of time remaining before the end of the project. If we do pursue this goal, we would try and obtain a dataset used in one of the primary literature papers listed above and recreate their results.