# SELECTING AN ACCURATE MODEL OF EVOLUTIONARY RATE HETEROGENEITY

JOHN ST. JOHN

Abstract. Determining the evolutionary history of genes and organisms is a critical aspect of fields as diverse as developmental biology, astrobiology, history, anthropology, and medicine. The understanding gained through these phylogenetic analyses provides valuable information about the past and present unavailable through other avenues. The most widely used modern phylogenetic techniques require some knowledge about how evolution works to ensure accurate results. Selecting the best evolutionary model available for a given set of data is thus a critical step of phylogenetic analyses. One aspect of evolution that is not commonly accounted for is when different sites on a sequence have varying rates of change that differ in location on the sequences being compared. This phenomena, known as heterotachy, has been shown to be modeled best by a Mixed Branch Length (MBL) model. The question then is how many heterotachous partitions are needed to adequately model the underlying evolutionary history. This question is answered through the empirical evaluation of modern model selection techniques on simulated data. The findings suggest that the Akaike Information Criterion ($AIC$) is the best technique for determining the number of heterotachous rate partitions to use in the MBL Model of evolution.

*Date*: June 3, 2009.

## Contents

## 1. Biological Background

**1.1. Evolution and Speciation.** Evolution occurs recursively through the process of speciation followed by independent changes occurring in the separate populations. Speciation starts with reproductive isolation, the cessation of interbreeding between populations. Reproductive isolation arises through populations being separated geographically, through sexual behavior, or through physical incompatibility. Although the definition of a species is debatable, it is generally accepted that once two populations cease to interbreed, and will not do so even if they have the opportunity, they are separate species. Once populations cease to interbreed, they can undergo changes independently of each other. Over time, this process led to the diversity seen in the world today, and much more that remains to be discovered.

**1.2. The Central Dogma of Biology.** DNA is the blueprint for every living thing currently known. Changes in DNA can lead to changes in the organism. Every cell within an organism contains its own copy of that organism's DNA. DNA is essentially a string of 4 types of molecules called nucleotides (Higgs and Attwood, 2005). These four nucleotides are called Adenine, Guanine, Cytosine and Thymine (Higgs and Attwood, 2005). They are typically abbreviated A, G, C and T respectively. An organism's DNA can range in length from several thousand, like some single celled bacteria and most viruses, to several billion like many multicellular organisms (Higgs and Attwood, 2005). DNA does its work through a complex series of regions called genes that encode one, or multiple RNA sequences that are translated into proteins. Genes may also encode RNA sequences that never end up being encoded into proteins, but are used instead for other purposes in the cell such as the regulation of which proteins are allowed to be produced in the cell. RNA is very similar to DNA except rather than having a Thymine, it has a Uracil in its place (abbreviated U) (Higgs and Attwood, 2005). These end products of Protein or RNA strands are sometimes responsible for cellular function, or the regulation of other genes. The functional interaction between proteins, RNA segments, and DNA, together form complex regulatory networks. These genes are controlled by a plethora of regulatory sites that are responsible for determining when those genes are turned on, the level to which they are turned on, or which proteins should be produced by those genes. Changes in the sequence that sufficiently disrupt a necessary gene, or gene activator lead to cell death; however many changes do not produce that effect due to redundancies in cellular and chemical networks in an organism (Higgs and Attwood, 2005).

### 1.3. How New Traits Arise.

Changes in DNA are caused by mutation. Mutation can come in many forms including single sites switching from one nucleotide to another, insertions of new nucleotides, deletions of one or more nucleotides, regions of DNA being copied and pasted, or regions of DNA being cut out and pasted in a new location without being copied. Changes in DNA can lead to changes in the organism, which is the method by which the vast differences between species that have been apart for long periods of time are achieved. Changes that are sufficiently detrimental are often selected out of existence before they become significant contributors to a population's gene pool; however if a population size is small, or the change doesn't have much of an effect on the organism, it is free to become fixed in a population over time (Higgs and Attwood, 2005).

1.4. **Which Traits Last.** Some changes lead to a decrease in an organism's ability to successfully reproduce, either through death before the organism has had the chance to reproduce, or any other decreased ability to pass on genetic material. These additions to the population's gene pool generally do not survive, depending on their severity. This is what is referred to as negative selection. Some changes lead to an increase in an organism's ability to pass on genetic material. These changes are advantageous, and organisms that have these changes are more likely to reproduce. This is referred to as positive selection. Neutral or even negative traits can sometimes become fixed in a population due to phenomena called genetic drift. Genetic drift allows for non grossly deleterious changes to not only survive but rise to fixation in a population. In other words the entire population may adopt the deleterious change through subsequent generations simply due to random circumstances. The probability of this happening decreases substantially as population size increases. Environmental changes, or changes in a populations sexual preference may change which traits, coded by the DNA, are the most successful.

1.5. **Summary.** Over the course of history, environments have changed, speciation has occurred, and populations have evolved to adapt to these changes through a combination of positive selection for beneficial changes, negative selection for deleterious changes, and the power of randomness fixing neutral and even slightly deleterious changes. Sites within DNA that were vastly important at some time in history may no longer be important, and thus will be lost or recruited to a new function. At the same time new sites that werent important in the past may become very important to a species survival.

## 2. Biological & Computational Problem

**2.1. Introduction.** Understanding how organisms evolve is a critical component of biology, and bioinformatics (Higgs and Attwood, 2005). For example understanding the evolution of disease-model organisms, and how they differ on the molecular level from humans, improves their usefulness for biomedical research (Gibbs et al., 2007). Phylogenetics is the study of evolutionary relationships, and is concerned with discovering the historical pattern of relatedness among species or genes. Properly modeling evolution, and designing automated methods of determining phylogenetic relationships, is a very important application of computer science to the field of biology. The most widely used computational techniques to determine phylogenetic relationships include maximum parsimony, and maximum likelihood.

**2.2. Sequence Alignments.** For the following sections, phylogenetic alignments are considered on sequences, and for the purpose of this paper DNA sequences will be considered. Following is an example of what a sequence alignment might look like:

```
Sequence 1:  C-CT
Sequence 2:  CAGT
Sequence 3:  CAG-
```

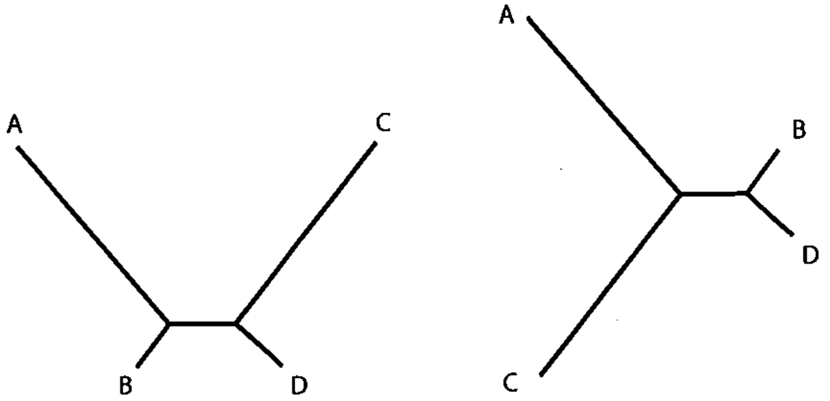
The computational problem of how to arrive at an alignment is beyond the scope of this paper, however a very rough idea is that sites in a sequence with shared ancestry should also share a column, dashes are inserted at certain points to account for mutations that inserted extra characters in one or more of the sequences. Arriving at the most probable alignment is an active area of research. In this paper, the focus is on the process at discovering the ancestry of organisms, and asumes that an oracle has provided the correct sequence alignment to accomplish this task.

In this paper a state refers to a single character. For example the first nucleotide of "Sequence 1" above is a state. A homologous state refers to states that are the same in multiple organisms within the same column of an alignment. Consider the first column in the above alignment, in this column the states are homologous between all sequences. Phylogenetic trees, such as those in figure 1, are derived from sequence alignments like those above using one of several methods.

2.2.1. *Maximum Parsimony.* Maximum Parsimony(MP) is the introductory technique many students get to phylogenetic analysis using genetic sequence data. According to Darwinian evolutionary theory, the only way that one can hope to quantitatively determine evolutionary relatedness is to study synapomorphies. Synapomorphies are homologous states with a common origin shared between two organisms or genes. A given state's homology doesn't necessarily infer synapomorphy. The trait could have independently arose in separate populations, which is called homoplasy. To get an idea of the problem of homoplasy in phylogenetics, consider two distantly related organisms where at one column in the alignment, one organism has an "A" and the other has a "C". Then imagine that the organism with the "C" undergoes a mutation that changes that site to an "A". If the ancestral state of both organisms were not known, and two "A"s were observed in the alignment, it would be impossible to know for sure whether the state is shared from a common ancestor, or if the state was arrived at independently in both organisms. The most parsimonious explanation for a phylogenetic relationship is the one that assumes the maximal number of synapomorphies, and thus makes the fewest "ad-hoc hypotheses" of states being shared via independent mutations (Farris, 1982). It is important to note that MP doesn't rely on the assumption that homoplasies aren't prevalent (Farris, 1982). Under evolutionary conditions in which the homoplasy is shared fairly evenly among organisms, the correct phylogeny is still the one with a maximal number of shared states, regardless of whether those states are synapomorphic or homplasic (Farris, 1982). However the parsimonious explanation doesn't necessarily perform well under certain structures of homoplasy.

Although parsimony works well in many cases, cases exist in which maximum parsimony will select the incorrect evolutionary tree with increasing certainty (Felsenstein, 1978). In fact, under those evolutionary conditions, maximum parsimony will select the incorrect tree with increasing support as more data is analyzed (Felsenstein, 1978). This phenomona is called long-branch attraction, and it occurs under evolutionary conditions called the Felsenstein Zone. Under these conditions, the number of homologous states between more distantly related sequences due to homoplasies may outnumber those due to true synapomorphies between more closely related sequences which leads MP astray. In the tree shown in Figure 1(a) the longer branches would tend to be grouped together under a parsimony analysis, as shown in Figure 1(b), with increasing probability as the sequence length of the alignment increases.

2.2.2. *Maximum Likelihood.* Maximum likelihood (ML) has been shown to be a more accurate phylogenetic technique under some conditions (Felsenstein, 1978). For that reason it is generally preferred to MP. Likelihood in phylogenetics is the probability of observing the sequence alignment, given the model of evolution and the phylogenetic



(a) An example of an unrooted phylogenetic tree that yields biased results under Maximum Parsimony but not under Maximum Likelihood. This tree configuration is called the "Felsenstein zone".

(b) This is how the tree in Part (a) is evaluated by Maximum Parsimony. The long branches attract to each other and A is falsely grouped with C. This phenomenon is also known as long branch attraction.

FIGURE 1. A case in which parsimony methods yield incorrect results.

tree (Foster, 2001; Bryant et al., 2005). Consider the following hypothetical DNA sequence alignment:

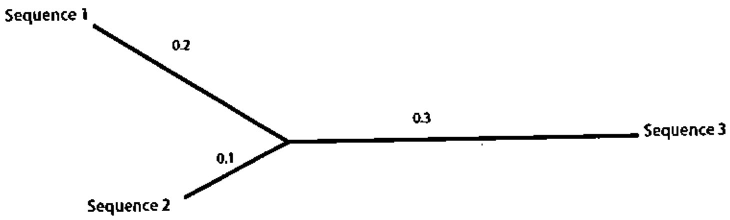
```
Sequence 1:  CCT
Sequence 2:  CGT
Sequence 3:  CAG
```

Now to calculate the likelihood one must assume a model, and an evolutionary tree. For the purpose of this example, the model will assume that each column in the alignment is independent from each other column. Additionally the model will assume the nucleotides, {A, C, G, T}, have the following probabilities of existence respectively.

$$\pi = [0.1, 0.4, 0.2, 0.3] \tag{1}$$

The symbol $\pi$ is typically used to denote the frequency of a character, and $\pi_A$ is used to specify the frequency of character A. Next an evolutionary tree must be defined. In phylogenetics, evolutionary trees are typically defined with branch lengths in units of the number of mutations per site. Thus when considering the sequence at the end of a branch with a length of 0.3, the probability that each site will have experienced a mutation event (A-A, A-C, A-G, A-T, and so on...) is 0.3. For this calculation the following tree will be assumed, where the numbers on the branches represent their length:
It is important to note that the tree above may or may not be the correct tree, likelihood can be calculated for any tree on the data.

Since even a simple model of evolution should encapsulate the probability of change from one character to another, a state change probability matrix is defined. The probability that one state remains the same, or changes to another state, logically should be different based on the number of times the sequence has undergone change.



The number of times a sequence has undergone change is expressed as a probability in the branch lengths discussed previously. Using some linear algebra techniques discussed in further detail in Bryant et al. (2005), one can calculate the state change probability matrix for any branch length. The following probability matrices (taken from Foster (2001)) are denoted $P(n)$ where $n$ is the number of expected substitutions per site taken from the assumed evolutionary tree.

$$P(0.1) = \begin{bmatrix} 0.886 & 0.047 & 0.031 & 0.036 \\ 0.012 & 0.918 & 0.024 & 0.046 \\ 0.016 & 0.047 & 0.902 & 0.036 \\ 0.012 & 0.062 & 0.024 & 0.902 \end{bmatrix} \tag{2}$$

$$P(0.2) = \begin{bmatrix} 0.787 & 0.089 & 0.057 & 0.067 \\ 0.022 & 0.847 & 0.045 & 0.086 \\ 0.029 & 0.089 & 0.815 & 0.067 \\ 0.022 & 0.115 & 0.045 & 0.819 \end{bmatrix} \tag{3}$$

$$P(0.3) = \begin{bmatrix} 0.7 & 0.126 & 0.08 & 0.094 \\ 0.031 & 0.786 & 0.063 & 0.119 \\ 0.04 & 0.126 & 0.74 & 0.094 \\ 0.031 & 0.159 & 0.063 & 0.747 \end{bmatrix} \tag{4}$$

The matrices $P(n)$ are arranged so that rows and columns are ordered (A, C, G, T) from left to right and top to bottom. The position in the matrix $P(0.1)_{A \to G}$ can be read as the probability of a change from A to G given 0.1 expected substitutions per site.

Calculating the likelihood of the data given the previously discussed tree, and model would be calculated as the product of the likelihoods for each column in the alignment. Given the column in the alignment $n$, the alphabet $\Psi$={A, C, G, T}, the set of branches mapping to lengths $\Phi = \{Sequence1 \mapsto 0.2,\ Sequence2 \mapsto 0.1,\ Sequence3 \mapsto 0.3\}$, the probabilities of existence in equation 1, and the character at column $i$ in branch $m \in \Phi$ denoted $\gamma_{i \in m}$ the likelihood can be calculated as follows (Foster, 2001):

$$L = \prod_{i=1}^{n} \sum_{\ell \in \Psi} \pi_\ell \prod_{m \in \Phi} P(m)_{\ell \to \gamma_{i \in m}}$$

substituting in the previously mentioned data, the following likelihood is calculated (Foster, 2001):

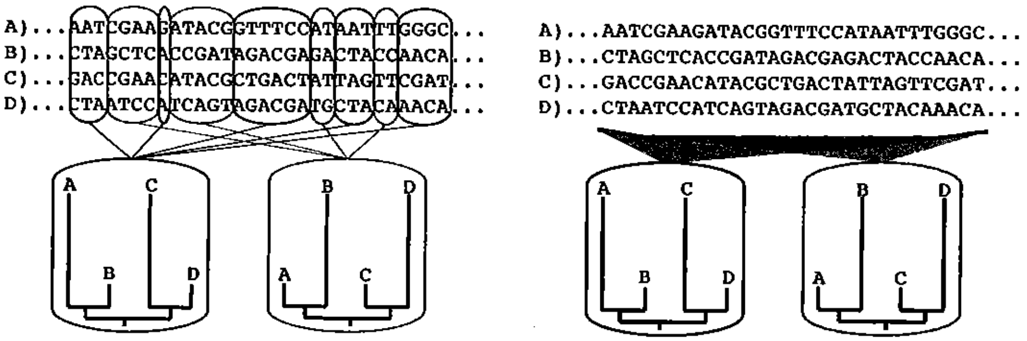
$$
\begin{aligned}
L = & \left( \begin{array}{lllll}
\pi_A & P(0.2)_{A \to C} & P(0.1)_{A \to C} & P(0.3)_{A \to C} & + \\
\pi_C & P(0.2)_{C \to C} & P(0.1)_{C \to C} & P(0.3)_{C \to C} & + \\
\pi_G & P(0.2)_{G \to C} & P(0.1)_{G \to C} & P(0.3)_{G \to C} & + \\
\pi_T & P(0.2)_{T \to C} & P(0.1)_{T \to C} & P(0.3)_{T \to C} &
\end{array} \right) \\
\times & \left( \begin{array}{lllll}
\pi_A & P(0.2)_{A \to C} & P(0.1)_{A \to G} & P(0.3)_{A \to A} & + \\
\pi_C & P(0.2)_{C \to C} & P(0.1)_{C \to G} & P(0.3)_{C \to A} & + \\
\pi_G & P(0.2)_{G \to C} & P(0.1)_{G \to G} & P(0.3)_{G \to A} & + \\
\pi_T & P(0.2)_{T \to C} & P(0.1)_{T \to G} & P(0.3)_{T \to A} &
\end{array} \right) \\
\times & \left( \begin{array}{lllll}
\pi_A & P(0.2)_{A \to T} & P(0.1)_{A \to T} & P(0.3)_{A \to G} & + \\
\pi_C & P(0.2)_{C \to T} & P(0.1)_{C \to T} & P(0.3)_{C \to G} & + \\
\pi_G & P(0.2)_{G \to T} & P(0.1)_{G \to T} & P(0.3)_{G \to G} & + \\
\pi_T & P(0.2)_{T \to T} & P(0.1)_{T \to T} & P(0.3)_{T \to G} &
\end{array} \right)
\end{aligned}
$$

Although calculating likelihood is fairly straightforward, the bulk of the computational complexity stems from the fact that likelihood must be calculated so many times in an ML evaluation. To calculate ML, one must independently vary all of the free parameters, and compare likelihoods generated from each variation, selecting the set of parameters that lead to the *maximum* likelihood. The free parameters may include tree topology, branch lengths, the probability of observing each character (or subsets of characters), and maybe others depending on the model of evolution used. Since there are many potential tree topologies, an infinite number of potential branch lengths, infinite values for many of the other parameters, and there is no known algorithmic way to directly arrive at the maximum likelihood parameters (Bryant et al., 2005; Felsenstein, 2003), various heuristics are used to approximate values for the parameters (Bryant et al., 2005; Felsenstein, 2003).

Given the correct model of evolution, Maximum Likelihood is guaranteed to be statistically consistent. Statistical consistency means that the correct result will always have the highest likelihood value (Felsenstein, 2003). This is good because it means that the tree that generates the highest likelihood given the data, and the correct model of evoution, is the optimal tree. The important note here is that these properties of ML are guaranteed only when the correct model of evolution is assumed (Felsenstein, 2003). Even if it can be shown that in some cases a given model of evolution performs well on some data, unless that model accurately represents the true model that generated the data, MLs consistency cannot be assumed, and the tree with the highest likelihood may not be the best tree. Selecting the correct model of evolution is a critical step in phylogenetic inference.

2.3. **The Problem of Heterotachy.** ML is guaranteed to provide consistent results if the model of evolution provided is correct. In most cases currently popular models of evolution produce highly accurate results when used with ML. However one case in which even ML is highly biased in favor of the wrong phylogeny is when there are certain types of unrepresented evolutionary rate heterogeneity (heterotachy) (Kolaczkowski, 2004). Figure 2(a) shows an example of how heterotachous evolution might look on a simple dataset. The idea behind heterotachy is that there are varying rates among sites in a sequence that vary between sequences. This differs from among site rate variation(ASRV) because among site rate variation locks in the overall relative rates of evolution between sequences. Utilizing ASRV, a site in a sequence that evolves faster overall than another sequence, is forced to evolve faster relatively to the same site in a slower evolving sequence. This is not a biologically accurate assumption. One evolutionary condition that heterotachy properly models and ASRV fails to capture is through the differential selective constraints on a protein that vary between organisms or proteins. Some changes that benefit, kill or have no effect on one gene or organism may have an entirely different effect on another gene or organism. It is not hard to imagine that varying levels of constraints on the DNA sequence itself would also have a similar effect of creating heterotachous evolutionary conditions. Some sites in quickly evolving organisms may in fact change at a slower rate than the corresponding sites in a more slowly evolving organism, and common phylogenetic techniques including ASRV do not allow for this intellectually obvious evolutionary scenario.

Real world examples have been found where unincorporated heterotachy leads to phylogenetic error. Heterotachy is likely responsible for the incorrect grouping of Microsporidia with Archaebacteria through all commonly used phylogenetic methods available in 2004 when examining Elongation Factor 1-alpha (EF1-alpha) (Inagaki et al., 2004). There are probably many other situations in which this form of heterotachy is currently causing incorrect phylogenies (Inagaki et al., 2004).


A)...AATCGAAGATACGGTTTCCATAATTTGGGC...
B)...CTAGCTCACCGATAGACGAGACTACCAACA...
C)...GACCGAACATACGCTGACTATTAGTTCGAT...
D)...CTAATCCATCAGTAGACGATGCTACAAACA...
A
C
B
D
B
D
A
C
A)...AATCGAAGATACGGTTTCCATAATTTGGGC...
B)...CTAGCTCACCGATAGACGAGACTACCAACA...
C)...GACCGAACATACGCTGACTATTAGTTCGAT...
D)...CTAATCCATCAGTAGACGATGCTACAAACA...
A
C
B
D
B
D
A
C

(a) An example of heterotachous evolution. Some sites evolve under the set of evolutionary rates on the left while others evolve under the evolutionary rates shown on the right.

(b) This figure shows how the evolutionary rates are proportionally summed over all the sites under the heterogenous rate model developed by Kolaczkowski & Thornton. This models the evolutionary conditions shown in Part (a).

FIGURE 2. Heterotachy and how it is modeled.

Modeling heterotachy when it exists in the patterns previously discussed is necessary to get better phylogenetic results out of maximum likelihood. Without first modeling this heterotachy and applying model selection techniques, it is impossible to know for sure whether or not it plays an important role in sequence evolution. Figure 3 (from Kolaczkowski and Thornton (2008)) shows the effectiveness of various phylogenetic techniques at evaluating the data generated by a particularly hard to compute heterotachous model of evolution. Kolaczkowski and Thornton (2008) found that when the internal branch length of the tree that generated the data is short, all models tested that did not properly incorporate heterotachy failed to determine the true phylogeny. Ideally one would create a model of evolution that perfectly partitions the columns of aligned genomic data into their various rate partitions as is displayed in Figure 2(a). However this information is often not known in advance. Luckily a well-known statistical process called mixture modeling can be used. With a mixture model, one calculates the weighted average of multiple evolutionary models over the entire sequence, as is shown in Figure 2(b). This average is weighted by the proportion of sites that are calculated to most likely fall under each branch length set, as calculated and optimized through maximum likelihood (Kolaczkowski and Thornton, 2008).

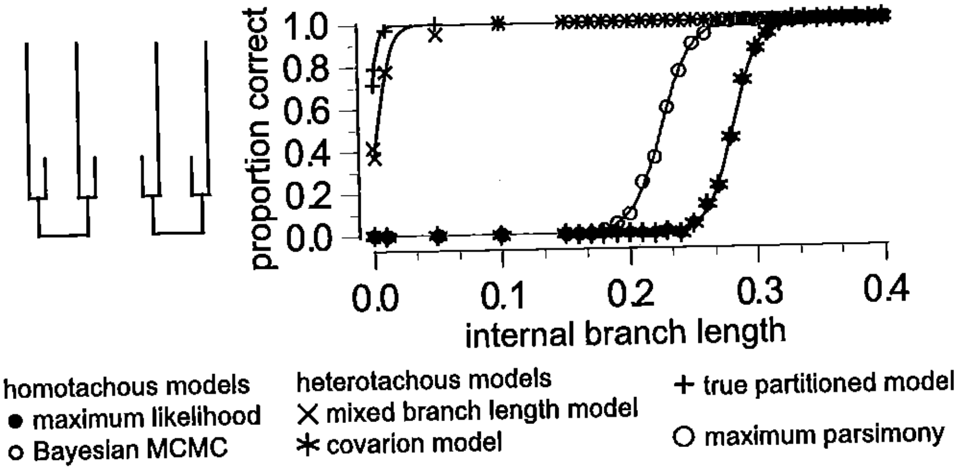


FIGURE 3. Figure from Kolaczkowski and Thornton (2008) showing the comparative performance of various phylogenetic models at determining the correct phylogeny. The two category heterotachous model of evolution used to generate the data is shown at the upper left. Sequences of 5000 nucleotides were generated under this set of branch lengths. The long terminal branch lengths were 0.75 while the short ones were 0.05. The graph shows the plot of the proportion of correct phylogenetic inferences verses increasing internal branch length. More accurate methods do not require as long an internal branch to reliably reconstruct the true phylogeny.

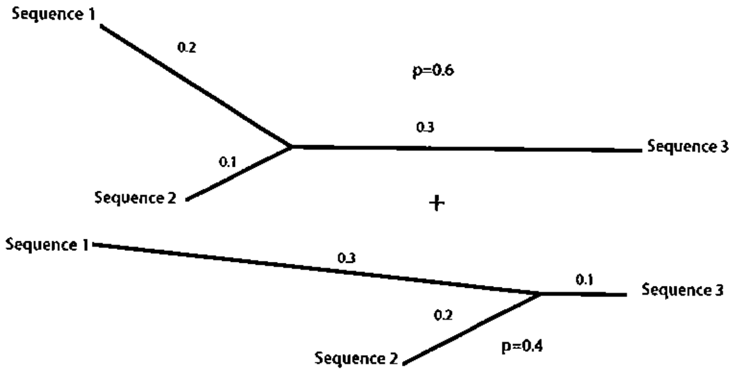
The mixed branch length model works by first grouping all of the branch lengths of a given topology into a set, independently optimizing each branch length within each set, and performing this operation on the sets independently of each-other. The likelihood given each set is calculated independently, and then summed proportionally with the likelihood given the other sets, where the proportion of sites falling under each set is also optimized by ML (Kolaczkowski and Thornton, 2008; Spencer et al., 2005). The equation is:

$$L(t|X) = \prod_{k=1}^{m} \sum_{i=1}^{n} \rho_i P(x_k|t, b_i) \tag{5}$$

where $X = (x_1, x_2, \ldots, x_m)$, $x_k$ is column $k$ in the alignment, $t$ is the tree, $m$ is the number of data points (columns in the alignment), $n$ is the number of rate partitions, $\rho_i$ is the proportion of sites that fall under rate partition $i$, and $P(x_k|t, b_i)$ is the probability of observing the alignment at column $x_k$ given tree $t$, and branch lengths $b_i$(Kolaczkowski and Thornton, 2008).

For a practical application of how one might compute likelihood given a heterotachous model consider the following sequence alignment:

| | |
|---|---|
| Sequence 1: | CCT |
| Sequence 2: | CGT |
| Sequence 3: | CAG |



Since it is not known a-priori which sites in the above alignment fall under which set of branch lengths, the formula for calculating likelihood proposed by Kolaczkowski and Thornton (2008) must be used. In the above branch length set are two heterotachous rate partitions, and the values of p denote the assumed proportion of sites that fall under one set of branch lengths or the other.

Additionally for this example consider the following equation which is an expansion of equation 5:

$$L(t|X) = \prod_{k=1}^{m} \sum_{i=1}^{n} \rho_i \sum_{\ell \in \Psi} \pi_\ell \prod_{q \in \Phi_i} P(q)_{\ell \rightarrow \gamma_{k \in q}} \tag{6}$$

where $X = (x_1, x_2, \ldots, x_m)$, $x_k$ is column $k$ in the alignment, $t$ is the tree, $m$ is the number of data points (columns in the alignment), $n$ is the number of rate partitions, $\rho_i$ is the proportion of sites that fall under rate partition $i$, $\Psi$={A, C, G, T}, the set of all branches mapping to lengths $\Phi = \{Sequence1 \mapsto [0.2, 0.3],\ Sequence2 \mapsto [0.1, 0.2],\ Sequence3 \mapsto [0.3, 0.1]\}$ ($\Phi_i$ represents the branch length at position $i$ in each mapped set), the probabilities of existence in equation 1, the probability of observing each character $\pi$ from equation 1, and the character at column $i$ in branch $m \in \Phi$ denoted $\gamma_{i \in m}$.

Substituting in values we get the following likelihood:

$$
\begin{aligned}
L = & \left( \begin{array}{lllll}
\pi_A & P(0.2)_{A\to C} & P(0.1)_{A\to C} & P(0.3)_{A\to C} & + \\
\pi_C & P(0.2)_{C\to C} & P(0.1)_{C\to C} & P(0.3)_{C\to C} & + \\
\pi_G & P(0.2)_{G\to C} & P(0.1)_{G\to C} & P(0.3)_{G\to C} & + \\
\pi_T & P(0.2)_{T\to C} & P(0.1)_{T\to C} & P(0.3)_{T\to C} &
\end{array} \right) \\
\times & \left( \begin{array}{lllll}
\pi_A & P(0.2)_{A\to C} & P(0.1)_{A\to G} & P(0.3)_{A\to A} & + \\
\pi_C & P(0.2)_{C\to C} & P(0.1)_{C\to G} & P(0.3)_{C\to A} & + \\
\pi_G & P(0.2)_{G\to C} & P(0.1)_{G\to G} & P(0.3)_{G\to A} & + \\
\pi_T & P(0.2)_{T\to C} & P(0.1)_{T\to G} & P(0.3)_{T\to A} &
\end{array} \right) \\
\times & \left( \begin{array}{lllll}
\pi_A & P(0.2)_{A\to T} & P(0.1)_{A\to T} & P(0.3)_{A\to G} & + \\
\pi_C & P(0.2)_{C\to T} & P(0.1)_{C\to T} & P(0.3)_{C\to G} & + \\
\pi_G & P(0.2)_{G\to T} & P(0.1)_{G\to T} & P(0.3)_{G\to G} & + \\
\pi_T & P(0.2)_{T\to T} & P(0.1)_{T\to T} & P(0.3)_{T\to G} &
\end{array} \right) \times 0.6 \\
+ & \left( \begin{array}{lllll}
\pi_A & P(0.3)_{A\to C} & P(0.2)_{A\to C} & P(0.1)_{A\to C} & + \\
\pi_C & P(0.3)_{C\to C} & P(0.2)_{C\to C} & P(0.1)_{C\to C} & + \\
\pi_G & P(0.3)_{G\to C} & P(0.2)_{G\to C} & P(0.1)_{G\to C} & + \\
\pi_T & P(0.3)_{T\to C} & P(0.2)_{T\to C} & P(0.1)_{T\to C} &
\end{array} \right) \\
\times & \left( \begin{array}{lllll}
\pi_A & P(0.3)_{A\to C} & P(0.2)_{A\to G} & P(0.1)_{A\to A} & + \\
\pi_C & P(0.3)_{C\to C} & P(0.2)_{C\to G} & P(0.1)_{C\to A} & + \\
\pi_G & P(0.3)_{G\to C} & P(0.2)_{G\to G} & P(0.1)_{G\to A} & + \\
\pi_T & P(0.3)_{T\to C} & P(0.2)_{T\to G} & P(0.1)_{T\to A} &
\end{array} \right) \\
\times & \left( \begin{array}{lllll}
\pi_A & P(0.3)_{A\to T} & P(0.2)_{A\to T} & P(0.1)_{A\to G} & + \\
\pi_C & P(0.3)_{C\to T} & P(0.2)_{C\to T} & P(0.1)_{C\to G} & + \\
\pi_G & P(0.3)_{G\to T} & P(0.2)_{G\to T} & P(0.1)_{G\to G} & + \\
\pi_T & P(0.3)_{T\to T} & P(0.2)_{T\to T} & P(0.1)_{T\to G} &
\end{array} \right) \times 0.4
\end{aligned}
$$

2.4. **Model Selection.** Choosing the number of rate categories to use in MBL is not a trivial problem. By simply examining the data, it is not clear how one would determine the number of rate categories that best explains the data. The current approach is to run the MBL on the data multiple times assuming different numbers of rate categories each time, and examine the results using a standard statistical model selection technique; the goal is to select the model that best explains the data without over-fitting.

Unfortunately likelihood scores alone are not sufficient to choose the best model. Taken at face value, likelihood scores will become better each time the analysis is performed with a more complicated model of evolution. Since not every site is perfectly explained by the true phylogeny and model of evolution, an overly complex model will be able to fit those sites that may support the wrong phylogeny, and in fitting those sites would have a greater likelihood. Over-fitting the data artificially reduces support for the ML phylogeny. Additionally when an overly complex model of evolution is used with this algorithm, there is a significant computational sacrifice. Each additional rate category squares the complexity of the model with one fewer rate categories.

Luckily there are statistical methods available to tackle the problem of model selection. The methods currently used in biology include the Likelihood Ratio Test ($LRT$) (Wilks, 1938), the Akike Information Criterion ($AIC$) (Akaike, 1974), a corrected version of $AIC$ called $AIC$ Corrected ($AIC_c$) (Shono, 2000), the Bayesian Information Criterion ($BIC$) (Schwarz, 1978), and Cross Validation ($CV$) (Zhou et al., 2007). $AIC$, $AIC_c$, and $BIC$ all work by applying a penalty for increasing complexity directly to the log likelihood score of the data, given each model. This penalty is simply a function of the number of parameters in the model, and in the case of $BIC$ and $AIC_c$, the amount of data as well. $LRT$ takes advantage of the fact that when the simpler model is true, then $2 \times ln(\frac{L_{simple}}{L_{complex}})$ is $\chi^2_k$ distributed if the models are nested, where $k$ is the difference in degrees of freedom between $lnL_{simple}$ and $lnL_{complex}$ (Felsenstein, 2003). The $LRT$ tests whether an increase in likelihoods given the number of parameters that are being added, is likely due to chance, or due to a model that is actually fitting the data better. Cross Validation works by evaluating a subset of the data given the model, and then applying the best-fit parameters to the remainder of the data. The idea with Cross Validation is that an overly complex model is likely to over fit the test data, and will be worse than the correct model at predicting the "left out" portion of the data.

$LRT$ is perhaps the most commonly used statistical model selection technique in biology. This method is based on the mathematical observation that, if model $m_1$ is a subset of the model $m_2$, then $2(\ln[L(D|m_1)] - \ln[L(D|m_2)])$ is roughly $\chi^2_k$ distributed with $k$ degrees of freedom, where $k$ is the difference in the number of free parameters between the two models (Felsenstein, 2003; Higgs and Attwood, 2005). To determine if the shift in log likelihood values obtained from Maximum Likelihood analysis using model $m_2$ is statistically significant, simply see where the likelihood ratio falls on the $\chi^2_k$ distribution (Higgs and Attwood, 2005). The problem with $LRT$ is that it requires models to be nested, and varying numbers of heterotachous rate categories in the mixed model are not nested. For example a mixed model with two rate categories is the strict boarder case of a mixed model with three categories, where one of the three categories effects 0% of the sites.

Unlike $LRT$, which requires nested models, $AIC$, $AIC_c$, and $BIC$ only require knowledge of the number of free parameters in each model that is being compared. This provides a logical reason to prefer other techniques to $LRT$, however it is important to test $LRT$ to see if it happens to work well. $AIC$, $BIC$ and $AIC_c$ are perhaps the easiest of the model selection techniques to implement. Each of the techniques generate scores by applying a function to the log likelihood result of an ML evaluation.

$AIC$ (Akaike, 1974) is determined by:

$$AIC = -2(\ln(L)) + 2k \tag{7}$$

In the above equation, $k$ is the number of free parameters in the model used to calculate $L$. In the case of the MBL model, $k$ is calculated as follows:

$$k = (b \times m) + (m - 1) \tag{8}$$

In the above equation, $b$ is the number of branches in the phylogenetic tree, and $m$ is the number of independent branch length categories that are assumed in the MBL model. The logic behind the selection of $k$ is that there are $b$ free parameters per tree, $m$ trees because each rate category assumes fully independent rates, and an added $m - 1$ parameters to solve for the proportions of sites that fall under each tree.

$AIC_c$ was developed due to the observation that on some kinds of data, when the number of data points is small, or the number of parameters approaches the number of data points, $AIC$ tends to select an overly complex model (Shono, 2000). This corrected version of $AIC$ is defined as follows:

$$AIC_c = AIC + \left(\frac{2k(k+1}{n-k-1}\right) \tag{9}$$

In the above equation, $n$ is the number of data points (number of columns in a sequence alignment for example). $AIC_c$ corrects the bias in $AIC$ by adding a penalty that decreases to zero as the sample size gets larger in relation to the number of parameters. This takes care of the case in which the number of parameters is near the count of data points, and also the error that occurs when there are a small number of data points.

$BIC$ is like $AIC$, but it applies an increasing penalty to the number of parameters, as the number of data points increase. $BIC$'s equation follows:

$$BIC = -2\ln(L) + k\ln(n) \tag{10}$$

$BIC$ tends to penalize more heavily than $AIC$, whenever $n \geq 8$. Because real world sequence data are almost always longer than 8 sites one can predict that $BIC$ will penalize more heavily than $AIC$, and favor simpler models in nearly all cases.

$CV$ works differently from the other statistical techniques in that it requires no knowledge about the underlying models assumed on the data. This benefit comes at a cost to computational complexity. The other techniques require only the resulting log likelihood scores from several runs with different models of evolution, while $CV$ requires iteratively sampling the data and calculating multiple likelihoods per model. The idea behind $CV$ is that even though an overly complex model will fit data better than the true model, if the overly complex model is trained on a subset of the data, it is less likely that it will be a good fit to the remainder of the data. The classic

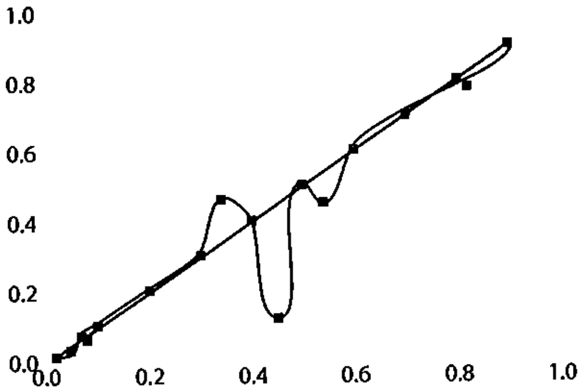


FIGURE 4. An eaxmple of how data may be overfit. The model that generated the points on the graph is the line in blue. The generating model had some error so some of the points were not exactly on the line. The red line shows the points fit by an overly complex model.

visual example of why $CV$ works is in the problem of regression. In Figure 4, the red line represents a model with high dimensionality that is over fitting points that were hypothetically created under a linear model depicted in blue. Imagine that several points on the line were left out of the training data set for $CV$. The regression line that $CV$ would draw then would perfectly match the points that it had, but would be far off from the remaining points that were on the line. If a linear model were used in the beginning to train the data, the few noisy data points would not be recognized as significant, and the line would be closer to the remaining points that were left out of the training data set. To analyze biological data, first one simply optimizes parameters using maximum likelihood or Bayesian methods on a randomly sampled portion of the aligned sequences. Then takes those parameters and uses them to calculate the likelihood score of the remainder of the data (the test data) (Zhou et al., 2007). The model chosen by $CV$ is the model that results in the lowest likelihood from the analysis of the test data. This process should be repeated several times for each alignment, and the likelihood from each of these runs should be averaged to provide the final $CV$ score for the given model of evolution(Zhou et al., 2007).

## 3. Methods

**3.1. Test Data.** To test the accuracy of various model selection techniques, simulated evolution scenarios were generated in sillico so the true model of evolution was known a-priori. This way the true model of evolution is known, and the selection techniques can be properly assessed for accuracy. Bryan Kolaczkowski wrote an in-house script called EVOL_E which was utilized to generate simulated sequence alignments under heterotachous conditions. EVOL_E works in much the same way as Seq_Gen(Rambaut and Grass, 1997) except that it allows for model heterogeneity among lineages as well as among sites. The transition/transversion ratio was modeled using JC69. Transitions are the rate of change between "A" & "G", and "C" & "T", while transversions are the rate of change between "A" or "G" to "C" or "T". The gamma rate distribution was not utilized, and instead all among site rate variation was allowed to be modeled using the heterotachous model.

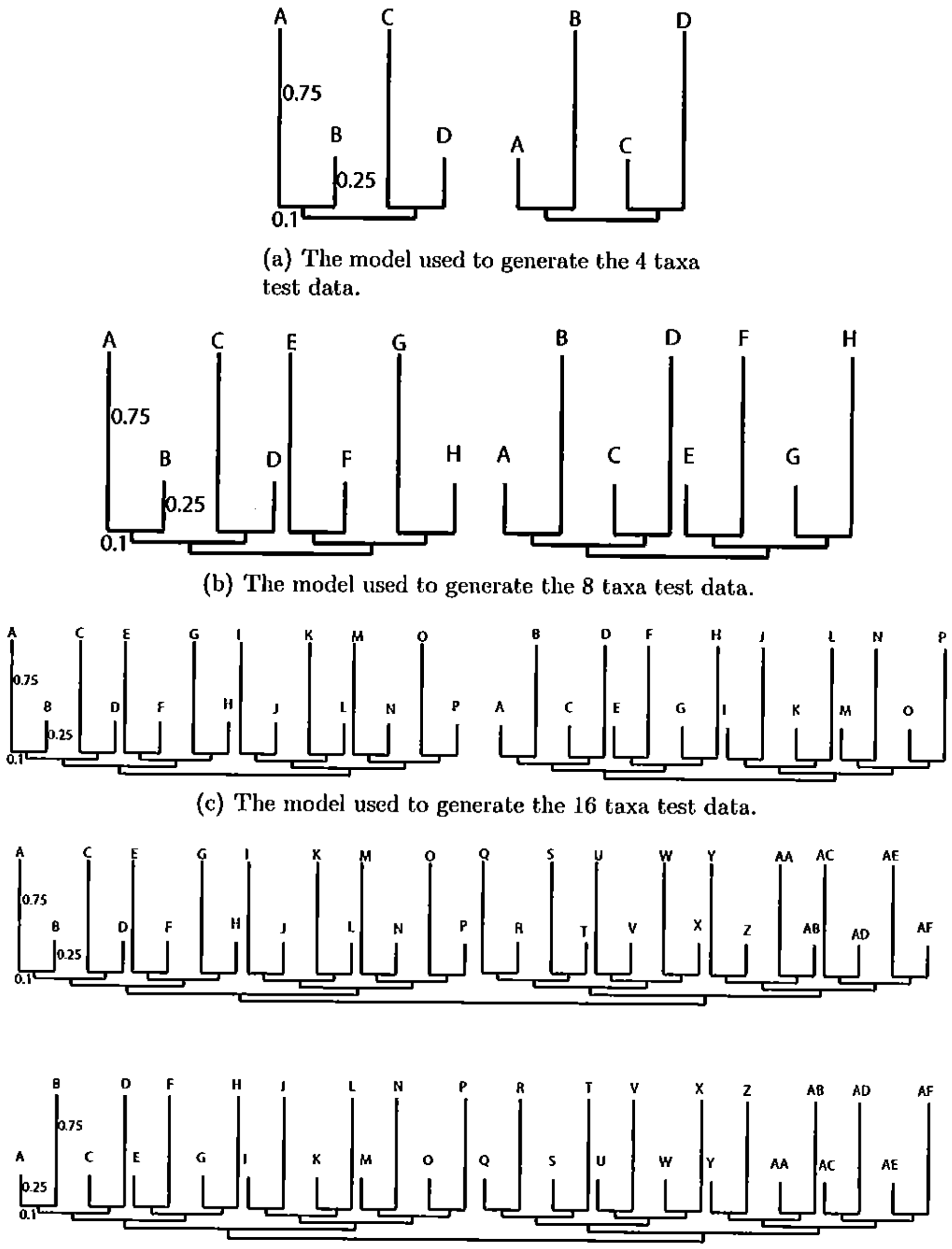
Two categories of experiments were performed. In the first, heterotachy was modeled using a two-rate category evolutionary tree where 50% of the sites were under one rate category or the other. Additionally the trees all were modeled with the same kind of heterotachy found to yield poor results if the MBL model is not used for phylogenetic reconstruction(Kolaczkowski and Thornton, 2008; Felsenstein, 1978), however the branch lengths were less extremely divergent, and the internal branch was significantly longer than the analysis done by Kolaczkowski and Thornton (2008).

The longer internal branch of 0.2 was in the zone where the mixed branch length model performs well, but other models still fail as shown in figure 3. Figure 5 shows the two branch length sets that were generated with the heterotachous simulated data utilizing EVOL_E given 4, 8, 16, and 32 taxa. Each set of branch lengths is applied to 50% of the sites in the alignment.

To determine which model selection technique performed best in the most conditions, the sequence length was varied in the alignments, and the number of taxa was also varied (as shown in Figure 5). 50 alignments were generated under each condition to be sure that the analysis was not biased by random error.

In the second main experimental procedure performed, the level of computational difficulty of the heterotachy was varied, and the internal branch length was very short (0.01 vs 0.2). The short internal branch makes evaluating heterotachy particularly difficult (Kolaczkowski and Thornton, 2008). Heterotachy was varied by varying the length of the long and short branches, and the proportion of sites that fell under one versus the other rate category. The short branch had a length between 0.0 and 0.5 with increments of 0.1, while the long branch had a length of 1 minus the length of the short branch. Similarly the proportions varied from 0.0 to 0.5 with increments of 0.1, and both proportions totaled to 1.0. The experimental procedure is depicted in figure 6.

3.2. **Data Analysis.** For the purposes of the first portion of this experiment, 1,2,3, and 4 evolutionary rate categories were assumed for each of the 50 alignments, and tested under each condition of sequence length and number of taxa. The various model selection techniques were utilized to choose the best set of assumptions. The results were compared with the true model, and examined for potential biases. *AIC*,



(a) The model used to generate the 4 taxa test data.

(b) The model used to generate the 8 taxa test data.

(c) The model used to generate the 16 taxa test data.

(d) The model used to generate the 32 taxa test data.

FIGURE 5. The models used to generate test data.

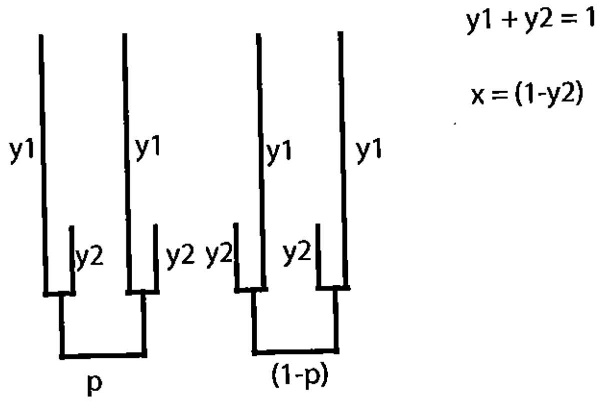


FIGURE 6. The experimental setup for the second experiment. $p$ represents the proportion of sites falling under the above branch length set. $y_1$ represents the length of the long branches, and $y_2$ the length of the short branches. $x = (1 - y_2)$ for readability purposes in the experiment shown in figure 10, since now when $x = 0$ the level of heterotachy is very light, while when $x = 0.5$ the level of heterotachy is much more extreme.

$AIC_c$, $BIC$, and $LRT$ were all performed on the log likelihood scores using Microsoft Excel. A custom program was written in python to perform the $CV$ analysis as described in the background section on $CV$. To run the $CV$ analysis as if a 2-fold $CV$ with 10 repetitions (Zhou et al., 2007) was performed, data was generated that was $\frac{1}{2}$ the original sequence length the following was performed. The program randomly chose 10 pairs of sequences, 50 times, from the pool of 50 $\frac{1}{2}$ length alignments. The ML tree was calculated from the first member of the pair, and the Likelihood of the second member of the pair was calculated from that tree. Then the log likelihoods generated through the 10 runs were averaged to give the $CV$ scores. The same heterotachous model of evolution, and the same evolutionary relationships were utilized for the full-length sequence as were for the $\frac{1}{2}$ length sequences.

Trends were plotted for the various model selection techniques in Microsoft Excel, and visually evaluated. In most cases the trends were very clear, and conclusions were drawn fairly easily.

For the second portion of the experiment 1,2, and 3 evolutionary rate categories were assumed for each of the 10 initial alignments, and tested under each combination of varied branch length and proportion of sites in one or the other model. For this very data intensive portion of the experiment, $AIC$, $AIC_c$ and $BIC$ were tested.

## 4. Results and Discussion

To make assumptions about the performance of the model selection techniques across an infinite number of possible conditions given a finite number of tests, one must examine the results for telling trends, and hypothesize that the trends more or less predict the untested conditions. Figure 7 shows the average number of rate categories selected with standard error bars as the sequence length increases given various model selection techniques. The general trend is that most model selection techniques become more accurate as the sequence length increases. The notable exception to this trend is $CV$ which starts off the closest to the correct generating model, two categories, and becomes more biased toward choosing an overly complex model of evolution as the sequence length increases. Additionally it can be plainly seen from the error bars that $CV$ also has the least certainty of all the techniques. Given that most researchers would not have the time or computing power necessary to re-run model selection many times over to be sure to account for the variance, $CV$

seems like a relatively poor choice especially given the apparent high performance, and low cost of the other techniques. Another notable piece of information presented is that $BIC$ is strongly biased toward selecting an overly simplified model of evolution at least when the sequence length is below 5000. There have been findings in the past that agree with this relative tendency of $BIC$ to select a more simple model of evolution than $AIC$, but in that case $AIC$ may have been selecting an overly complex model of evolution(Alfaro and Huelsenbeck, 2006). At least for selecting the number of evolutionary rate categories, $AIC$, $AIC_c$, and $LRT$ outperform $BIC$ and $CV$ in the case of four taxa under the computationally difficult Felsenstein zone like tree as described in the Methods section.

It is not clear from the empirical analysis of $AIC$ and $AIC_c$ which performs better. However the trend is quite clear that at low sequence length and low taxa, $AIC$ and $AIC_c$ both penalize too much and select an overly simplified model of evolution. Given that even $AIC$ over-penalizes at low sequence length and taxa, the additional penalty $AIC_c$ provides at short sequences no longer makes sense.

A more telling viewpoint of $CV$'s variance is perhaps its distribution of models selected under different techniques. Figures 8(a) and 8(b) show the model selected and the selection technique on the bottom axes, and the number of times that model was selected on the vertical axis. As can be seen, when the sequence length is 5000 (Figure 8(a)) all methods other than $CV$ and $BIC$ select the correct model every time. $CV$ has the unique distinction of being the most variable, and of selecting an overly complex model of evolution at all sequence lengths tested given 4 taxa.

The performance of the techniques as the number of taxa increases was examined to see if any interesting selection biases emerged. As shown in Figure 9 an interesting

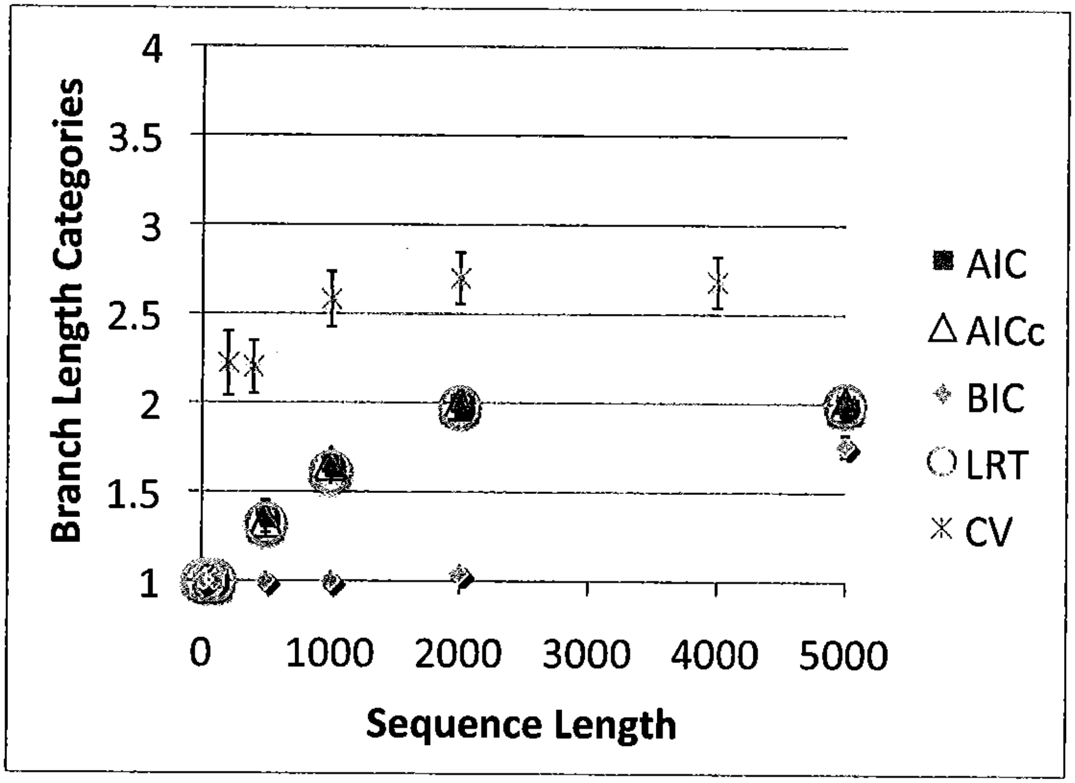
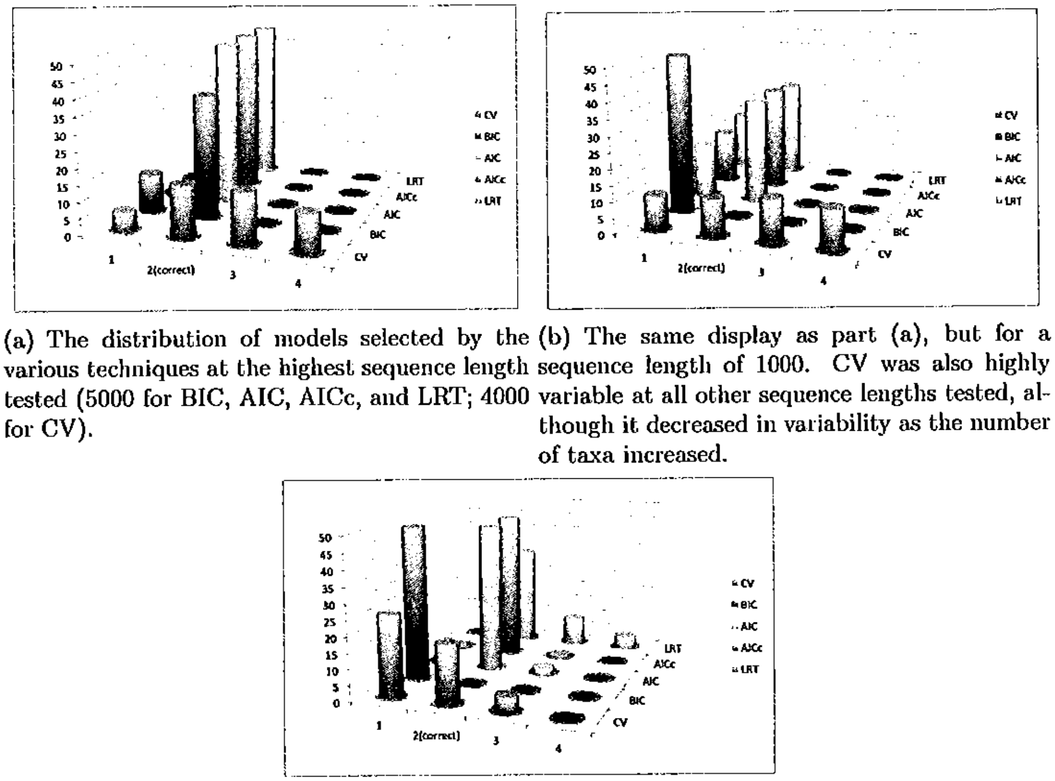


FIGURE 7. Results from the experiments testing the 2 rate category heterotachous model on four taxa as sequence length increases from 10 nucleotides to 5000 nucleotides. Error bars show standard error measurements. All the data for this graph was generated using the model of evolution posed in Figure 5(a).

pattern did emerge. It seems that *LRT* has a tendency to select increasingly complex models of evolution as the number of taxa increases and the sequence length remains fixed and relatively short at 500 nucleotides. Through the same trend $AIC$ and $AIC_c$ converge on the correct result, and $BIC$ remains overly conservative. Additionally $CV$ became less variable as shown in figure 8(c), and also, on average, converged on



(a) The distribution of models selected by the various techniques at the highest sequence length tested (5000 for BIC, AIC, AICc, and LRT; 4000 for CV).

(b) The same display as part (a), but for a sequence length of 1000. CV was also highly variable at all other sequence lengths tested, although it decreased in variability as the number of taxa increased.

(c) The same type of graph but a sequence length of 500 with 16 taxa.

FIGURE 8. The variance in the model selected by $CV$. In these figures the $z$ axis represents the simple count out of 50 trials that each value was observed.

the true model as shown in Figure 9. Still, the form of $CV$ tested did not perform as well as the computationally simpler technique of $AIC$, and $AIC_c$.

The form of $CV$ tested can be ruled out as an optimum technique due to its exceedingly high variance compared to other techniques (see Figures 7, 9, and 8) even though it performed slightly better at higher taxa (see Figure 9 and 8(c)). $CV$'s computational cost does not warrant its use for this problem given the higher

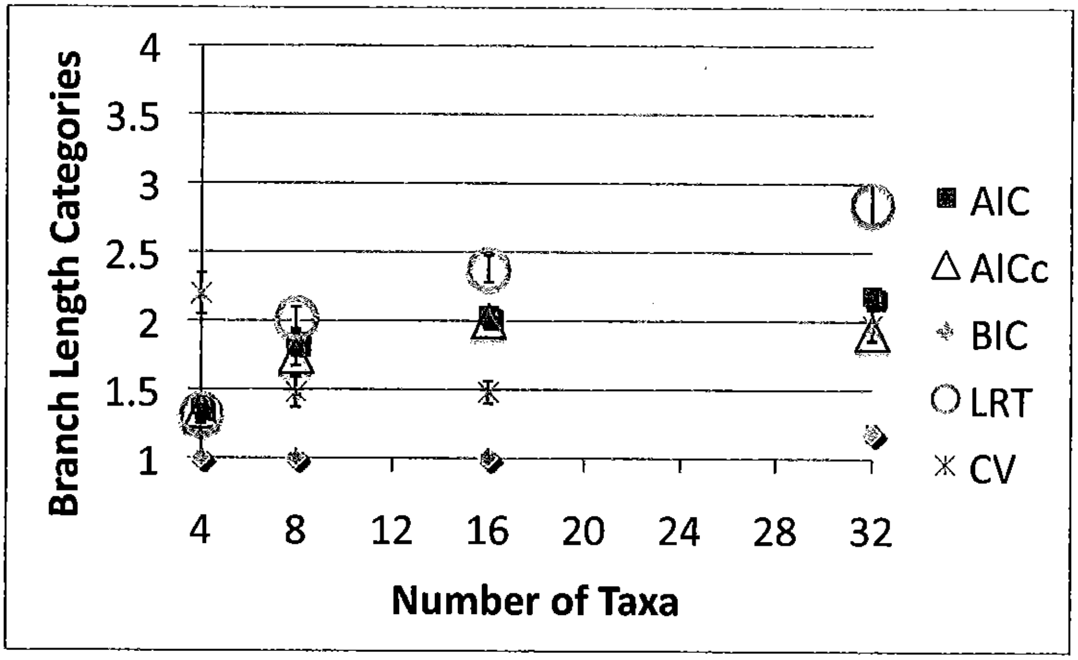
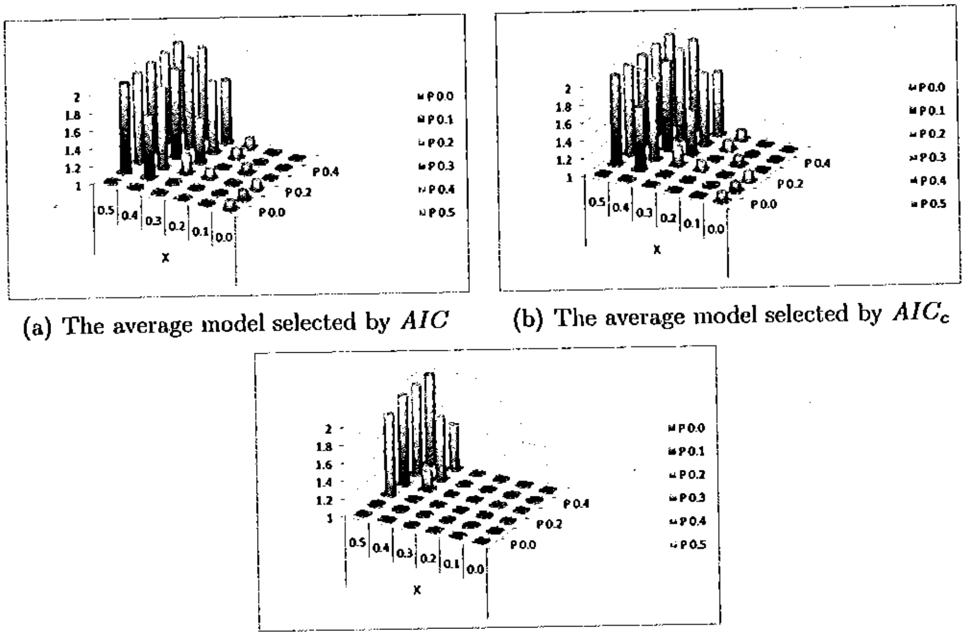


FIGURE 9. This figure examines the trend of model selected as the number of taxa are increased but the sequence length remains the same. The sequence length is fixed at 500 for all of these data-points. See Figure 5 to see the trees used to simulate data for each number of taxa on this graph. Note that in all cases 2 branch length categories is the correct value.

accuracy of the computationally simpler *AIC*. Additionally other forms of *CV* may perform better in general for the problem of selecting the number of branch length categories.

Given more time and resources, data with longer sequence lengths and higher numbers of taxa should also be explored. In these conditions would *LRT* have converged on the correct result like *AIC* and $AIC_c$ or would it have remained biased toward an overly complex model of evolution? Similarly would *BIC* have eventually converged on the correct result from being overly conservative? Also would *CV* exhibit less variation and converge on the correct result while following this trend? These are open questions that could be answered given more resources.



(a) The average model selected by *AIC*

(b) The average model selected by $AIC_c$

(c) The average model selected by *BIC*

FIGURE 10. p and x are derived as shown in figure 6. Note that in the above graphs, when x = 0 and p = 0 that represents the lightest form of heterotachy, while when x and p = 0.5 that represents the most extreme form of heterotachy. It is also important to note that the model used to generate this graph had a very short internal branch length of 0.005 rather than 0.1 as used in the rest of the experiments. This gives a more extreme case of heterotachy for which the correct model of evolution is theoretically more critical. In these graphs the $z$ axis is the average number of heterotachous partitions selected over 10 repetitions.

Figure 10 shows the results from the other main part of this experiment. In this analysis, $AIC$, $AIC_c$, and $BIC$ were compared to see which selects the correct model across the most diverse set of conditions described in figure 6 and 10. Note that when

X = 0.5, there is essentially no heterotachy because all branches from both rate sets are of length 0.5. Also note that there is essentially no heterotachy when P=0.0 because in that case all sites are generated by one of the two possible rate sets. This figure shows that under the most extreme forms of heterotachy, that is when X and P are high, $AIC$ and $AIC_c$ choose the correct model of evolution in more cases than does $BIC$.

## 5. Conclusion

Given the data examined, $AIC$ is the best candidate for a model selection technique to use for the problem of selecting the number of evolutionary rate categories for use in an MBL model. $AIC$ does not exhibit increasing biases toward overly complex models like $LRT$ does at shorter sequence lengths, and increasingly higher taxa(Figure 9). Also $AIC$ is not overly conservative like $BIC$(Figures 7, and 9). Additionally the $AIC$ does not have the variability problems that $CV$ has especially at lower taxa(Figures 8(a), and 8(b)). Also compared to $BIC$, $AIC$ selects the correct model across varying extremities in the computational complexity of the heterotachy tested (Figure 10).$AIC$ has an added bonus of being extremely computationally cheap if preliminary likelihood scores are already calculated. That in mind $AIC$ seems to be the best choice although more work should be done across a more diverse range of conditions to be sure of this finding.

$AIC$ and $AIC_c$ performed very similarly on the data tested. Since $AIC_c$ is correcting for a liberal bias at low sequence length, $AIC_c$ can be disregarded as the optimal technique for this problem. At low sequence lengths, $AIC$ does not have a tendency to select an overly complex model for this problem, in fact it selects an overly simplified model as seen in Figure 7. Thus at the low sequence lengths where $AIC_c$ makes a difference, it would actually bias $AIC$ further toward selecting an overly simplified model. For this reason $AIC$ is the optimal model selection technique for use in selecting the number of branch length categories for use in the mixed branch length model.

## 6. Future Work

This paper presents the examination of the effectiveness of various model selection techniques at determining the number of heterotachous partitions under one computationally difficult scenario. The work does not say for sure whether or not the same model selection techniques that performed well under the conditions tested would also perform well under other conditions such as varied sequence length and taxa, and/or more heterotachous partitions. To get a more thorough idea of how the techniques perform, one should examine other conditions as well, and see how the techniques perform in those circumstances. Although it would be important to evaluate a wider range of evolutionary conditions, more weight should always be placed on models shown to lead to phylogenetic error if an incorrect number of evolutionary rate categories were chosen. Thus any future analyses of the performance of model selection techniques should also include an analyses of the performance of the selected model at uncovering the true phylogeny as well.

Another important future research project is to check which model selection techniques lead to a selected model of evolution that generates the true phylogeny in the majority of all cases. The analyses presented were concerned with selecting the generating model of evolution, rather than determining whether the model of evolution chosen, if incorrect, was at least good enough to generate the true phylogeny.

Since this analysis was performed on a very specific case of heterotachy with long branches of 0.75 and short ones at 0.25, it would be important to examine the accuracy of the mixed branch length model while varying the length long and short branches. Additionally the majority of the analysis was performed with data generated using an equal proportion of all sites falling under each rate category. It be important to examine more data in which the proportion of sites falling under the rate categories is varied, and examine the accuracy of the model selection techniques.

Despite the fact that $CV$ performed very poorly when the implementation as described by Zhou et al. (2007) was used, the analysis should be re-run using different proportions of sites in the training set versus the test set. In this paper, only half of the sequence was used to approximate parameters that are applied to the remainder of the sequence. It is possible that if the proportion of sites in the training set were greater, and test set smaller, that CV would perform more optimally.

Finally the sequence length was varied with the number of taxa fixed at four, and the number of taxa was varied with the sequence length fixed at 500, but more work should be done with further variation of sequence length and taxa to see if any other patterns turn up. For example many model selection techniques choose an overly complex model of evolution as the number of taxa grew; would these methods converge on the correct answer as the sequence length grows as well, or would the greater sequence length cause the techniques to choose an even more complex model?

Regardless of the model selection technique that ultimately proves to be the best fit for the MBL model, the MBL model of evolution is an important development in phylogenetic analysis. The added accuracy of the model provides Maximum Likelihood with a much more realistic model of evolution to work with, pushing ML analyses closer to the goal of statistical consistency. If more work were put into optimizing a maximum likelihood package incorporating the MBL model, it should surely become one of the more important tools available to phylogeneticists.

## 7. Acknowledgments

Project advisors: Joe Thornton, and John Conery.

Project mentor: Bryan Kolaczkowski

## References

Akaike, H. (1974, Jan). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*(6), 716 – 723.

Alfaro, M. and J. Huelsenbeck (2006, Jan). Comparative performance of bayesian and aic-based measures of phylogenetic model uncertainty. *Systematic Biology.*

Bryant, D., N. Galtier, and M. Poursat (2005). Likelihood calculation in molecular phylogenetics. *Mathematics of Evolution and Phylogeny*, 33–62.

Farris, J. (1982). The logical basis of phylogenetic analysis. *Conceptual Issues in Evolutionary Biology. MIT Press, Cambridge, MA*, 675–702.

Felsenstein, J. (1978, Jan). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology.*

Felsenstein, J. (2003). *Inferring Phylogenies. Sunderland, MA.* Sinauer Press. Chapters.

Foster, P. G. (2001, Jul). The idiot's guide to the zen of likelihood in a nutshell in seven days for dummies, unleashed. (July 28), 8.

Gibbs, R. A., J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock, E. R. Mardis, K. A. Remington, R. L. Strausberg, J. C. Venter, R. K. Wilson, M. A. Batzer, C. D. Bustamante, E. E. Eichler, M. W. Hahn, R. C. Hardison, K. D. Makova, W. Miller, A. Milosavljevic, R. E. Palermo, A. Siepel, J. M. Sikela, T. Attaway, S. Bell, K. E. Bernard, C. J. Buhay, M. N. Chandrabose, M. Dao, C. Davis, K. D. Delehaunty, Y. Ding, H. H. Dinh, S. Dugan-Rocha, L. A. Fulton, R. A. Gabisi, T. T. Garner, J. Godfrey, A. C. Hawes, J. Hernandez, S. Hines, M. Holder, J. Hume, S. N. Jhangiani, V. Joshi, Z. M. Khan, E. F. Kirkness, A. Cree, R. G. Fowler, S. Lee, L. R. Lewis, Z. Li, Y.-S. Liu, S. M. Moore, D. Muzny, L. V. Nazareth,

D. N. Ngo, G. O. Okwuonu, G. Pai, D. Parker, H. A. Paul, C. Pfannkoch, C. S. Pohl, Y.-H. Rogers, S. J. Ruiz, A. Sabo, J. Santibanez, B. W. Schneider, S. M. Smith, E. Sodergren, A. F. Svatek, T. R. Utterback, S. Vattathil, W. Warren, C. S. White, A. T. Chinwalla, Y. Feng, A. L. Halpern, L. W. Hillier, X. Huang, P. Minx, J. O. Nelson, K. H. Pepin, X. Qin, G. G. Sutton, E. Venter, B. P. Walenz, J. W. Wallis, K. C. Worley, S.-P. Yang, S. M. Jones, M. A. Marra, M. Rocchi, J. E. Schein, R. Baertsch, L. Clarke, M. Csuros, J. Glasscock, R. A. Harris, P. Havlak, A. R. Jackson, H. Jiang, Y. Liu, D. N. Messina, Y. Shen, H. X.-Z. Song, T. Wylie, L. Zhang, E. Birney, K. Han, M. K. Konkel, J. Lee, A. F. A. Smit, B. Ullmer, H. Wang, J. Xing, R. Burhans, Z. Cheng, J. E. Karro, J. Ma, B. Raney, X. She, M. J. Cox, J. P. Demuth, L. J. Dumas, S.-G. Han, J. Hopkins, A. Karimpour-Fard, Y. H. Kim, J. R. Pollack, T. Vinar, C. Addo-Quaye, J. Degenhardt, A. Denby, M. J. Hubisz, A. Indap, C. Kosiol, B. T. Lahn, H. A. Lawson, A. Marklein, R. Nielsen, E. J. Vallender, A. G. Clark, B. Ferguson, R. D. Hernandez, K. Hirani, H. Kehrer-Sawatzki, J. Kolb, S. Patil, L.-L. Pu, Y. Ren, D. G. Smith, D. A. Wheeler, I. Schenck, E. V. Ball, R. Chen, D. N. Cooper, B. Giardine, F. Hsu, W. J. Kent, A. Lesk, D. L. Nelson, W. E. O'brien, K. Prufer, P. D. Stenson, J. C. Wallace, H. Ke, X.-M. Liu, P. Wang, A. P. Xiang, F. Yang, G. P. Barber, D. Haussler, D. Karolchik, A. D. Kern, R. M. Kuhn, K. E. Smith, and A. S. Zwieg (2007, Apr). Evolutionary and biomedical insights from the rhesus macaque genome. *Science 316*(5822), 222–234.

Higgs, P. and T. Attwood (2005). *Bioinformatics and molecular evolution.* Blackwell Oxford.

Inagaki, Y., E. Susko, N. M. Fast, and A. J. Roger (2004, Jul). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in ef-1alpha phylogenies. *Molecular Biology and Evolution 21*(7), 1340–9.

Kolaczkowski, B. (2004, Jan). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*.

Kolaczkowski, B. and J. W. Thornton (2008, Jun). A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution 25*(6), 1054–66.

Rambaut, A. and N. Grass (1997, Jan). Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic .... *Bioinformatics*.

Schwarz, G. (1978, Jan). Estimating the dimension of a model. *The annals of statistics*.

Shono, H. (2000, Jan). Short paper efficiency of the finite correction of akaike's information criteria. *Fisheries Science*.

Spencer, M., E. Susko, and A. J. Roger (2005, May). Likelihood, parsimony, and heterogeneous evolution. *Molecular Biology and Evolution 22*(5), 1161–4.

Wilks, S. (1938, Jan). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*.

Zhou, Y., N. Rodrigue, N. Lartillot, and H. Philippe (2007, Jan). Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evolutionary Biology 7*(1), 206.