2012

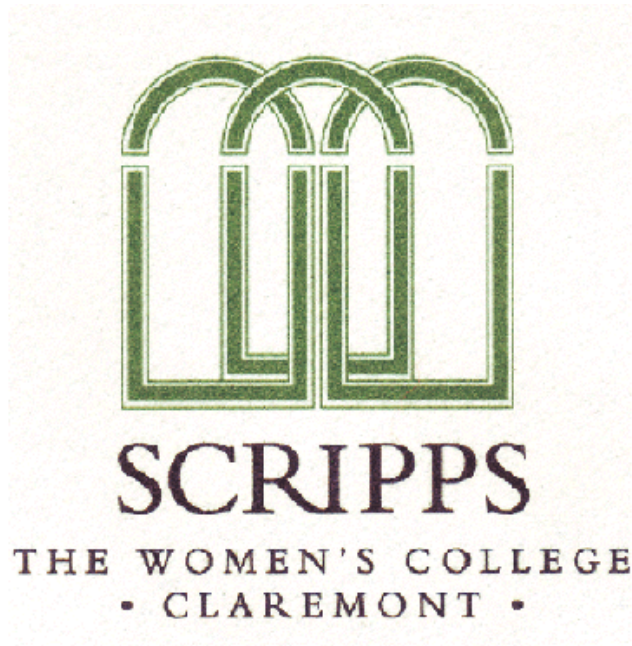# Constructing Phylogenetic Trees Using Maximum Likelihood

Anna Cho
Scripps College

Constructing Phylogenetic Trees using Maximum Likelihood

Anna Cho

Christopher Towse, Advisor

Ami Radunskaya, Reader

Submitted to Scripps College in Partial Fulfillment
of the Degree of Bachelor of Arts

March 9, 2012

Department of Mathematics

Abstract

Maximum likelihood methods are used to estimate the phylogenetic trees for a set of species. The probabilities of DNA base substitutions are modeled by continuous-time Markov chains. We use these probabilities to estimate which DNA bases would produce the data that we observe. The topology of the tree is also determined using base substitution probabilities and conditional likelihoods. Felsenstein [2] introduced this method of finding an estimate for the maximum likelihood phylogenetic tree. We will explore this method in detail in this paper.

Contents

Acknowledgments

I'd like to thank my parents for their love and support. I wouldn't have gotten to the point I am at today without them. I'd also like to thank my advisor, Professor Towse, for all of the advice and emotional support that he offered me on my thesis. Thank you for for your energy and helping me see the beauty of math. I'd also like to thank my reader, Professor Radunskaya, for all of her brilliant advice. I would have been lost without your help.

Chapter 1

The Stochastic Process of DNA Base Substitutions

More and more DNA sequences are being analyzed today than ever before. With this increase in the accumulation of DNA sequences comes a demand for the study of the ancestry of organisms and their phylogenetic trees. Scientists are interested in how closely related one species is to another. Studying phylogenetic trees and the evolutionary processes that they model allow scientists to gain a better understanding of how organisms have arrived at the state they are in today. Phylogenetic trees give us the ability to see how species evolve and adapt throughout different time periods with different conditions and needs. Studying these evolutionary processes is clearly important to the advancement of biology, but finding the correct phylogenetic tree for a set of related species is very difficult considering that we are only given the data that we can observe today, namely the DNA sequences of those species. We have overcome much of this difficulty using statistical inference. Statistical models and Markov models allow us to estimate how similar a phylogenetic tree is to the actual, unknown phylogenetic tree for a given set of DNA sequences. We use the maximum likelihood method to infer what the true phylogenetic tree of our set of data looks like. Maximum likelihood uses an explicit evolutionary model. We assume that the data we observe is identically distributed from this model.

Before defining maximum likelihood, we review some of the terminology used in this statistical approach. We have been using the term *phylogenetic tree* to indicate a branching diagram describing a set of species and their common ancestors. The terms *phylogenetic tree* and *evolutionary tree* are often used interchangeably in the field of computational biology. In

this paper, we will be using the term *phylogenetic tree* exclusively. The set of data, namely the DNA sequences of the species we are observing, will be at the *tips* of the phylogenetic tree. The internal nodes of the tree represent the DNA sequences of the ancestors of the species we are examining. The segments of the diagram that connect one DNA sequence to another are called the *branches* of the tree. Finally, the *root* of the tree represents the DNA sequence of the sole common ancestor of all of the species we observe in our set of data. Each species at the tip of the tree can be traced back to this common ancestor at the root of the tree. Figure 1.1 shows an example of a phylogenetic tree.
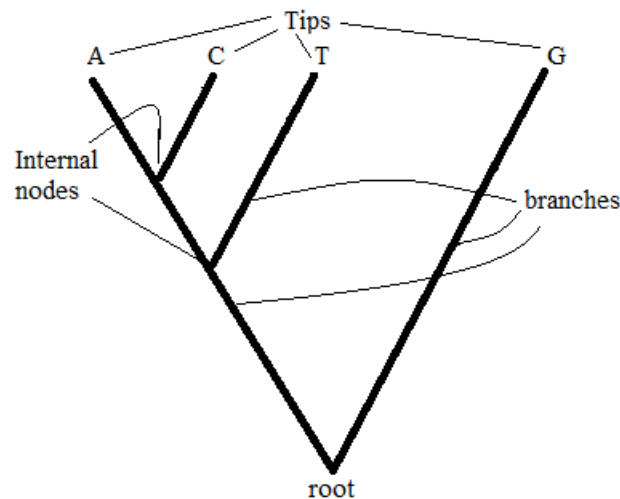
Figure 1.1: An example of a phylogenetic tree. Note: This tree represents the phylogenetic tree for one site in the DNA sequences (i.e., the DNA base of each species' DNA sequence that is located at the same place).

We refer to the tree shape or the way in which the tips, nodes, and root are connected by branches as *topology*. It is important to note that this use of the word topology is different from the branch of mathematics known as topology. To topologists, each phylogenetic tree would have the same, trivial topology and would be indistinguishable. In evolutionary biology, two topologies are considered different from one another if one topology cannot be cannot be recreated from the second topology without disassembling a connection between two nodes or between a node and a tip. Figure 1.2 shows both an example of equivalent topologies and an example of topologies that are different from one another.

Figure 1.2: Tree A and Tree B have equivalent topologies. Tree C has the same number of tips as Tree A and Tree B, but the branches connecting the two middle tips cannot be changed to look like Tree A and Tree B without detatching them. Thus, Tree C has a different topology from Tree A and Tree B.
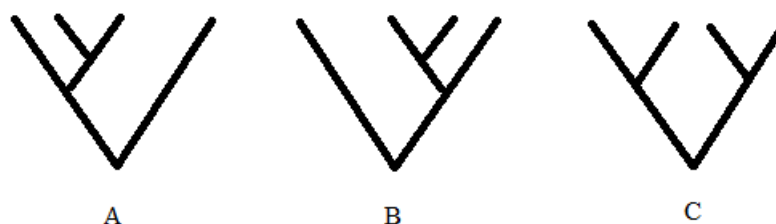
1.1   Likelihood

The process of finding a phylogenetic tree using maximum likelihood involves finding the topology and branch lengths of the tree that will give us the greatest probability of observing the DNA sequences in our data. After each step, we take the likelihood of each tree that we examine. The tree that gives us the largest likelihood is then chosen to be examined in the next step. We will describe this process in more detail in Chapters 2 and 3.

Definition 1. The *likelihood* of a set of data, D, is the probability of the data, given a hypothesis, θ. The hypothesis will usually come in the form of different parameters. We denote the likelihood, L, of a set of data, D, as $L = P(D \mid \theta)$.

This definition seems quite simple, but we also need to be careful not to use the term, *likelihood* as we otherwise would in English. When we say, "the likelihood of a phylogenetic tree," we are not referring to the probability of seeing that particular tree. Rather, we are referring to the probability of seeing the DNA sequences that we have in front of us, given that phylogenetic tree.

In order to see how likelihood works, we will first consider a very simple example.

Example 1. Suppose there is a gameshow that uses a robot to decide whether or not a contestant will win a prize of $5,000. If the robot raises its left arm, then the contestant wins the prize, and if the robot raises its right arm the contestant goes away with nothing. The robot is controlled by a computer

that randomly chooses the probability with which the robot will raise its left arm for the entire season of the game show. Out of ten episodes of the gameshow, six contestants left with a prize while 4 contestants went away with nothing. One ambitious frequent viewer of the show is interested in finding the likelihood of the robot's decision to raise its left arm for this season. Let p be the probability that the robot raises its left arm, and let X be the proportion of times the robot raises its left arm. One hypothesis we could consider is that the robot is fair. Then, the likelihood that $p = 0.5$ is $P(X = 3/5 \mid p = 0.5)$. Clearly, this probability is less than one. Another hypothesis we could consider is that the robot has a 3/5 probability of raising its left arm and a 2/5 probability of raising its right arm. The likelihood of this probability is $P(X = 3/5 \mid p = 3/5) = 1$. Then, $P(X = 3/5 \mid p = 0.5) < P(X = 3/5 \mid p = 3/5)$. The viewer's best guess is that the probability of the robot raising its left arm is $\frac{3}{5}$.

Similarly, when inferring phylogenetic trees using maximum likelihood, we are searching for the tree that gives the highest probability of producing the observed DNA sequences. For another example see [3], page 249.

1.2   Evolution and DNA mutations

Before we go any further into detail about phylogenetic trees, we describe the evolutionary process of DNA mutations that phylogenetic trees represent. DNA is short for deoxyribonucleic acid, and this genetic material makes up the chromosomes that give organisms the characteristics they have. DNA is made up of nucelotides. The four different kinds of nucleotides are distinguished by their different nitrogenous *bases*: adenine, guanine, cytosine, and thymine. We will denote these four different bases with A, G, C, and T, respectively. When an organism has offspring, its DNA is replicated and passed on to its offspring.

During DNA replication, changes in the DNA, which we call mutations, can occur. The changes in the genes that the offspring inherits gives them a different phenotype from their parents. The changes in the DNA occur at the level of the bases; different bases are substituted for the bases that the parents' DNA originally had. These alterations change the characteristics of the offspring and eventually, after a few generations have passed, may lead to the production of a new species [8].

In this paper, we look at the DNA sequences of a set of species and use maximum likelihood methods to determine the how closely related the

species are to each other. We model the probability of a DNA base substitution as a continuous time Markov process (see [4]). We will describe this in more detail in the following section.

1.3   Stochastic Processes

Recall that for a specific site in a DNA sequence, the bases (A, C, T, or G) observed in each sequence of DNA in our set of data are placed at the tips of the phylogenetic tree. In order to compute the likelihood of a phylogenetic tree producing the set of bases seen in our data, we must have a probability model describing the event of a mutation from one base to another. In other words, for each site in a DNA sequence, we must find a model describing the evolution of a site from a previous DNA sequence into a base we see at the tip of our tree.

   The next issue we encounter is the difficulty of putting all of the DNA sites together to form the DNA sequences that we are actually concerned with in our data. In order to make the computation of a tree's likelihood more feasible, we assume that the probability of mutation in one site of a sequence of DNA is *independent* of the probability of mutation in another site of that sequence of DNA. This means that given any information about site n, the probability of site m having state s is not effected by the information we have received about site n. The converse is also true.

Definition 2. We say that event A is *independent* of event B when the occurrence of B gives no information and does not change the probability that A will occur. In other words,

$$P(A \mid B) = P(A).$$

By the definition of conditional likelihood, $P(A \mid B) = P(A \cap B)/P(B)$. When A is independent of B, this is equivalent to

$$P(A \cap B) = P(A)P(B).$$

   This allows us to compute the likelihood of a given set of DNA sequences one site at a time. Once we have computed the likelihoods of each site of the sequences, the product of the likelihoods of each individual site gives us the likelihood of the set of DNA sequences as a whole. Thus, the bulk of the maximum likelihood method lays in finding the optimal phylogenetic tree for each single site of DNA.

For a single site of DNA, we denote the probability of a site in state i evolving to state j in t units of time as $P_{ij}(t)$. States i and j can be either 1, 2, 3, or 4, which correspond to the bases A, C, G, or T, respectively. We say that the *state space*, $S = \{1, 2, 3, 4\}$. When a site undergoes a mutation, its state is called a *random variable* in a *stochastic process* [5].

Definition 3. A *random variable* is a real-valued function of the outcome of an experiment, or a process described by a probablistic model.

Definition 4. A *probability mass function* (PMF), denoted p_X, of a random variable, X, is the function $p_X(x) = P(X = x)$.

Example 2. Suppose that at site n of the DNA sequence of a unicorn there are equal probabilities of finding base A, C, T, or G. Just as we did earlier, we will represent A, C, T, and G with 1, 2, 3, and 4, repectively. Then, we are applying a real-valued function to the possible outcomes for the base found at site n. The base we find at site n is a random variable which we will call X. The PMF for X is $p_X(x)$. We have that $p_X(1) = 0.25$, $p_X(2) = 0.25$, $p_X(3) = 0.25$, and $p_X(4) = 0.25$.

Definition 5. A *stochastic process* is an indexed family of random variables. In other words, it is a set of random variables whose values are assigned a t from the index, T. For each $t \in T$, the state of the stochastic process at step t of the process is a random variable denoted by $X(t)$. We denote a stochastic process by $\{X(t) : t \in T\}$.

When the indexing parameter, t, can take on a continuous range of values, we call the process a *continuous-time process*. The random variable $X(t)$ can take on a continuous or a discrete range of values.

Example 3. In the case of base substitutions, at any time, t, during the evolutionary process of a site of DNA, the state of the site can take on the random variable, 1, 2, 3, or 4. The process of substituting bases is indexed by the time, T, which has a continuous range. Thus, the process is a continuous-time process. The stochastic process, $\{X(t) : t \in T\}$ can take on the values 1, 2, 3, and 4 at each time, t, in our index, T.

The probability of a base substitution occuring at time t, which we denote as $P_{ij}(t)$, is the probability that a base, i, will undergo a mutation and be substituted by base, j, at time, t. Again, the states, i and j, are random variables, and the index of these random variables is time, T. This reflects a process with the *Markov property*: given the present state, the future states

are not dependent on past states. This means that a future state only depends on the present state, $X(s)$. It is not affected by any past state, $X(u)$, where $0 \leq u < s$. Since there are not any discrete indicators of when a base will mutate and the index, T, takes on a continuous range, the process of base substitution is a *continuous-time Markov process*, as described below.

Definition 6. A stochastic process, $\{X(t) : t \geq 0\}$, is called a *continuous-time Markov processes* when it posesses these properties:

1. Each event of the process is independent of previous events. (In other words, the process has the Markov property).

2. When entering a state, the process will stay in that state for a random amount of time before transitioning to another state.

We denote the probability that a continuous Markov process currently in state i will be in state j after t time units as

$$P_{ij} = P(X(t + s) = j \mid X(s) = i).$$

Matrices are used to represent Markov processes with rows representing all possible current states and columns representing all possible future states. Each term, P_{ij}, of a matrix representing a Markov process represents the probability of the process moving from state i to state j in a span of time, t.

Example 4. Suppose the robot in Example 1 works without any problems for a period which is exponentially distributed with parameter, λ. Then, it breaks down and a replacement robot will have to substitute for the broken robot on the gameshow for a period which is exponentially distributed with parameter, μ. Since the time that each robot spends on the gameshow is exponentially distributed with different parameters, the times that each robot spends on the show are independent of each other. Let $X(t) = 1$ if the robot is working at time t. Let $X(t) = 2$ if a replacement robot is being used at time t. Then, $\{X(t) : t \geq 0\}$ is a continuous-time Markov process with each event being independently distributed and the following transition probabilities: $P_{11} = P_{22} = 0$, $P_{12} = P_{21} = 1$.

As we stated earlier, sometimes, we represent base substitution probabilities in a matrix called a *transition matrix*. The rows of the matrix represent the initial state, and the columns represent the state after a base substitution has occurred. Our transition matrix for this problem would be:

$$P_{ij} = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

1.4   The Memoryless Property

Suppose that a continuous-time Markov chain enters state i at time 0. Next, suppose that the process does not leave state i in the following ten minutes. Then, what would the probability of the process remaining in state i in the following 20 minutes (30 minutes total) be? Since we have a continuous-time Markov chain, the process has the Markov property. Thus, the probability of the process remaining in state i from the time interval [10, 30] is simply the probability that it stays in state i for 20 minutes. In other words, if we let T_i represent the amount of time that the process remains in state i before transitioning into a different state, then

$$P(T_i > 30 \mid T_i > 10) = P(T_i > 20)$$

and in general,

$$P(T_i > s + t \mid T_i > s) = P(T_i > t)$$

for all s and $t \geq 0$. We say that a random variable X with this property is *memoryless*. The process's determination of what state it is in at time $t + s$ is not affected by the state it was in at time t for all $0 \leq t < s$; it does not remember its prior states. If we restate this property using the definition of conditional properties, we have

$$P(X > s + t \mid X > t) = \frac{P(X > s + t \cap X > t)}{P(X > t)} = P(X > \quad t)$$

which is equivalent to

$$P(X > s + t \cap X > t) = P(X > s + t)P(X > t). \tag{1.1}$$

We will use a stochastic process called the *Poisson process* in conjunction with Markov processes to model the base substitution probability $P_{ij}(t)$.

Definition 7. A stochastic process is a *Poisson* process with rate, λ for some $\lambda \geq 0$, if it posesses these properties:

1. At time $t = 0$, the number of events that have occurred is 0.

2. The time increments $t \in T$ are independent of each other.

3. The number of events in any interval of length t is a random variable with a Poisson distribution with mean λt.

If the *Poisson process* has $N(t)$ events at time t, then we represent the process by:

$$P(N(t+s) - N(s) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad \text{for } n = 0, 1, \ldots.$$

We will assume that in a small interval of time of length dt, there is a probability μdt that the current base at a site will transform, where μ is the rate of base substitution per unit of time. It is clear that at time $t = 0$, there will have been 0 base substitutions. Thus, the first property of our definition of a Poisson process is fulfilled. The probability of transitioning bases, μdt, is the same for all intervals of time with length dt. This is the second property of a Poisson process. Further, the number of base substitutions during a particular time interval is independent of the history of changes outside of this interval, so each time increment is independent from other. time increments Finally, the probability of a change in base in a time interval is very small and the number of base changes can be modeled by a Poisson distribution, satisfying the third property of a Poisson process. Thus, the process of changing from the current base to another base in a time interval dt is a Poisson Process. If we let $N(t)$ be the number of transitions from base i to base j at time t, then $P(N(t) = k) = e^{-\mu t}\frac{(\mu k)^k}{k!}$. Then, $P(N(t) = 0) = e^{-\mu t}$ and $P(N(t) > 0) = 1 - e^{-\mu t}$. If we let X be the time of the first event, then the probability of the first event occuring before a time $t > 0$ is $P(X \leq t) = P(N(t) > 0) = 1 - e^{-\mu t}$. The complement of this probability, the probability that no mutation occurs before a time $t > 0$, is $1 - (1 - e^{-\mu t}) = e^{-\mu t}$. Then,

$$P_{ij}(t) = e^{-\mu t}\delta_{ij} + (1 - e^{-\mu t})\pi_j \tag{1.2}$$

where π_j is the probability that a mutation will result in the current base being replaced with base j [2]. The δ is taken from the Kronecker delta formula in which

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

In other words,

$$P_{ij}(t) = \begin{cases} e^{-\mu t} + (1 - e^{-\mu t})\pi_j & \text{if } i = j \\ (1 - e^{-\mu t})\pi_j & \text{if } i \neq j \end{cases}.$$

1.5   Stationary Distributions

When determining the likelihood of the root of a phylogenetic tree, it is helpful to know the proportion of base substitutions that results in a specific base. In this section, we will be examining general ideas about Markov models. We will see that for some continuous Markov chains, a limiting probability, $lim_{t\to\infty}P_{ij}(t)$, which we denote, π_j, exists. This means that as time goes on to infinity, the probability of switching from state i to state j in time t approaches a constant.

The stationary distribution of an ergodic Markov process, π_i, is also interpreted as the proportion of time the process is in state i.

Definition 8. *Ergodicity* is the property that the limiting probability $lim_{t\to\infty}P_{ij}(t)$ exists and is independent of the initial state, i. We call a Markov process with this property *ergodic*.

Definition 9. Let $\{X(t) : t \geq 0\}$ be a continuous Markov process with state space S and transition probability matrix P. A state j is said to be *accessible* from state i if there is a positive probability that starting from i, the Markov process will be in state j after a finite time. In other words, $P_{ij}(t = n) > 0$, for some $n < \infty$.

Example 5. At the Pacific Science Center, there is an exciting exhibit that allows visitors to learn about Markov chains. When the line for the exhibit is empty, there is a positive probability that after n time units, the line will have one visitor. Then, if we represent the state of the line when it is empty with 0 and when it has one visitor with 1, we have that $P_{01}(t = n) > 0$. Thus, 1 is accessible from 0.

Definition 10. Let $\{X(t) : t \geq 0\}$ be a continuous Markov process with state space S and transition probability matrix P. If $i, j \in S$, and i and j are accessible from *each other*, then we say that i and j *communicate*.

Example 6. Suppose that in Example 5, there is also a positive probability that when the line has one visitor, there will be no visitors in line after m time units. Then, we have that $P_{10}(t = m) > 0$ along with $P_{01}(t = n) > 0$. Thus, 0 and 1 communicate.

Definition 11. If all of the states of a Markov process communicate with each other, then we call the Markov process *irrreducible*.

Example 6 is an example of an irreducible Markov process.

Definition 12. Let $\{X(t) : t \geq 0\}$ be a continuous Markov process with state space S and transition probability matrix P. Let $P_{ii}(n)$ be the probability, that starting from state i, the process will return to state i, for the first time, after finite time n. Let p_i be the probability that, starting from state i, the process will return to state i after a finite number of transitions. Then, $P_i = \sum_{n=1}^{\infty} P_{ii}(n)$. If $P_i = 1$, then the state i is called *recurrent*.

Example 7. In the Pacific Science Center problem presented in Example 5 and Example 6, It is clear that $P_{01} < 1$ and $P_{10} < 1$. This implies that there is a positive probability, that starting with an empty line, the line will remain empty for a finite amount of time, n. In other words, $P_{00}(n) > 0$. There is also a positive probability that after starting with an empty line, and having one visitor in the line, the line will be empty again after a finite amount of time, m. Then, $P_{11}(m) > 0$. Thus, at each time that the line is in state 0, after a finite amount of time, there is a chance that it will return to being in state 0. Then, $P_i = \sum_{n=1}^{\infty} P_{ii} = 1$ and state 0 is recurrent.

Definition 13. Let i be a recurrent state of a Markov process. The state i is called *positive recurrent* if the *expected* amount of time between two consecutive returns to i is finite.

The Pacific Science Center examples are all positive recurrent since the line will not stay in state 0 or state 1 for an infinite amount of time. For a nonexample, consider the robot gameshow example once more. Suppose that after a robot malfunctions, the gameshow completely replaces the robot forever. Then, if we let 1 denote the state that robot 1 is being used on the gameshow, the expected amount of time between two consecutive returns to state 1 is infinite. We never see robot 1 again since it is replaced. Therefore, state 1 is not positive recurrent in this case.

Consider a continuous Markov process, $\{X(t) : t \geq 0\}$, with state space, S. Suppose that $P_{ij} > 0$ for each $i, j \in S$. Restated, this means that for all $i, j \in S$, the states, i and j, are accessible to each other. Thus, all of the states in S communicate with each other, making the Markov process irreducible. Also, suppose that, starting from state i, the process will return to state i with probability 1, and the expected number of transitions before a first return to i is finite. This means that the Markov process is recurrent. *An irreducible continuous Markov process in which each state is positive recurrent is ergodic* [5].

To understand the reasoning behind this argument we will consider a continuous Markov process. For a continuous Markov process, we call the limiting probabilities for transitions into state j, denoted π_j, *stationary*

distributions. We want to show that $\pi_j = lim_{t\to\infty} P_{ij}(t)$. Suppose that the Markov process is reducible, and that there are two irreducible partitions of S, say s_1 and s_2. Let j be an element of s_2. Also, suppose that $lim_{t\to\infty} P_{ij}(t)$ converges to a limiting probability for an i in s_1. But since none of the elemetns of s_1 are accessible from s_2 and vice versa, $P_{ij}(t) = 0$ for all $t \geq 0$. This means that depending on whether the initial state i is in s_1 or s_2, the limiting probabilities could differ. The limiting probabilities would not be independent of state i, a necessary property for an ergodic Markov process. Therefore, the Markov process will not be ergodic. The Markov process must also be recurrent. Otherwise, if i is the initial state, $P_{ii}(t) = 0$ for all $t \geq 0$ and $lim_{t\to\infty} P_{ii}(t) = 0$. Again, the limiting probability depends on whether or not the initial state is i. Hence, the Markov process needs to be irreducible, and each of its states must be positive recurrent.

1.5.1 Base Frequencies

Our model for base substitution probabilities represent a continuous Markov process. We will assume that this Markov process is ergodic, so as t approaches ∞, the probability that the DNA site is in some state, j, is non-zero and independent of the starting state, i. In other words, there are positive values, $\pi_1, \pi_2, \pi_3,$ and π_4, such that, for all i and j in our state space $S = \{1, 2, 3, 4\}$,

$$lim_{t\to\infty} P_{ij}(t) = \pi_j.$$

Remember that we are representing DNA bases A, C, T, and G here with 1, 2, 3, and 4, respectively. Furthermore, for all $t \geq 0$, these values satisfy

$$\pi_j = \sum_{i \in S} \pi_i P_{ij}(t). \tag{1.3}$$

We call π_j the *base frequency* of base j because it represents the proportion of base substitutions that result in base j.

Chapter 2

The Likelihood of a Phylogenetic Tree

When calculating likelihood in Example 1, we did not know the probability with which the robot would raise its left or right arm. Similarly, we do not know the probability with which a base will mutate. Thus, calculating the likelihood of a tree is equivalent to finding the probability of the data we see at the tips of our tree given a hypothesis for the shape and bases of the nodes in the tree. In this chapter, we will focus on describing how to calculate the likelihood of a phylogenetic tree for a set of species.

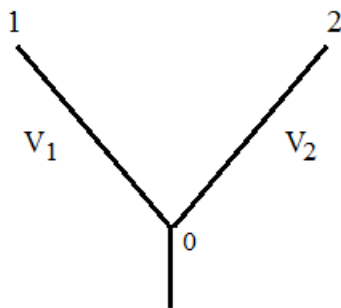## 2.1 Computing the Likelihood of a Phylogenetic Tree

Figure 2.1: 2-Species Phylogenetic Tree.

Let us first consider a tree with tips, 1 and 2, root, 0, and branch lengths,

v_1 and v_2, as in Figure 2.1. If the state, S_0, of node 0 was known, the likelihood of the tree would simply be the product of the probabilities of base substitution in each tree branch and the base frequency, π_{S_0}, of state, S_0:

$$L = \pi_{S_0} P_{S_0 S_1}(v_1) P_{S_0 S_2}(v_2)$$

Here, we are simply multiplying the independent events that would give us the tree we see in Figure 2.1, given the hypothesis that node 0 has state S_0. More specifically, these independent events are the event that the root of the tree is in state S_0, with probability π_{S_0}, and the event that the root bifurcates to give us species 1 and species 2 at the tips of our tree. The event that the root transitions from S_0 to the state of tip 1, S_1, in a branch length, v_1, has probability, $P_{S_0 S_1}(v_1)$. Similarly the event that the root transitions to the state of tip 2, S_2, in a branch length, v_2, has probability, $P_{S_0 S_2}(v_2)$.

Since, in reality, we will not know the *state* of the interior node, 0, we must sum the likelihoods for each possible state, S_0. The state, S_0, can take on the values 1, 2, 3, or 4, which correspond to A, C, T, and G, respectively. So, our likelihood calculation for the tree in Figure 2.1 will actually look like this:

$$L = \sum_{S_0} \pi_{S_0} P_{S_0 S_1}(v_1) P_{S_0 S_2}(v_2) = \sum_{1}^{4} \pi_{S_0} P_{S_0 S_1}(v_1) P_{S_0 S_2}(v_2).$$

The states S_1, S_2, v_1, and v_2, are given (from the tree), and π_i, for $i = \{1, 2, 3, 4\}$, is given as a fixed constant, independent of the tree. Notice that when we are calculating the likelihood of a phylogenetic tree, we are only able to determine the likelihood of the shape of the tree (i.e, the likelihood of the locations of the branches and nodes of the tree). Thus, we cannot determine the states of interior nodes simply by calculating the likelihood of the tree.

Example 8. If we know that tip 1 is an A and that tip 2 is a G in the tree in Figure 2.1, $S_1 = 1$ and $S_2 = 4$. Then, the likelihood of the tree is:

$$L = \sum_{S_0} \pi_{S_0} P_{S_0 1}(v_1) P_{S_0 4}(v_2).$$

Now, let us consider a slightly more complicated example in Figure 2.2. For this tree, the likelihood is the product of the sums of the transition probabilities for all possible states for the interior nodes and π_{S_0} (recall that
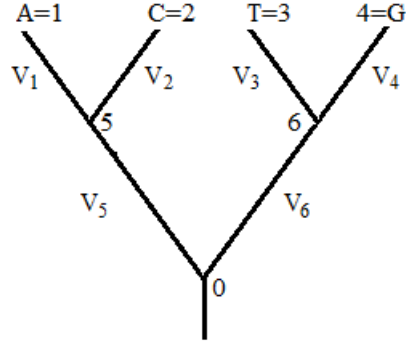
Figure 2.2: 4-Species Phylogenetic Tree.

states 1, 2, 3, and 4 correspond to DNA bases A, C, T, and G, respectively).
In this tree, we are given observable data at the four tips. Thus, our likeli-
hood caculation will look like this:

$$L = \sum_{S_0=1}^{4} \sum_{S_5=1}^{4} \sum_{S_6=1}^{4} \pi_{S_0} P_{S_0 S_5}(v_5) P_{S_5 1}(v_1) P_{S_5 2}(v_2) P_{S_0 S_6}(v_6) P_{S_6 3}(v_3) P_{S_6 4}(v_4)$$

(2.1)

This calculation sums over all four possible states for each node with an
unknown state.

Unfortunately, this calculation is extremely long. The likelihood calcu-
lation for Figure 2.1 has 4 terms, while the calculation for Figure 2.2 has 64
terms. Phylogenetic trees with more species have even more terms. It is
helpful to move each summation as far to the right in the likelihood cal-
culation as possible, allowing us to find the likelihoods of each individual
segment of the tree. If we do this with (2.1), our calculation would look like
this:

$$L = \sum_{S_0} \pi_{S_0} \sum_{S_5} [P_{S_0 S_5}(v_5)(P_{S_5 1}(v_1))(P_{S_5 2}(v_2))]$$

$$\times \sum_{S_6} [P_{S_0 S_6}(v_6)(P_{S_6 3}(v_3))(P_{S_6 4}(v_4))] \quad (2.2)$$

We have placed the parentheses and brackets in our equation [( )( )][( )(
)] so that they model the topology of our tree. Each segment of the tree, rep-
resented in our equation by the terms in the parentheses ( ), contains a base

substitution probability $P_{ij}(v)$. Within each set of parentheses, we have the likelihood for each branch connected to a tip. Within each bracket, we have the likelihood for each of the two branches that stem from the root, node 0. It should be clear that this relationship between the equation and the topology of the tree is a result of our construction of the likelihood calculation for the tree, but this construction also allows us to compute likelihood using the conditional likelihoods of each segment of the tree.

We will use the notation, $L_s^{(k)}$, for the likelihood of the tree at and above node k on the tree, given that node k has state s. Then, each $L_s^{(k)}$ corresponds to segments of the tree beginning with the tips of the tree. For example, in Figure 2.2,

$$L_S^{(5)} = \sum_{s_5=1}^{4} (P_{S_5 1}(v_1) L_1^{(1)})(P_{S_5 2}(v_2) L_2^{(2)}). \tag{2.3}$$

Since the tips of our tree contain our set of data, we know the states of the tips of the tree. Thus, if k is a tip of our phylogenetic tree, $L_s^{(k)}$ will be 0 for all states except for its observed state. If s_* is the observed state at tip k, then $L_{s_*}^{(k)} = 1$. Remember that this simply means that the probability of tip k having its observed state s_* is 1 given the hypothesis that tip k has probability s_* and 0 given any other hypothesis. For example, in Figure 2.2, $L_1^{(1)} = 1$, $L_2^{(2)} = 1$, $L_3^{(3)} = 1$, and $L_4^{(4)} = 1$. Now that we have an easy calculation for the tips of our tree we are able to begin our likelihood calculation for the entire tree in Figure 2.2 at its tips.

Since our likelihood calculation has been reduced to conditional likelihood calculations and we can easily find the likelihoods of the tips of the tree, we begin the computation from the tips of the tree and work our way down to the root. We can compute the conditional likelihoods of the nodes in the tree with tips as their immediate descendents. In Figure 2.2, nodes 5 and 6 satisfy this property.

Example 9. For node 6 in Figure 2.2, the likelihood that node 6 has state S is:

$$L_S^{(6)} = \sum_{S_6=1}^{4} \left[\left(\sum_{S_1=1}^{4} P_{S_6 S_3}(v_3) L_{S_1}^{(1)} \right) \left(\sum_{S_6=1}^{4} P_{S_6 S_4}(v_4) L_{S_4}^{(4)} \right) \right]$$

But we know that $L_1^{(1)} = 1$ and $L_2^{(2)} = 1$, so the calculation reduces to:

$$L_S^{(2)} = \sum_S [(P_{S_k 1}(v_1)(1))(P_{S_k 2}(v_2)(1))].$$

Now, we can state the process of finding conditional likelihoods as a general formula. For any node x, whose immediate descendants are tips y and z, we can compute $L_s^{(x)}$:

$$L_S^{(x)} = \sum_{S=1}^{4} \left[\left(\sum_{S_y=1}^{4} P_{S_x S_y}(v_y) L_{S_y}^{(y)} \right) \left(\sum_{S_z=1}^{4} P_{S_x S_z}(v_z) L_{S_z}^{(z)} \right) \right]$$

Once we have found the conditional likelihoods for all nodes with tips as immediate descendants, we can think of those nodes as our new "tips" by using their conditional likelihoods to compute the likelihoods of their ancestor nodes. Thus, for any node x with immediate descendants y and z, the conditional likelihood at node k is

$$L_S^{(x)} = \sum_s \left[\left(\sum_{S_y} P_{S_x S_y}(v_y) L_{S_y}^{(y)} \right) \left(\sum_{S_z} P_{S_x S_z}(v_z) L_{S_z}^{(z)} \right) \right]. \qquad (2.4)$$

We continue calculating conditional likelihoods, replacing our new "tips", and finding conditional likelihoods of ancestor nodes until we reach the root of our tree. At the root, node 0, of the tree, we will also compute the conditional likelihood $L_{S_0}^{(0)}$ given each possible state for root 0. Then, the overall likelihood of the tree for the DNA site we are considering is

$$L = \sum_{S_0} \pi_{S_0} L_{S_0}^{(0)}.$$

The base frequency for the root of our tree, π_{S_0}, gives us the probability that the root is in state S_0.

Example 10.

$$P_{ij} = \begin{pmatrix} 0.8 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.8 \end{pmatrix}$$

Suppose the terms in the matrix above represent the base substitution probabilities of a phylogenetic tree with two species that are both one branch length unit away from a common ancestor. As we have seen before, i will represent a prior state, and j will represent a later state for the DNA site under consideration. Both i and j can take on values 1, 2, 3, and 4 coresponding to DNA bases, A, C, T, and G, respectively. Suppose we are studying orcs and trolls in our set of data. For the DNA site that we are examining, orcs have an A, and trolls have a T. Then, the likelihood of seeing our

observed data, or in other words, the probability of seeing our observed data given a hypothesis for the base seen at the DNA site in the common ancestor, will be as follows:

$$L_S^{(0)} = P_{S1}P_{S3}$$

where 0 is the node in our tree representing the common ancestor and S is the state of node 0. We take the base substitution probabilities from our matrix and see that $L_1^{(0)} = 0.08$, $L_2^{(0)} = 0.01$, $L_3^{(0)} = 0.08$, and $L_4^{(0)} = 0.01$. Now, say that the common ancestor is the root of our tree. Then, we simply need to find the sum of the products of these conditional likelihoods and the base frequencies for each hypothesized ancestor base. If the base frequencies are 0.25 for each base, then

$$L = \sum_{S=1}^{4} \pi_s L_s^{(0)} = \pi_1 L_1^{(0)} + \pi_2 L_2^{(0)} + \pi_3 L_3^{(0)} + \pi_4 L_4^{(0)} = 0.045$$

In order to get a tree with a greater likelihood, we could find a branch length that would provide us with greater base substitution probabilities, increasing the likelihood of the subsequent tree. We will see how to do this in Chapter 3.

2.2 Time Reversibility

While computing the likelihoods of the examples in the previous section, we actually made quite a few assumptions. For instance, in Example 5, we assumed that the root of the tree was at node 0 in order to compute the likelihood of the entire tree. In reality, when given a set of DNA sequences, we oftentimes will not know where the root of the tree lays. Another assumption we made was that the DNA bases at the tips of our tree were exactly one branch length away from their common ancestor. Again, in actual families of species, we will have to search for this information. One property that helps us find the actual branch lengths and location of the root of our tree is *time reversibility*.

Definition 14. A Markov chain is *time reversible* if the rate at which it goes from state i directly to state j is the same as the rate at which it goes from state j directly to state i.

Recall that the base frequency of base i is π_i, and this term represents the proportion of base substitutions that will result in base i. We can restate

our definition of time reversibility, using our model of base substitution, as $\pi_i P_{ij}(t) = P_{ji}(t)\pi_j$ for all i, j, and t. In order to determine whether or not our Markov process is time reversible, we have to show that these two products are equivalent. This is made clear by utilizing (1.2). By (1.2),

$$\pi_i P_{ij}(t) = \pi_i e^{-\mu t}\delta_{ij} + (1 - e^{-\mu t})\pi_i \pi_j$$

We also see that,

$$\pi_j P_{ji}(t) = \pi_j e^{-\mu t}\delta_{ji} + (1 - e^{-\mu t})\pi_i \pi_j$$

The only difference between these two equations is the δ_{ij} and δ_{ji}. If we recall the origin of these terms in (1.2), the Kronecker delta function, we know that the values of δ_{ij} and δ_{ji} are either both 1 or 0 depending on whether $i = j$ or $i \neq j$, respectively. Then, $\delta_{ij} = \delta_{ji}$ for all i and j. Hence, we have that the Markov process modeling base substitution is in fact time reversible.

But where exactly does (1.2) depend on t? The probability of a base substitution from base i to base j, is only dependent on the product μt. Suppose we divided μ in half and doubled t. Then, we would have $\frac{\mu}{2}2t = \mu t$. Next, suppose we divided μ by 3 and tripled t. We would be left with $\frac{\mu}{3}3t = \mu t$. We see that if we divide the rate μ by an amount and multiply the time t by the same amount, we get the same product μt. This makes it difficult for us to determine the values of μ and t separately in (1.2). All we are able to take away from (1.2) is the product, μt.

If we assume that the rate of base substitution, μ, is the same for all branches of the tree and all times, then μt should be proportional to the time that has passed so far during the Markov process, where time is represented in the number of mutations that have occurred. Since we do not have any indication of what μ or t could be, we will assume that $\mu = 1$. This allows us to think of t as units of expected numbers of substitutions.

2.3   The Pulley Principle

The time-reversibility and ambiguity of branch lengths turn out to be quite useful in finding the best phylogenetic tree for a set of DNA sequences. We can actually place the root of our tree anywhere in our tree. We will look at our calculation of the likelihood of the tree seen in Figure 2.2 to see why. The last two steps of our algorithm for computing the likelihood of the tree involved nodes 0, 5, and 6. In the last step of the algorithm, we had

$$L = \sum_{s_0=1}^{4} \pi_{s_0}((P_{s_0 s_5}(v_5)L_{s_5}^{(5)})(P_{s_0 s_6}(v_6)L_{s_6}^{(6)})). \tag{2.5}$$

What would happen to the likelihood of the tree, L, if we added a length x to v_5 and subtracted x from v_6? We will try rewriting (2.5) using what we know about time reversibility to find out. With the time reversibility property, we have

$$\pi_{s_0} P_{s_0 s_5}(v_5) = \pi_{s_5} P_{s_5 s_0}(v_5)$$

which allows us to rewrite (2.5) as

$$L = \pi_{s_5} L_{s_5} \sum_{s_0=1}^{4} P_{s_5 s_0}(v_5) P_{s_0 s_6}(v_6). \tag{2.6}$$

This form of our equation allows us to experiment with different branch lengths. In order to do this, we must first introduce the *Chapman-Kolmogorov equation*. We will consider a continuous Markov process $\{X(t) : t \geq 0\}$ with state space S and transition probability matrix P. If the process begins in state i and moves to state j in t units of time, then we have

$$P_{ij}(t) = P(X(s + t) = j \mid X(s) = i), \quad \text{for } i, j \in S; s, t \geq 0.$$

Since $X(0) = i$, it should be clear that after s units of time, the Markov process will enter another state, we will call it k, before entering state j in $s + t$ units of time. This gives us the *Chapman Kolmogorov* equation:

$$P_{ij}(s + t) = \sum_{k=0}^{\infty} P_{ik}(s) P_{kj}(t). \tag{2.7}$$

We will prove (2.7) using the Law of Total Probability. Since $\{X(s) = k \mid k \in S\}$ is a set of mutually exclusive events, we can apply the Law of Total Probability to get:

$$P_{ij}(s+t) = P(X(s+t) = j \mid X(0) = i)$$

$$= \sum_{k=0}^{\infty} P(X(s+t) = j \mid X(0) = i, X(s) = k)P(X(s) = k \mid X(0) = i)$$

$$= \sum_{k=0}^{\infty} P(X(s+t) = j \mid X(s) = k)P(X(s) = k \mid X(0) = i)$$

$$= \sum_{k=0}^{\infty} P_{kj}(t)P_{ik}(s) = \sum_{k=0}^{\infty} P_{ik}(s)P_{kj}(t)$$

Now that we have the Chapman-Kolmogorov Equation, we can rewrite the summation $\sum_{s_0=1}^{4} P_{s_5 s_0}(v_5)P_{s_0 s_6}(v_6)$ from (2.6) as $P_{s_5 s_6}(v_5 + v_6)$. This means that the likelihood of our tree is only dependent on the total lenth of the tree branches; it only depends on the sum, v_5+v_6. This makes it possible for us to increase and decrease the lengths of v_5 and v_6 as long as the sum, $v_5 + v_6$, remains constant. Thus, the root of the tree can be placed anywhere between node 5 and node 6. Since these nodes were chosen arbitrarily, we are able to think of each branch of our tree in a similar manner and place the root of the tree anywhere in the tree. Since Section 2.1 only described how to find the likelihood of a rooted tree, how do we find the likelihood of an unrooted tree? By the Pulley Principle, we can root the tree anywhere and get the same likelihood. Figure 2.3 shows three trees that are equivalent because of the pulley principle.
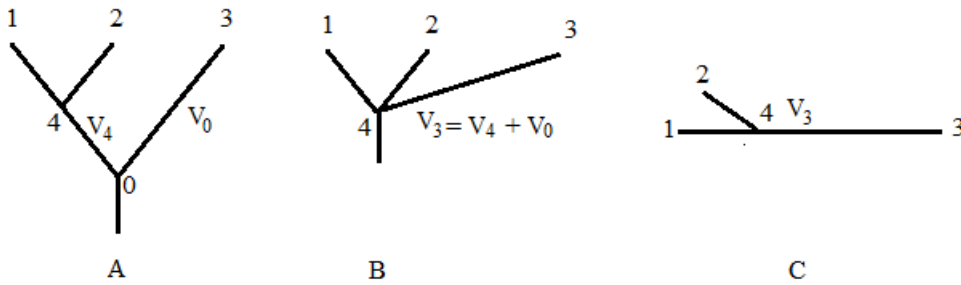

Figure 2.3: Tree A, Tree B, and Tree C are all equivalent due to the pulley principle. Tree A shows the tree rooted at node 0. Tree B demonstrates how the tree root can be moved. Tree C shows the unrooted tree.

Chapter 3

Finding a Maximum Likelihood Tree

The Pulley Principle forces us to think of each segment of the phylogenetic tree as a possible location for the root of the tree. This actually becomes a very useful property of the tree because it allows us to manipulate or change branch lengths and the location of the root of the tree. This is helpful because it simplifies the process of finding the tree with the maximum likelihood out of the set of trees that we test for a set of DNA sequences. In Chapter 2, we described a long but still practical process for finding the likelihood of a phylogenetic tree. However, in order to find the tree that produces the largest likelihood out of all of the trees for which we calculate a likelihood, we still need to experiment with possible topologies and branch lengths and then, evaluate the likelihoods of the trees they produce.

   After we have found all possible topologies for a particular site in a set of DNA sequences, we can evaluate the likelihood of each topology. As described in Chapter 2, in each computation for the likelihood of a phylogenetic tree, we get the base substitution probability, $P_{ij}(v_j)$, of a site in state i transitioning to state j in v_j time units by finding the transition probabilities specified for a time of length v_j. But so far we have only found possible tree topolgies; we do not know what v_j is. Furthermore, direct search would not be a feasible method of finding that v_j because it would require us to evaluate the likelihood of the tree under consideration for each distinct v_j. The Pulley Principle gives us an algorithm which allows us to change one branch length, v_j, at a time (see section 3.2). Each branch length is altered to the value that provides the highest likelihood out of all of the other branch lengths tested. Once we find each branch length, v_j, we can

complete the likelihood calculations for each possible topology. Finally, the topology producing the greatest likelihood out of all of the other topologies we examine is the tree that we choose as the best phylogenetic tree for our set of data.

3.1   How Many Possible n-species Trees are There?

Now that we have a general idea of how to find a tree with a good likelihood, we still need to find a feasible way to find a good tree topology without examining every possible tree topology. For the purposes of this paper, we will be focusing on finding the topology of bifurcating trees. Although it is possible to find every tree topology available for a set of DNA sequences, there would be too many possible tree topologies to test since we would have to compute the likelihood for each topology. This direct search strategy would require us to look at each possible unrooted tree topology, iterate each branch length to its optimal value for each branch, and then choose the topology that has given us the greatest likelihood. However, by looking at even a simple tree, we see that this method is not ideal. We will later see that the number of unrooted bifurcating trees with n labelled tips is $(2n - 5)!/[(n - 3)!(2^{n-3})]$ (see section 3.1.1). For a tree with 10 labelled tips, for example, there are more than 2 million such topologies to evaluate. This is much too big a number of trees to evaluate the likelihood for.

Therefore, we need to use a more focused strategy to search among possible tree topologies. Felsenstein [2] suggests building the tree starting with a 2-species tree and successively adding one species at a time to the tree until all of the species are on the tree. One difficulty with this method is determining where to place each species as it is added to the tree. I will describe the process for deciding where to place a species on both a rooted bifurcating phylogenetic tree and an unrooted bifurcating phylogenetic tree. For both types of trees, we will add one species at a time in some predetermined order (e.g. alphabetical order by species name).

3.1.1   Rooted Bifurcating Trees

In Chapter 1, both Figure 1.1 and Figure 2.1 show examples of rooted bifurcating trees. Each of the phylogenetic trees has a root at the bottom of the tree which can be traced back to from any node and tip, and when a base undergoes a mutation, there are two possible bases that are substituted for the original base. Figure 3.1 also shows some rooted bifurcating trees. How

can we determine whether or not Figure 3.1 contains all possible labeled, rooted bifurcating trees for a tree with three tips?
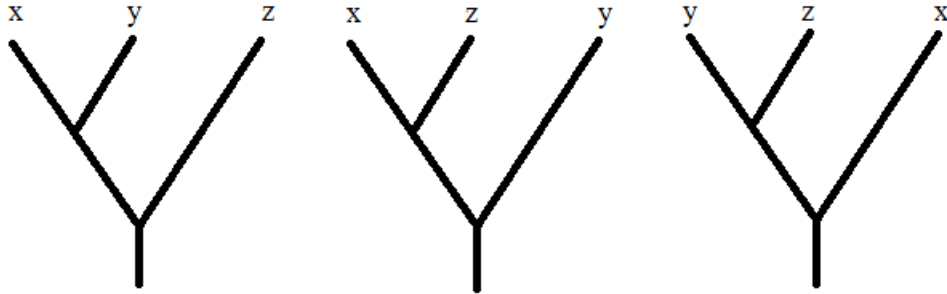

Figure 3.1: All possible labeled, rooted bifurcating trees for three species.

We use the following straightforward reasoning. Consider constructing one possible phylogenetic tree by beginning with a tree with two of the n species and then adding one species at a time to that tree in some predetermined order (e.g. species name in alphabetical order). It is clear that if we add one more species to a tree that already has n species added to it, we only need to place the $(n + 1)$-st species in each possible location in order to get the total number of trees for an $n + 1$-species tree. Since the tree is bifurcating before and after each node, and hence before and after the location of the addition, we are unable to add the new species to a node. Instead, the new species must be added to an existing branch, resulting in the creation of a new node and two new branches. This implies that each existing branch in the n-species tree is a possible location for the addition of the new species. Our reasoning for this is depicted in Figure 3.2.

How can we be sure that this process will give us all possible labeled rooted, bifurcating trees? With the information we have so far, we also cannot be completely sure whether or not adding a new species to two different branches will produce two different trees. In order to understand this further, consider the process of adding species k to a tree with $k - 1$ species and also the process of removing species k from a k-species tree. Suppose we have an n-species tree. If we remove $n - k$ species in the reverse order of which they were added to the tree, we will be left with a k-species tree. This k-species tree will actually be the tree that we should have obtained before when we were adding species k to the tree. This indicates that there exists a particular sequence of places to add species $k + 1, k + 2, k + 3, \ldots$ onto the k-species tree in order to return to the n-species tree we had before
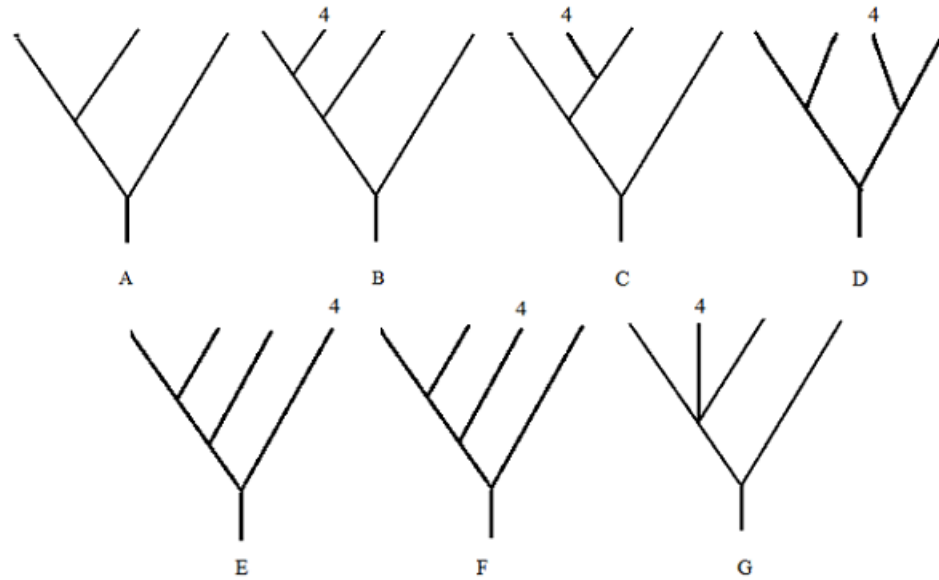
Figure 3.2: Tree A shows a 3-species tree. When we want to add species 4 to the tree, we must add it to a new node on an existing branch in the 3-species tree as shown in trees B-F. If we decide to add species 4 to an existing node, as we have done in Tree G, then our tree will no longer be a rooted bifurcating tree.

we began removing species. In addition, no other k-species tree can turn into that same n-species tree after adding the $n - k$ species that are missing. Suppose there was another k-species tree that could produce the same n-species tree. Then, this k-species tree should also be produced through the removal of the $n - k$ species that were added after it. However, this presents a contradiction, since the same sequence of removals cannot yield two different trees. Therefore, any n-species tree can be produced from one and only one k-species tree [3].

This result implies that each possible addition sequence leads to a different n-species tree. Since everytime we add a new species to a tree, it can only be added to one of the existing branches, the number of ways in which we can add that species to the tree is equal to the number of branches in the existing tree. This includes the branch connected to the root of the tree.

Example 11. In Figure 3.2, tree A has 5 branches, so species 4 can be added in 5 ways. Notice that when we add species 4 to tree A, as shown in tree B, there are 2 new branches and 1 new interior node. Then, when we add

species 5 to tree B, there will be 7 ways that we can add it.

Using this reasoning, we see that there are $3 \times 5 \times 7 \times 9 \times 11 \times \cdots \times (2n-3)$ different ways to add species to a rooted bifurcating phylogenetic tree in order to produce an n-species tree. The product, $3 \times 5 \times 7 \times 9 \times 11 \times \cdots \times (2n-3)$, looks similar, but not equal, to $(2n-3)!$. The even divisors have simply been removed from $(2n-3)!$. In order to correct for this, we take out the even divisors by dividing by $2 \times 2^2 \times \cdots \times 2^{n-1} = 2^{n-1}(n-1)!$. This shows that $3 \times 5 \times 7 \times 9 \times 11 \times \cdots \times (2n-3) = \frac{(2n-3)!}{2^{n-1}(n-1)!}$

Now, we are able to check that at each addition of a new species to a phylogenetic tree, we have added the new species to each possible location. For example, if we had 20 species ($n = 20$), the total possible number of trees would be:

$$\frac{(2n-3)!}{2^{n-1}(n-1)!} = 8,200,794,532,637,891,559,375.$$

This means that if a computer was able to evaluate the likelihood of a 20-species tree in one hour, it would take us about 1.64×10^{23} hours, or about 1.87×10^{19} years. Remember that this is only for the calculation of the likelihood of each tree, without testing various branch lengths.

Obviously, there are way too many total possible trees to examine all of them when we have a tree with more than ten species. We must also keep in mind that we have to examine all possible trees for each site of the DNA sequences, making this process even less feasible. Thus, we must use a different algorithm for examining as many possible tree topologies as possible.

3.1.2   Unrooted Bifurcating Trees

As we saw in section 2.2, we do not actually know where the root of our tree is. Therefore, instead of considering rooted, bifurcating trees as we did in the previous section, we must consider unrooted, bifurcating trees. Then, how many possible unrooted, bifurcating trees would we have to consider?

Well, if we root the tree at one of its species, we will have a rooted, bifurcating tree. Suppose we had n tips on an unrooted tree. Then, after rooting our tree at one of those n species, we would end up with $n-1$ tips. This resulting tree would be a rooted, bifurcating tree with $n-1$ tips. Thus, the number of possible trees would be:

$$3 \times 5 \times 7 \times \cdots \times (2(n-1)-3) = 1 \times 3 \times 5 \times 7 \times \cdots \times (2n-5)$$

This equation gives us the same number of possible rooted, bifurcating trees with $n - 1$ tips. We see that every rooted tree with $n - 1$ labeled tips is analogous to an unrooted, bifurcating tree with n tips. Likewise, every unrooted tree with n tips is analagous to one rooted tree with $n - 1$ tips. Figure 3.3 shows an example of rooting an unrooted, bifurcating tree at one of its species.
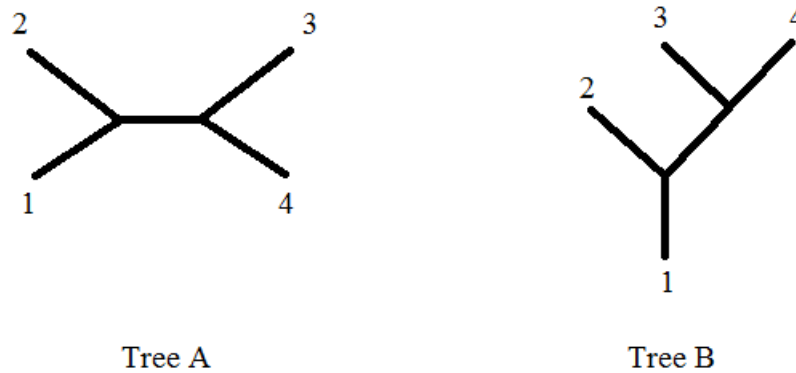

Tree A   Tree B

Figure 3.3: Tree A shows an labeled unrooted, bifurcating tree. We can root Tree A at species 1, resulting in Tree B, a rooted, bifurcating tree.

3.2   Which Tree Topology?

Even though there are fewer possible unrooted, bifurcating trees for n species than there would be for rooted, bifurcating trees with n species, it is not by much. Clearly, we must introduce a method that is not as computationally expensive. Felsenstein [2] has come up with a method that is much more simple. When the k-th species is being added to the tree, there will be $2k - 5$ different branches to which it could be added. Each of these different locations is tried, and the likelihood of each resulting topology is evaluated. The topology that produces the largest likelihood out of all of the other likelihoods computed is accepted, and the rest are thrown out.

In order to ensure that the order in which we add each species is not limiting the number of topologies we produce using this strategy, we use local rearrangements of the tree (described in more detail in the following section). These local rearrangements will not lead us to every topology, for that would give us the same problem presented in 3.1.1 and 3.1.2: there would be way too many trees to examine. Instead, local rearrangements

give us a *greedy* algorithm. In a greedy algorithm, we decide when the algorithm has produced desirable enough results to stop performing the algorithm. We determine how "greedy" we want to be in assuming that when the results are sufficient. With local rearrangements, we are simply checking to see if those topologies that are similar to the topology we begin with can result in a higher likelihood. As the likelihoods of the successive trees increase, we can determine when to stop performing local rearrangements.

If the tree has more than four species, before we add the next species, local rearrangements are done to see if any of the resulting topologies increases the likelihood of the tree. If a topology resulting from these local rearrangements does in fact increase the tree's likelihood, it is accepted. These local rearrangements continue until a tree is found for which no local rearrangement can increase the likelihood significantly.

3.2.1 Local Rearrangements

Since the order in which we add a new species to a tree will affect the maximum likelihood tree we find, we use local rearrangements in an attempt to correct for this. This means that if we use a different sequence of adding species to our tree, we may end up with a different maximum likelihood tree.

Remark 1. The tree with the likelihood that we ultimately decide on after performing local rearrangements will only provide a local maximum likelihood. This means that there may be a possible topology with an even greater likelihood that we have not considered because its topology is so different from the original topology that we began with. We cannot be guaranteed that we have found the tree with the greatest likelihood of all.

There are several different kinds of local rearrangements that we can use. Since our goal is simply to understand how these local rearrangements can guide us to a tree with a local maximum likelihood, we will use a simple local rearrangement process called Nearest Neighbor Interchange (NNI). This process basically involves switching adjacent branches with each other. More specifically, we are switching adjacent subtrees connected to an interior branch. We erase the interior branch and the branches connected to it at each end. For example, in an unrooted, bifurcating tree, we erase a total of five branches. The four adjacent trees are then disconnected from each other. They can then be reconnected into a tree in three possible ways. These are the only three possible ways to reconnect the tree without

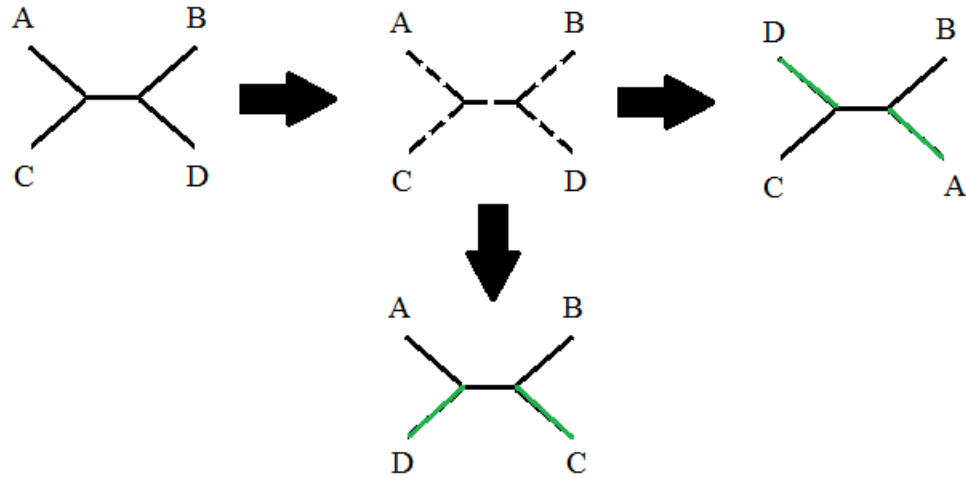repeating a topology. These rearrangements are depicted in Figure 3.4.

Figure 3.4: *Nearest Neighbor Interchange.* Here we see our original tree with subtrees A, B, C, and D. For the particular interior branch that we are concerned with, we erase that particular branch along with all of the branches directly connected to it. Then, we reassemble the subbranches in the two different ways shown above. This gives us a total of three possible ways to construct our tree, including the tree before local rearrangements.

Once we have completed this local rearrangement at one of our interior branches, we evaluate the likelihood of each of the three resulting trees. We accept the tree that gives us the highest likelihood and throw out the other two possiblities. We then continue using NNI on all of the other interior branches and accept those rearrangements that give our tree a higher likelihood. We continue this process until no more local rearrangements can increase the likelihood of the tree. Once we have reached this point, we add the next species on to our tree.

Before we go any further, we can't forget that we need branch lengths in order to calculate the likelihood of a phylogenetic tree. The term, $P_{S_k S_i}(v_i)$ in the general equation for the likelihood calculation (2.4) is a transition probability specific to the branch length, v_m. But we do not yet know the branch lengths that make up our phylogenetic tree. This means that before we can actually evaluate any of the likelihoods we have discussed above, we need to figure out each branch length.

3.3   Maximum Likelihood Branch Lengths

For each of the topologies we will consider when looking for the maximum likelihood tree topology, we have to find the branch lengths that will maximize the likelihoods of those topologies. Felsenstein [2] created an iteration technique that simplifies this at first sight, daunting task.

We will consider Tree C in Figure 2.3 in order to understand this strategy better. This is the unrooted phylogenetic tree for these three species, and we do not know where the root of the tree is located. Consider branch v_3. Suppose the root is located somewhere along v_3. Then, we can use the method presented in Chapter 2 to compute the likelihood of the tree. We will assume that the root is directly to the right of node 4. Then, the likelihood of the tree for one site of DNA is:

$$L = \sum_{S_0} \sum_{S_4} \sum_{S_3} \pi_{S_0} (P_{S_0 S_3}(0) L_{S_4}^{(4)})(P_{S_0 S_3}(v_3) L_{S_3}^{(3)}) \tag{3.1}$$

$$= \sum_{S_0} \pi_{S_0} L_{S_0}^{(4)} (\sum_{S_3} P_{S_0 S_3}(v_3) L_{S_3}^{(3)}) \tag{3.2}$$

The equality in (3.2) comes from the fact that $P_{S_0 S_4}(0) = 1$. Now, if we substitute (1.2) into (3.2), letting $\mu = 1$, we get:

$$L = e^{-v_3} \sum_S \pi_S L_S^{(4)} L_S^{(3)} + (1 - e^{-v_3}) \left[\sum_{S_4} \pi_{S_4} L_{S_4}^{(4)} \right] \left[\sum_{S_3} \pi_{S_3} L_{S_3}^{(3)} \right] \tag{3.3}$$

Now, we must figure out how to find $L_{S_4}^{(4)}$ and $L_{S_3}^{(3)}$ without knowing the branch lengths in the tree. Recall that in Section 2.1, for a tip, k, of our tree with base, i, $L_i^{(k)} = 1$, while $L_{j \neq i}^{(k)} = 0$. Since species 1, 2, and 3 are all located at the tips of our tree, they have a known conditional likelihood, $L_{S_k}^{(k)}$. In the 2-species tree that we added species 3 to in order to get Tree C, we should have evaluated branch lengts, v_1 and v_2. We can then use these branch lengths and the known conditional likelihoods of the tips of our tree to compute $L_{S_4}^{(4)}$.

Recall that this calculation for likelihood only gives us the likelihood for the phylogenetic tree representing one DNA site. For all K DNA sites in the sequences of our data set, the likelihood is:

$$\prod_{i=1}^{K} (A_i q + B_i p) \tag{3.4}$$

where $q = e^{-v_3}$ and $p = 1 - q = 1 - e^{-v_3}$. A_i and B_i represent the following:

$$A_i = \sum_S \pi_S L_S^{(4)} L_S^{(3)}$$

and

$$B_i = \left(\sum_{S_4} \pi_{S_4} L_{S_4}^{(4)} \right) \left(\sum_{S_3} \pi_{S_3} L_{S_3}^{(3)} \right)$$

for the i-th DNA site in the set of sequences. We are interested in finding the value of v_3 that will maximize the likelihood. We can find this value of v_3 by finding the value of p that will maximize (3.3). Then, we can solve for v_3 since $v_3 = -\ln(1 - p)$.

Now, we will explore some properties that follow from these equations that will help us understand Felsenstein's iteration formula for finding segment lengths that will maximize the likelihood. If we take the logarithm of (3.3), we get

$$\ln(L) = \sum_{i=1}^{K} \ln(A_i q + B_i p).$$

If we take the derivative of this equation and set it equal to zero, we get

$$\frac{d\ln(L)}{dp} = \sum_{i=1}^{K} \frac{B_i - A_i}{(A_i q + B_i p)} = 0. \tag{3.5}$$

Now, notice that if we have K sites total,

$$K = \sum_{i=1}^{K} 1 = \sum_{i=1}^{K} \frac{A_i q + B_i p}{A_i q + B_i p} = \sum_{i=1}^{K} \frac{B_i - (B_i - A)q}{A_i q + B_i p}. \tag{3.6}$$

We can use (3.5) to take out the terms in the numerator of (3.6) containing q. We end up with

$$\sum_{i=1}^{K} \frac{B_i - (B_i - A)q}{A_i q + B_i p} = \sum_{i=1}^{K} \frac{B_i}{A_i q + B_i p} - q \sum_{i=1}^{K} \frac{B_i - A_i}{(A_i q + B_i p)}.$$

The last term is 0 because of (3.5). Then,

$$K = \sum_{i=1}^{K} \frac{B_i}{A_i q + B_i p}. \tag{3.7}$$

The equation presented in (3.7) must be satisfied at a maximum in the likelihood function for the tree, or equivalently, when $d \ln(L)/dp = 0$. Now, by multiplying both sides of (3.7) by p, we get the iteration-formula:

$$p^{(k+1)} = \frac{1}{K} \sum_{i=1}^{K} \frac{B_i p^{(k)}}{A_i q^{(k)} + B_i p^{(k)}} \tag{3.8}$$

where $q^{(k)} = 1 - p^{(k)}$. This iteration expression is a specific case of the general EM algorithm presented by Dempster, et al [1]. The first step of the algorithm is to make an estimated for the value of p, calling it $p^{(1)}$. This is simply an educated guess. This value of $p^{(1)}$ generates another estimate for p, namely $p^{(2)}$. This value of $p^{(2)}$ is then used to find $p^{(3)}$, and this process continues until the estimates for p converge to some $p^{(k)}$. It can be shown that the successive segment lengths produced never decrease the likelihood of the tree [1]. Thus, we can iterate this process until the $p^{(k)}$s converges. This iteration is another greedy algorithm because it is only ensured to lead to a branch length that produces a local maximum likelihood. Once the $p^{(k)}$ converges, we may continue the iteration to see whether or not it will lead us to an even larger likelihood, or we may choose the $p^{(k)}$ we already have.

It is important to observe that summing over many (K) DNA sites is significant here. Let us consider what happens when $K = 1$. In this case, the iteration equation will be:

$$p^{(k+1)} = \frac{B p^{(k)}}{A q^{(k)} + B p^{(k)}}.$$

Now, we will substitute an x for $p^{(k)}$. Then, we have a function in terms of x:

$$f(x) = \frac{Bx}{A(1-x) + Bx} = \frac{Bx}{(B-A)x + A}.$$

Notice that this equation is in the form of a fractional linear transformation such that it has only two fixed points. These two fixed points are $x = p = 0$ and $x = p = 1$. Then, the iteration will converge to either $p = 1$ or $p = 0$. Hence, if we use this iteration algorithm for the single site, we will get that the branch length we are estimating is either 0 or undefined since $p = 1 - e^{-v}$, for branch length, v. If $p = 0 = 1 - e^{-v}$, then $v = 0$. If $p = 1 = 1 - e^{-v}$, then $v = \infty$. Therefore, in practice, we only use the EM algorithm with DNA sequences with two or more sites.

We must do this iteration technique for each of the branches on each topology. Once we iterate and optimize for v_x, we fix this value of v_x and

move on to optimize for another branch length, v_y. We continue optimizing branch lengths in this fashion until the branch lengths each converge to some value.

3.4   Putting Topologies and Branch Lengths Together to find the Maximum Likelihood Tree

Now that we have a method for finding all of the branch lengths that will maximize each topology's likelihood, we can use the resulting branch lengths to actually compute the likelihoods of each topology we are interested in. Once we compute each topology's likelihood, we follow the procedure described in 3.2. We will accept those topologies that increase the likelihood of the tree until the topology cannot be altered to increase the likelihood of the tree. This will give us a sequence of trees with increasingly better likelihoods. Remember that these topologies will all be unrooted bifurcating trees. The Pulley Principle allows us to compute the likelihoods of unrooted trees. The tree that results in the highest likelihood is accepted. Once we find a phylogenetic tree whose topology and branch lengths cannot be altered to significantly increase the likelihood of the tree, we choose that topology as our maximum likelihood tree. Figure 3.5 illustrates this entire process for us.
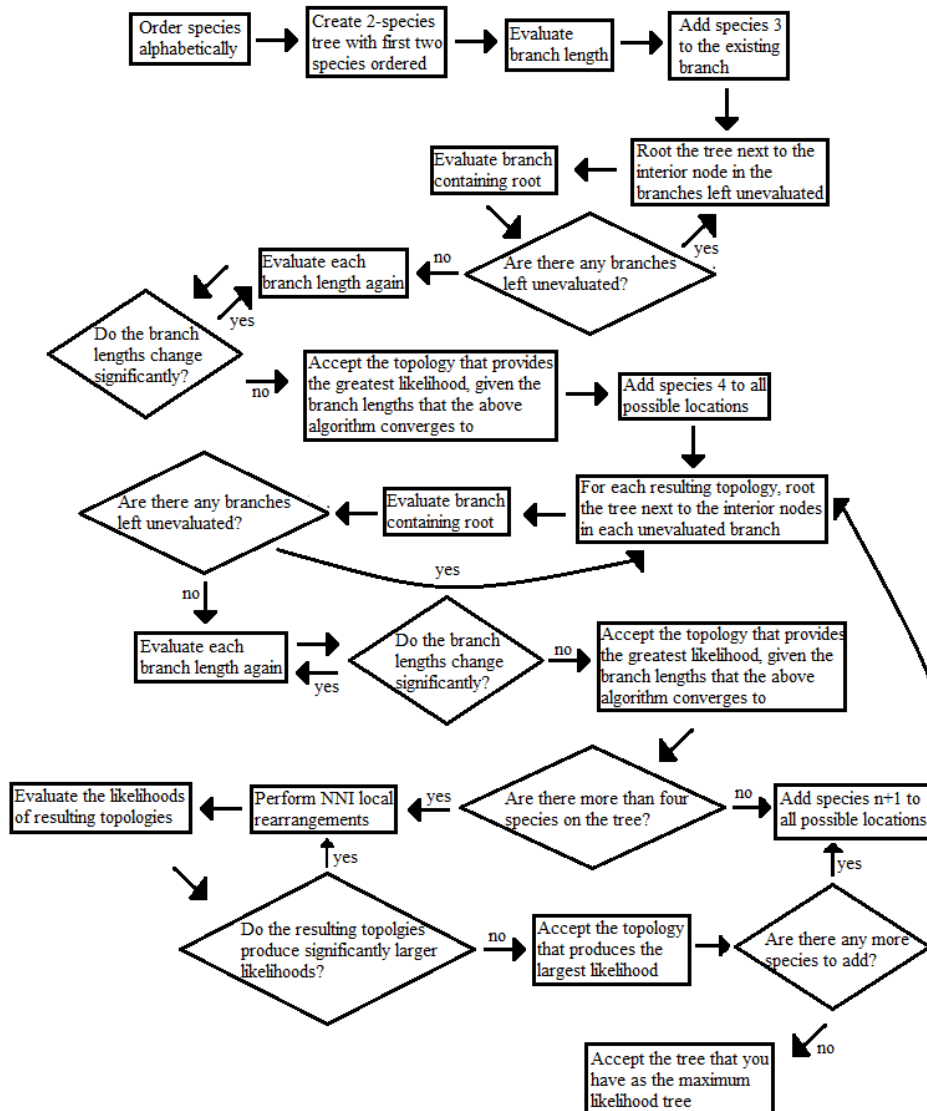
Figure 3.5: Finding an estimate for the maximum likelihood tree.

Chapter 4

Finding the Maximum Likelihood Tree - An Example

In this chapter, we will demonstrate the process of finding a maximum likelihood phylogenetic tree using a relatively simple example. We will only consider two DNA sites for a family of five species, namely the wizards, elves, hobbits, dwarves, and humans from Middle Earth. At site 1, we set the DNA data to be A, C, T, G, and T, respectively in our five species. At site 2, let us set our DNA data to be A, A, C, G, and G. In other words, the first species has base A at site 1 and base A at site 2. The following table illustrates which base each species has at each DNA site.

| DNA Site	Wizard	Elf	Hobbit	Dwarf	Human
1	A	C	T	G	T
2	A	A	C	G	G

We will assume that we can model base substitution probabilities using the Jukes-Cantor model (see [7]). Under this model, base frequencies are all equal. In other words, $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1/4$.

First, we will focus on site 1 of the DNA sequences. We will add each base to the tree in alphabetical order (A, C, G, T, T). Our 2-species tree will look like this:

$$1 = A \quad \rule{3cm}{1pt} \quad C = 2$$

Now, since there is only one branch in our tree, there is only one possible location for us to add species 3. Then, our 3-species tree will look like

this:
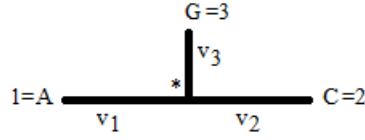
In order to begin the process of finding branch lengths, suppose the root of the tree lays along v_1. Suppose it lays directly to the left of the node labelled with *. Then, as in Section 3.3, the likelihood of the tree is:

$$L = \sum_{S_0} \sum_{S_*} \pi_{S_0} (P_{S_0 S_*}(0) L_{S_*}^{(*)})(P_{S_0 1}(v_1)).$$

Since $P_{S_0 S_*}(0) = 1$, this is equivalent to:

$$L = \sum_{S_0} \pi_{S_0} L_{S_0}^{(*)} (P_{S_0 1}(v_1)).$$

Now, we can substitute (1.2) for $P_{S_0 1}(v_1)$ to get:

$$L = \pi_1 L_1^{(*)}(e^{-v_1} + (1 - e^{-v_1})\pi_1) + \sum_{S_0 \neq 1} \pi_{S_0} L_{S_0}^{(*)}(1 - e^{-v_1})\pi_1 \qquad (4.1)$$

$$= e^{-v_1} \pi_1 L_1^{(*)} + (1 - e^{-v_1}) \sum_{S_0} \pi_{S_0} L_{S_0}^{(*)} \pi_1. \qquad (4.2)$$

Our equation is now in the same form as (3.3), so we can label our terms as we did in (3.7) for site 1. In this example, term $A_1 = \pi_1 L_1^{(*)}$, term $B_1 = \sum_{S_0} \pi_{S_0} L_{S_0}^{(*)} \pi_1$, term $q = e^{-v_1}$, and term $p = 1 - q = 1 - e^{-v_1}$.

Next, we consider DNA site 2, so that we can find A_2 and B_2. If we order the bases in site 2 in the same order as we ordered the bases in site 1, we have A, A, G, C, and G. The phylogenetic tree for site 2 will look like this:
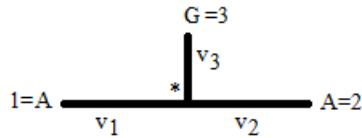
Then, the likelihood for the phylogenetic tree for site 2 is:

$$L = e^{-v_1}\pi_1 L_1^{(*)} + (1 - e^{-v_1})\sum_{S_0}\pi_{S_0}L_{S_0}^{(*)}\pi_1.$$

In this equation, $A_2 = \pi_1 L_1^{(*)}$ and $B_2 = \sum_{S_0}\pi_{S_0}L_{S_0}^{(*)}\pi_1$.

Since the number of sites in the DNA sequences of our data is $K = 2$, our iteration formula will simply be:

$$p^{(k+1)} = \frac{1}{2}\sum_{i=1}^{2}\frac{B_i p^{(k)}}{A_i q^{(k)} + B_i p^{(k)}} \tag{4.3}$$

In order to use this iteration formula, we must first determine what B_i and A_i are for $i = 1$ and for $i = 2$. For each A_i, we know that $\pi_1 = 1/4$ under the Jukes-Cantor model. Then, we are left with finding $L_1^{(*)}$. Similarly, for each B_i, we know that $\pi_{S_0} = \pi_1 = 1/4$. Thus, we are left with the task of finding $L_{S_0}^{(*)}$. By (2.4), we have that

$$L_1^{(*)} = (P_{13}(v_3))(P_{12}(v_2))$$

and

$$L_{S_0}^{(*)} = \sum_{S_0}(P_{S_04}(v_3))(P_{S_02}(v_2))$$

at site 1 in our DNA sequence.

At site 2 in our DNA sequence,

$$L_1^{(*)} = (P_{14}(v_3))(P_{11}(v_2))$$

and

$$L_{S_0}^{(*)} = \sum_{S_0}(P_{S_04}(v_3))(P_{S_01}(v_2)).$$

Here we find that we cannot calculate these likelihoods without knowing v_1 and v_2. Recall that the algorithm described in Section 3.3 is a special case of the EM algorithm that never produces branch lengths that will decrease the likelihood of the tree. Thus, we are able to set v_2 and v_3 at an initial estimate of the branch lengths while trying to find v_1. We will estimate that $v_2 = v_3 = 1$. Now, we are able to use these values in (1.2) to get

$$L_1^{(*)} = ((1 - e^{-1})(1/4))^2 \approx 0.025$$

in A_1,

$$L_1^{(*)} = ((1 - e^{-1})(1/4))(e^{-1} + (1 - e^{-1})(1/4)) \approx 0.083$$

in A_2, and

$$L_{S_0}^{(*)} = 2((1 - e^{-1})(1/4))^2 + 2((e^{-1} + (1 - e^{-1})(1/4))((1 - e^{-1})(1/4))) \approx 0.216$$

in both B_1 and B_2.

Now, we must take an initial estimate of $p^{(1)}$. We will also estimate that $p^{(1)} = 1 - e^{-1}$. Once we run the EM algorithm a few times, it converges to $p \approx 0.999653874$. Since $p = 1 - e^{-v_1} \approx 0.999653874$, $v_1 \approx 7.969$. We still need to find v_2 and v_3, and since the calculation of v_1 depended on these two branch lengths, it may change again.

Next, we must do the same computation for v_2. We will root the tree directly to the right of node * so that the node lays in v_2. Then, for site 1, the likelihood of the tree is:

$$L = e^{-v_2} \pi_2 L_2^{(*)} + (1 - e^{-v_2}) \sum_{S_0} \pi_{S_0} L_{S_0}^{(*)} \pi_2.$$

This means that $A_1 = \pi_2 L_2^{(*)}$ and $B_1 = \sum_{S_0} \pi_{S_0} L_{S_0}^{(*)} \pi_2$.

For site 2, the likelihood of the tree is:

$$L = e^{-v_2} \pi_1 L_1^{(*)} + (1 - e^{-v_2}) \sum_{S_0} \pi_{S_0} L_{S_0}^{(*)} \pi_1.$$

This means that $A_2 = \pi_1 L_1^{(*)}$ and $B_2 = \sum_{S_0} \pi_{S_0} L_{S_0}^{(*)} \pi_1$.

We are again left with the difficulty of calculating $L_2^{(*)}$ and $L_{S_0}^{(*)}$ for site 1 and $L_1^{(*)}$ and $L_{S_0}^{(*)}$ for site 2. Fortunately, we now have a better estimate for v_1. We will estimate that $v_1 = 7.969$ and that $v_3 = 1$. Then, for site 1,

$$L_2^{(*)} = [(1 - e^{-7.969})(1/4)][(1 - e^{-1})(1/4)] \approx 0.0395$$

and

$$
\begin{aligned}
L_{S_0}^{(*)} = {} & 2[(1 - e^{-7.969})(1/4)][(1 - e^{-1})(1/4)] \\
& + [e^{-7.969} + (1 - e^{-7.969})(1/4)][(1 - e^{-1})(1/4)] \\
& + [(1 - e^{-7.969})(1/4)][e^{-1} + (1 - e^{-1})(1/4)] \approx 0.250. \quad (4.4)
\end{aligned}
$$

For site 2,

$$L_1^{(*)} = [e^{-7.969} + (1 - e^{-7.969})(1/4)][(1 - e^{-1})(1/4)] \approx 0.395$$

and

$$
\begin{aligned}
L_{S_0}^{(*)} = {}& 2[(1 - e^{-7.969})(1/4)][(1 - e^{-1})(1/4)] + [e^{-7.969} \\
& + (1 - e^{-7.969})(1/4)][(1 - e^{-1})(1/4)] \\
& + [(1 - e^{-7.969})(1/4)][e^{-1} + (1 - e^{-1})(1/4)] \approx 0.250. \quad (4.5)
\end{aligned}
$$

Now, we have all of the components to find v_2 using our iteration formula (4.3). Again, we will estimate $p^{(1)} = 1 - e^{-1}$. Once we run the EM algorithm a few times, it converges will converge to a number as it did for v_1. We will then be able to use this number to find v_2.

We follow the same procedure to find an estimate for v_3. Using our new estimates for v_1 and v_2, the EM algorithm will converge to some value. We will use this value to find an estimation for v_3. We use these approximations of v_1, v_2, and v_3 in the EM algorithm again until each of the values converges.

We continue to use the resulting branch length estimates from the EM algorithm to produce better estimates of the branch lengths until the successive estimates begin to converge. Then, the branch lengths that the iteration converges to are accepted as the best estimates for the actual branch lengths of the phylogenetic tree in this step of the maximum likelihood process.

We can now proceed to add our fourth species to the tree. Since there are three branches in the 3-species tree, there are three possible locations for species 4. For site 1 of the DNA sequences, the 4-species trees that can result from the 3-species tree are:
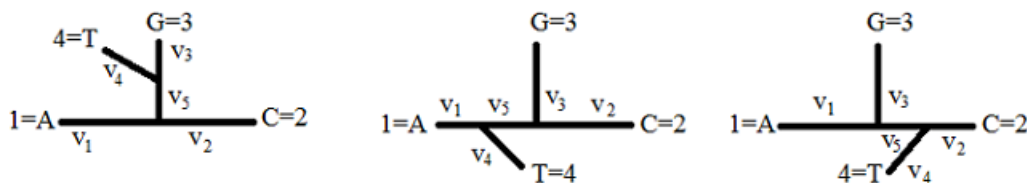
Notice that a new node and two new branches are created after the addition of species 4 to the tree.

Now, we proceed as we did with the three-species tree. For each new tree produced, we evaluate the branch lengths using the EM algorithm. After the iteration converges and gives an estimate of each branch length, we evaluate the likelihoods of each topology. We accept the topology that gives us the greatest likelihood and dismiss the other two topologies. After we have found our optimal 4-species tree, we can add species 5 to each possible location in the accepted 4-species tree. After evaluating the branch lengths and likelihoods of each resulting 5-species tree topology, we have a 5-species tree.
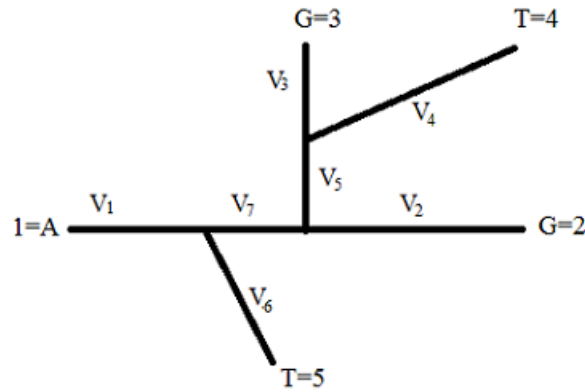

Figure 4.1: The 5-species tree resulting in the greatest likelihood before local rearrangements.

Before we can accept this tree as our maximum likelihood tree, we must perform local rearrangements to help test some possible phylogenetic trees that we may have missed due to the order in which we added each species to the tree. Suppose the tree in Figure 4.1 is the tree we obtain. For each interior branch of the tree, we can perform Nearest Neighbor Interchange (NNI). For interior branch, v_7. the resulting topologies are:

Next, we evaluate the branch lengths and likelihoods of these two topologies, accepting the tree that results in the largest likelihood. If this tree's likelihood is greater than the likelihood of the tree in Figure 4.1, we accept the new tree. In the tree that we accept, whether it be the tree in Figure 4.1 or the new tree, we continue to perform NNI on any remaining interior branches. Each time we do NNI, we evaluate the lengths of the branches and find the likelihood of the new trees. Then, we accept each tree that increases the likelihood. Remember that this is a greedy algorithm, so we continue this process of local rearrangements until we feel that a tree gives
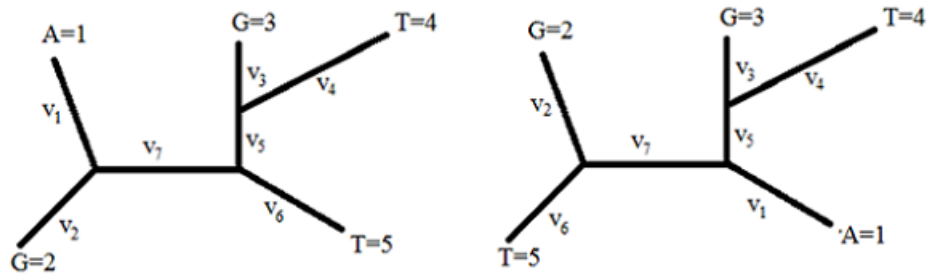
us a sufficiently large likelihood for us to choose it as our maximum like-
lihood tree. The tree that we choose is our best guess for the phylogenetic
trees of the species from Middle Earth. This tree allows us to see how
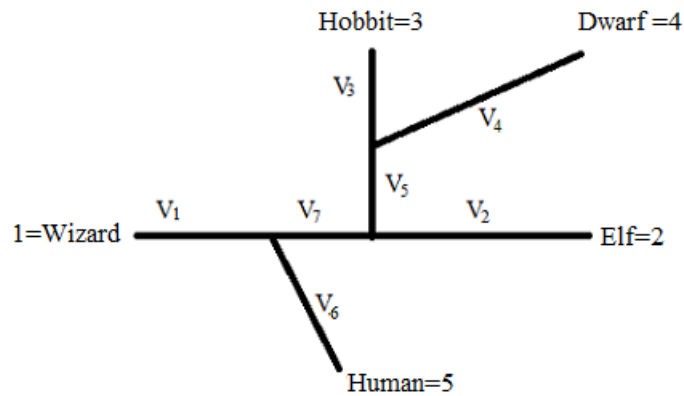closely related each of the species is to one another. Figure 4.2 illustrates
this.

Figure 4.2: Suppose this is our maximum likelihood tree for wizards, elves,
hobbits, dwarves, and humans. The branch lengths, $v_1, v_2, \ldots, v_7$, are
known from the EM algorithm calculations. They help us see how closely
related one species is to another.

Chapter 5

The Limitations of Our Model

We have examined how probability and statistics can describe the process of estimating the most likely phylogenetic tree for a set of DNA sequences. As with any mathematical modeling problem, there are several assumptions that we made in order to ease the computational burden of finding the maximum likelihood tree. It is important to consider the assumptions that we have made and investigate ways in which to make our approach better.

In Chapter 4, we took quite a large leap in modeling base substitution with the Jukes-Cantor model. There are several models of DNA base substitution, including Kimura, Felsenstein, and Tamura (see [3] for descriptions of these models). We did not collect any actual information about the DNA sequences in our data besides the bases that we could observe. After running a few experiments, one is able to determine which model of base substitution best fits the set of species one is examining. For example, in some DNA sequences, the rate of base substitution of a transition can be different from the rate of base substitution of a transversion. Transitions and transversions categorize base substitutions. Both states before and after the substitution can be purines (A or G) or pyramidines (C or T), in which case we call the mutation a transversion. If the base was a purine and was substituted with a pyramidine, or vice versa, we call the base substitution a transition (note: the name, transition, does not relate to the transition probabilties we discussed earlier). There are also cases in which we are unable to observe which specific base we have at the tips of our trees (i.e., in our set of DNA sequences). Sometimes, we are only able to observe whether or not we have a purine or a pyrimadine at the tips of the tree. In order to determine which model of base substitution best de-

scribes the set of DNA sequences we have, we must perform experiments and observations on the DNA.

Our model of DNA mutations also neglects the fact that deletions and insertions can occur. This means that if we begin with K DNA sites, mutations can cause an addition or removal of a DNA site. Then, we may end up with $K + 1$, $K - 1$, $K + 2$, etc. DNA sites.

Another limitation of our model is that it does not consider whether or not rates of substitution vary. We assumed that each DNA site mutated independently from other DNA sites. Gillespie and Langley [6] observed that the number of substitutions can vary in the two branches descending from a common node depending on the location of the DNA site being inspected. This means that depending on how close together two DNA sites are, they may not be independent of each other.

We also made the assumption that the phylogenetic tree is bifurcating. This is not always the case. We may have a multifurcating tree, in which case, we would have to think of a different method for finding the possible locations to add a new species when constructing a tree.

A few of these ambiguities can be solved using our current model of base substitution. For example, when we have a set of DNA sequences in which we are only able to observe whether we have purines or pyrimadines at the tips of our branches, we can adjust the likelihoods of the tips of our tree. If we know that tip i is a purine, then $L_1^{(i)} = L_4^{(i)} = 1$ and $L_2^{(i)} = L_3^{(k)} = 0$. Then, to accomodate for some of the unknown parameters presented here, we can either tweak the Jukes-Cantor model or use a totally different model of base substitution.

The computational burden for the simple example we presented in Chapter 4 was large, and factoring in these limitations can make the computation even more difficult. Fortunately, this method of maximum likelihood estimation has been programmed in PASCAL by Mark Moehring. It estimates the terms of the $p^{(k)}$ during the branch length iteration process. Unfortunately, the program is known to be quite slow. However, since the process of iterating $p^{(k)}$s can become extremely long and tedious, this program is very useful (for more information about the program, see [2].

In this paper, we have described a method to find an estimate of the topology of a set of DNA sequences for a group of species. In our example in Chapter 4, we were able to find a maximum likelihood topology for the family of species (see Figure 4.2). This information allows us to analyze how closely related one species is to another in relation to genetic makeup. We have not, however, determined the DNA of the species' com-

mon ancestors. The likelihoods of each topology must be analyzed further to determine the states of each internal node of the tree.

Bibliography

[1] Dempster, A. P., Laird, N. M., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

[2] Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, (17):368–376.

[3] Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

[4] Gascuel, O., editor (2007). *Mathematics of evolution and phylogeny*. Oxford University Press, Oxford.

[5] Ghahramani, S. (2000). *Fundamentals of Probability*. Prentice Hall, Upper Saddle River, NJ, second edition.

[6] Gillespie, J. H. and Langley, C. H. (1979). Are evolutionary rates really variable? *Journal of Molecular Evolution*, pages 27–34.

[7] Jukes, T. H. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, pages 21–123),.

[8] Sadava, D., Heller, C., Orians, G., Purves, B., and Hills, D. (2008). *Life the Science of Biology*. The Courier Companies, Inc., Sunderland, MA, eighth edition.