Phylogeny of Five Taxa in the Felsenstein and Farris Zones

Eric Trung Lam

Thesis submitted to the Faculty of Science in partial fulfillment of the requirements
for the degree of
Masters in Statistics with Specialization in Bioinformatics[1]

Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Mathematical conditions which showed where parsimony was not consistent for four taxa were first introduced by Felsenstein in 1978. This was subsequently labelled the "Felsenstein zone". Following Felsenstein's findings, 'frequentists' conjectured that for five taxa there would also be a region in parameter space where parsimony is not consistent. In response, 'cladists' claimed that parsimony was consistent in a different region of parameter space, which is called the "Farris zone". However, no analytical description of the region in which this consistency occurs has been made. Furthermore, no mathematical extensions of this Felsenstein theory to five taxa or more has been made. The same is true for the Farris zone. In this thesis, we give a complete account for the Felsenstein zone and Farris zone for four and five taxa and interpret these in terms of the shape of the phylogenetic tree.

Dedications

I would like to dedicate this to my parents and sister, whose love and support has helped me throughout my university career.

Acknowledgements

Firstly, I would like to express my thanks and gratitude to my supervisor, David Sankoff, without whom this thesis would not have been completed. I cannot thank him enough for giving me the opportunity to join his lab and gain invaluable experience and knowledge.

I would also like to thank everyone at the Sankoff Lab. Everyone there has been a wonderful mentor, colleague, and friend. I wish everyone success in their academic endeavours.

Lastly, I would like to thank my wonderful girlfriend, Julie, who has been nothing but supportive during the whole process of this thesis by providing a plethora of baked goods for motivation. Our cat, Yzma, has also been very supportive by attending my Zoom calls and keeping me company while I wrote, without feeling the need to step all over my keyboard.

Contents

List of Figures

Chapter 1

Introduction

Parsimony is one of several of criteria used in phylogenetic inference to infer the topology and branch lengths of a phylogeny, or species tree. In biological systematics, a version of parsimony was introduced by Camin and Sokal in 1965 [1], and the basic principles were outlined by Farris in 1970 [5] and Fitch in 1971 [10]. Parsimony infers evolutionary phylogenetic trees from discrete character data, obtains a reconstruction of the evolutionary changes in a given set of characters on a given tree, and counts the minimum amount of times that character changes must have occurred on that tree to account for the data. It then uses this minimum evolution as the measure of the adequacy of the evolutionary tree. The idea is to find the evolutionary tree which requires the fewest evolutionary changes to explain the observed data. This is called the most parsimonious tree.

Camin-Sokal parsimony [1] is perhaps the simplest method for parsimony, it assumes for each character that the ancestral state is known, and also assumes that character states are binary [6]. However, character state changes will only occur from 0 to 1, and cannot revert to state 0 [1]. The characters are also assumed to

have evolved independently, but according to the same evolutionary processes [6]. Counting the amount of character state changes is an easy task since a node with its immediate descendants in state 0 must also be in state 0 [6]. Otherwise, if any immediate descendants are in state 1, then the node is also in state 1. In contrast to this, Farris's method for inferring unrooted Wagner trees [5] is under the assumption that these character state changes are reversible.

Other than parsimony, there are a variety of reconstruction methods. In many of these, the species (or "taxa") and the ancestors are represented by points or nodes in a metric space. The unweighted pair group method with arithmetic mean (UPGMA), is a simple clustering algorithm which assumes that there is a constant rate of evolution (i.e. a molecular clock assumption), and quickly produces a tree phylogeny. Trees which are considered "clock-like" are rooted, meaning the tree includes a node which represents the most recent common ancestor acting as the parent node for all other nodes [8, p. 161]. These clock-like trees, or often referred to as "ultrametric", require that the total branch length from the root be the same for all species in the tree. A distance matrix D is ultrametric if and only if the "three-point condition" is satisfied, that is for any three taxa, we can label them i, j, k such that

$$D(i,j) \leq \max\{D(i,k), D(k,j)\} \tag{1.0.1}$$

Many other methods are also based on a distance matrix D comparing each pair of species. Instead of evaluating each character separately, these methods start with an overall distance or dissimilarity between species.

If evolution proceeds at an uneven pace, so that the rate of evolution may be accelerated in one branch and depressed in another, the data may violate the basis of the UPGMA method and may lead to the inference of the wrong tree topology

[23]. Another distance-based method, neighbor-joining (NJ) [20], is also popular for phylogenetic reconstruction. Neighbor-joining is also a clustering method, but does not assume a molecular clock and instead approximates the minimum evolutionary distance [20]. This method does not require the molecular clock assumption, but still needs the data to have some metricity, in this case, additivity. Pairwise distances are called additive if they satisfy the "four-point condition":

$$D(i,j) + D(m,n) \leq \max\{D(i,m) + D(j,n), D(j,m) + D(i,n)\}, \tag{1.0.2}$$

where i, j, m, n represent species on a tree, and $D(i,j)$ is the distance between species i and j [11, p. 456]. The four-point condition is a generalization of the triangle inequality (e.g. take $m = n$) and a distance matrix can be well represented by a phylogenetic tree if and only if it is additive, or approximately so [11, p. 456]. In many cases however, the data may seriously violate this four-point rule. Moreover, neighbor-joining may obtain negative branch lengths [23].

Traditional statistical methods are also used to estimate phylogeny. Perhaps the most widely used is maximum likelihood (ML) [2]. Maximum likelihood methods for estimating phylogeny make use of the already existing statistical framework of maximum likelihood. Edwards and Cavalli-Sfroza (1964) [4, pp. 67–76] introduced ML methods for estimating phylogenies for gene frequency data. Simply, given a set of data D, ML methods for phylogenetic inference will choose the hypothesis H which maximizes the likelihood function for D [4, pp.67-76]. Tuffley and Steel (1997) [28] defines the likelihood method as follows: "*Given a set* $X = \{\chi_i\}$ *of* K *r-state characters, likelihood chooses the unrooted tree and vector pair or pairs* (T, p)

maximizing

$$\mathcal{L}[(T,p) \mid X] = \Pr[X \mid T,p] = \prod_{i=1}^{K} \Pr[\chi_i \mid T,p]''. \tag{1.0.3}$$

The tree(s) inferred by this maximization is the maximum likelihood tree(s). One benefit of using the statistical framework set by likelihood methods is that it is consistent, meaning that as the amount of data increases, the maximum likelihood estimate becomes more likely as it converges to the true value [8, p. 249].

In contrast to distance-based methods, ML is able to explain character distributions as it requires an explicit model of character evolution with associated probabilities for each transformation from one character state to another [2]. On the other hand, parsimony can explain character distributions across the taxa of what is minimally required by the observed data [23].

Although it is one of the simplest and earliest methods for inferring phylogenies, the use of parsimony has been heatedly debated [25] between its proponents [23] and those who advocate for maximum likelihood. While parsimony looks to find the tree topology that requires the fewest changes in character state to explain the character distribution across taxa, ML looks to find the tree topology with the highest probability of observing the data given a model [7]. Both methods provide an ordering of tree topologies from best to worst given the data set. However, authors like Siddall (1998) [23] and Steel and Penny (2000) [26] have noted that parsimony and maximum likelihood will sometimes disagree on which tree is the best supported by the evidence, given that the tree model used to create the observed data is different than the model used in the ML inference. Siddall's study (1998) [23] suggests that parsimony recovers the correct tree more than 95% of the time, while ML methods may only have a recovery rate of 30% for the correct tree 50% of the time. Arguments for and against

parsimony and ML have been made ever since the 1960s and the advocates for both parsimony and ML claim superiority for their method. One of the main objections to ML is that it requires the user to adopt an evolutionary model process that they believe to be true, even if that assumption may be unrealistic [25]. Others have also noted that ML is sensitive to changes in the observed data and instead of searching for the most economical explanation for the data, it searches for the most probable estimate [23]. On the other hand, those who oppose parsimony make the claim that it fails to take into account any knowledge, such as mutational mechanism [23]. The supporters of parsimony and likelihood have since then separated themselves into their self-proclaimed groups: cladists and frequentists respectively [25].

The most infamous problem that plagues parsimony is that it may perform poorly when unrelated taxa have high rates of change. Under this condition known as 'long-branches attract', parsimony may not lead to a consistent estimate and fail to converge to the true tree [6]. It was Felsenstein who discovered the mathematical constraints on tree branch lengths in the region where parsimony is not consistent in the case of three and four taxa, and a number of papers have built upon this idea set out by him [6]. These studies [14], [15], [16], [22], [23], [31], [32] have aptly dubbed certain topologies where parsimony is not consistent as 'Felsenstein-type trees' and certain regions where parsimony fails to converge to the correct tree, as a 'Felsenstein zone'. Of particular interest here, following Felsenstein's 1978 seminal paper, Zharkikh and Li [32] have shown analytically that parsimony is not consistent in the case of five taxa under a molecular clock assumption. However, there still is no publication outlining any analytical conditions for which parsimony is not consistent in the case of five taxa under a Felsenstein-type tree when two unrelated taxa have elevated rates of evolution.

In Felsenstein's original work, he illustrated with accelerated evolution in two non-sister taxa. On the other hand, it has been shown [16] that when it is sister taxa that have high rates of change in a topology called a 'Farris-type tree', parsimony is consistent in the case of four taxa. As counterpoint to the Felsenstein zone, this space of trees is called the 'Farris zone' [23]. However, studies on the Farris zone [23], [26], [29], [30] have not found mathematical constraints on tree branch rates in the same way Felsenstein did to find regions where parsimony is not consistent. They instead generate data from nucleotide substitution models (JC69 [18], K80 [19], HKY85 [13]) for each possible probability of transforming one character state to another, and apply different character change rates in computer simulations to find the proportion of correctly estimated trees.

In this thesis, the consistency of the four and five taxa cases in both Felsenstein and Farris-type trees is investigated. In Chapter 2, Farris' unrooted Wagner tree, which is the same bifurcating tree model used in Felsenstein's paper [6], is introduced. This will lay the foundation of how parsimony will be represented for phylogenetic trees.

It is important to remark that, for example, a four-taxa tree has four independent dimensions, and authors such as Yang (1996) [30] and Siddall (1998) [23] utilise a four-valued state space to develop their simulation studies, more relevant for DNA data. However, for the purposes of this study, we are only looking at two dimensions with parameters P and Q for the sake of showing how to determine boundaries between conditions in which parsimony may or may not be consistent within a certain set of tree topologies. Next, an outline of Felsenstein's method [6] to find the conditions for when parsimony is not consistent is summarized. Thereafter, in Chapter 3, since no formal analytical description or conditions have been discussed about the

consistency of parsimony in a Farris zone, we apply Felsenstein's method for finding the constraints for branch lengths to prove that parsimony is consistent in a Farris-type tree, as conjectured by the simulation results by Siddall [23]. Lastly, in Chapter 4, since no analytical conditions have been extended for greater than four taxa, the same approach is applied again to show the constraints and the regions of consistency of five taxa in Felsenstein and Farris-type trees. The result found in a five-taxa Felsenstein-type tree are analogous to the result of Felsenstein's four-taxa study in 1978 [6], and the result found in a five-taxa Farris-type tree extends and proves the consistency of parsimony suggested by the simulation of four taxa in the Farris zone done by Siddall [23]. In all cases, new analytical conditions for the branch rates are found and these conditions are visualized in a plot of the consistency in parameter space.

Chapter 2

Investigating Consistency in Felsenstein-Type Trees in the Case of Four Taxa

2.1 Parsimony for Phylogenetic Trees

The principle of parsimony leads into two sub-problems: *(i) the small parsimony problem*, which is how to calculate the amount of character change, i.e., the total length of the branches on a given tree, and *(ii) the large parsimony problem*, which is how to find the tree that minimizes this length over all possible topologies [8, p. 9]. The small problem admits to rapid solution algorithms, but in the large problem, searching among all possible trees for the most parsimonious ones is hard [3].

To count the number of state changes on a given phylogeny, four similar dynamic programming algorithms are most often mentioned: the Farris algorithm (1970) [5], the Fitch algorithm (1971) [10], the Hartigan algorithm (1973) [12], and the Sankoff-

Rousseau algorithm (1975) [21]. These dynamic programming algorithms, given n species and k character states, have $\mathcal{O}(nk)$ complexity runtime and evaluate a phylogeny character by character. Each character is considered on the same rooted tree, i.e., the tree has a parent node which is the most recent common ancestor for all other nodes. Then, as information is updated down the tree, the total number of state changes for the tree is found at the bottom (the most recent species) [8, pp. 11-18]. While these are efficient exact algorithms for the small parsimony problem, the large parsimony problem, i.e., finding the best tree topology, is computationally hard and has no efficient solution since it is NP-complete [3]. The most obvious and straightforward method would be to consider all possible trees, evaluating each, keeping a list of the most parsimonious trees, and if any trees are more parsimonious than the previous, start a new list based on that new tree [8, p. 19]. Rather than searching through all possible trees, greedy methods such as a "heuristic search", use an *ad hoc* approach by making an initial estimate of the tree, and commits to this search path by making small branch rearrangements to reach neighbouring trees [8, pp.37-55]. At the end, the search reaches a tree where no rearrangements can improve the current tree. While this search is vastly more time efficient by ignoring a majority of possible trees, there is no guarantee that the final tree is a global maximum and not stuck at a local optimum [8, pp. 37-38]. There are multitudes of other heuristic search methods such as nearest neighbour interchange, local searches, or a divide and conquer approach like "branch-and-bound", which may take exponential time and is only feasible for 12-25 taxa [8, pp. 59-64].

2.1.1 Statistical Consistency in Parsimony

In statistics, an estimator is called consistent if it converges in probability towards the true value as the amount of data collected grows. As mentioned, parsimony is not statistically consistent as it is not guaranteed to produce the true tree with high probability given sufficient data [22]. Felsenstein [6] first showed that parsimony methods, such as Camin-Sokal, Farris' unrooted Wagner tree method, and the compatibility method, are not consistent under the particular case of the Felsenstein zone. He outlined the mathematical conditions for consistency of parsimony on the branches of three and four species tree with binary character states. In his paper, he computes the probabilities of various character configurations for the possible unrooted bifurcating phylogenies. As more characters are scored, he can predict which of the bifurcating trees parsimony will converge to as more data is accumulated [6]. Felsenstein [6] showed that all of the methods he studied will fail to converge to the true tree when unrelated external branches (see Figure 3.1) have a higher probability of sharing the same character state than the internal branches.

This phenomenon called 'long-branches attract' is of great importance and has been widely discussed in phylogenetic inference literature [6], [14], [22], [23], [27], [32]. Many of these papers [6], [15], [23], [24], [27] study the case when parsimony is not consistent in the Felsenstein zone. In this case, parsimony tends to group unrelated taxa together as a sister group when estimating tree phylogenies. In all of these studies [6], [32], [27], the authors all conclude that when there is a sufficient disproportion between long and short segments of the tree branches, parsimony methods tend to converge to the wrong tree as more and more data is collected. They [6], [22], [27], [32] direct their readers to the idea that to avoid the systematic errors in estimating a phylogeny from parsimony, selecting slowly evolving sequences will alleviate the

problem of 'long-branches attract'.

The rest of this thesis will be an investigation of consistency in Felsenstein and Farris-type trees in the case of four and five taxa. As no formal analytical constraints have been made about consistency in the Farris zone, a formulation will be presented in Section 3.1. Additionally, no studies have extended any analytical conditions beyond four taxa. Hence, formulations for consistency in five-taxa Felsenstein and Farris-type trees will be developed in Chapter 4.

2.1.2 Unrooted Bifurcating Trees

To study the consistency of parsimony in the case of four taxa, Felsenstein [6] utilized Farris' unrooted tree model [5]. In this model, phylogenetic trees are unrooted and bifurcating, meaning that they are connected, undirected graphs with no cycles, and with three edges emerging at each internal node. He takes advantage of the model's simplicity to find the most parsimonious tree which allows evolutionary state reversals. The following is the characterization of unrooted bifurcating trees in Felsenstein's paper [6].

Definition 2.1.1. *Let* $X = \{A, B, C, D\}$ *represent the set of present-day taxa, and the unrooted evolutionary tree,* T*, on* X*. Each present-day taxa has two character states,* 0 *or* 1*, where* 0 *is the ancestral state, and* 1 *is the derived state. Assume that the character state changes are reversible, i.e., it is possible for the state of the population to change from* 0 *to* 1*, and it is possible to revert from state* 1 *to state* 0

Now, suppose that the four observed species, A, B, C, and D, and that the true (but unknown) phylogeny is given by Figure 2.1. Furthermore, suppose that we know for each branch in the tree, not its length, but the net probability that the character

will change in that branch. Next to each branch in Figure 2.1, is the net probability of change along it, P and Q. Note that P and Q cannot be greater than 0.5, since the chance that a state at the end of a branch is different than the state at the beginning is only $1/2$, even if we had an arbitrarily long branch [8, p. 111]. Furthermore, since character state changes are assumed to be reversible, namely the characters that were originally in state 0 changing to state 1, have the same probability as reverting from state 1 to state 0 [6].

Figure 2.1 is constructed with longer branch lengths, which indicate a faster rate of evolution, whereas the shorter branch lengths are slower [6]. Additionally, the root placement does not need to be specified since the probability of any character pattern on the tree of Figure 2.1 is the same no matter where the root is placed, by the symmetry of the tree [8, p. 111].

Following Felsenstein, our exposition is phrased in terms of probabilities. We could as well have developed this work in terms of evolutionary time. In the simplest case, time t would satisfy

$$P = \lambda e^{-\lambda t} + \frac{1}{2}(1 - \lambda e^{-\lambda t}) \tag{2.1.1}$$

for some rate parameter λ. Monotonicity ensures that the two ways of looking at things are equivalent.

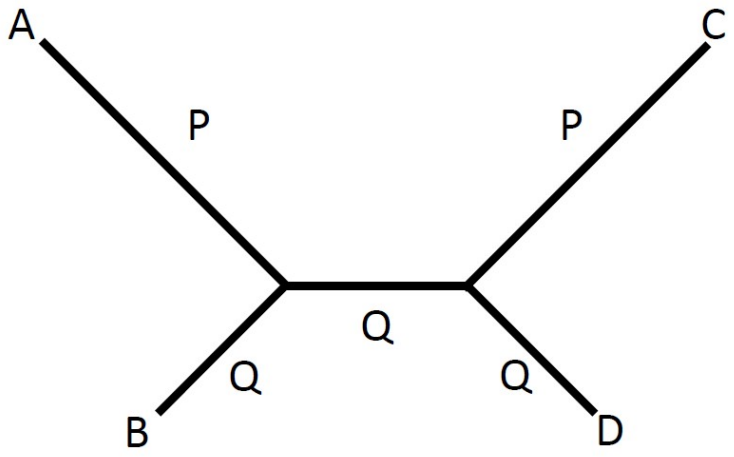

Figure 2.1: The true unknown phylogeny in the case of four taxa in a Felsenstein-type tree. Unrelated taxa, A and C, separated by an internal branch, have longer branches lengths, indicating an accelerated rate of evolution, while shorter branch lengths indicate a slower rate. Next to each branch is the net probability of character change along it, P or Q.

2.2 Felsenstein's Example with Four Taxa Shows Parsimony is not Consistent

Assume that the N characters in these four species have evolved independently (i.e. sampled independently) and all have the same probabilities. In the case of four taxa, we can count how many characters are in each of the $2^4 = 16$ possible combinations: $0000, 0001, \ldots, 1111$ with probabilities $P_{0000}, \ldots, P_{1111}$. These probabilities are given for the ordered set of species P_{ABCD}. The counts of these characters are defined by $n_{0000}, n_{0001}, \ldots, n_{1111}$ and are drawn from a multinomial distribution. Felsenstein applies the Strong Law of Large Numbers so that as $N \to \infty$,

$$\frac{n_{ijkl}}{N} \to P_{ijkl} \tag{2.2.1}$$

for all combinations $\{i, j, k, l\} \in \{0, 1\}$ [6]. For an unrooted tree with four taxa, there are only three possible bifurcating phylogenies that represent unique configurations, (AB)(CD), (AC)(BD), and (AD)(BC), where sister taxa groups are separated by parentheses. For example, the Felsenstein-type tree in Figure 2.1 is represented by the phylogeny (AB)(CD) since taxa A and B are grouped as sister taxa on the left, while taxa C and D are sisters on the right.

Consider the unrooted phylogenetic tree in Figure 2.1 and suppose that this is the true, but unknown, phylogeny. Moreover, recall from beginning of this section that the N characters have evolved independently and that there is a probabilistic model of evolutionary change of the characters. From this model, one can calculate the expected frequencies of each of the 16 character patterns. In this simplified case with binary character states and a symmetric model of evolutionary change, a formula

can be written for the probability that a character will change states given the branch length. With this tree and probabilistic model of character change, we can calculate the frequency that each configuration is expected to be seen.

Here, we give an example of how to calculate the probability of obtaining the pattern 1100 in the case of four taxa. Figure 2.2 illustrates this scenario, where the branches are labelled with the probability of change, P and Q, from $0 \to 1$ along each branch. For the pattern 1100, we can start at the left interior node of the tree. There are four possible ways we can assign states to both the left and right interior nodes. These combinations are 00, 01, 10, and 11. For the probability of pattern 1100, we can sum the probabilities of the four possible ways of obtaining this pattern. Starting at the left interior node in Figure 2.2, the probability that the node is in state 0 is $1/2$, by symmetry of the model tree. The probability of a change from 0 to 1 on the upper-left branch and lower-left branch is P and Q respectively, and the probability of no change on the interior branch between interior nodes is $(1 - Q)$. Likewise, given that the state of the right interior node is 0, the probability of no change on the upper-right branch is $(1 - P)$, and the probability of no change on the lower-right branch is $(1 - Q)$.

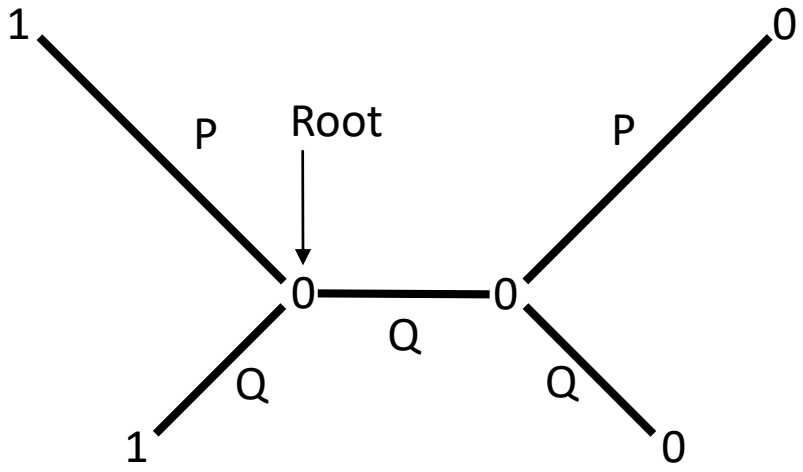

Figure 2.2: One scenario for the pattern 1100 on a Felsenstein-type tree for four species. The branches of the tree are labelled with the net probability of character change along it, P or Q, which indicates the probability of change from state 0 to 1. Starting at the root, which is in state 0, there is a change in both left branches of the tree. There is no change between internal nodes, and no change in both the right branches of the tree.

Since independence of evolutionary processes in different lineages and in different segments of the same lineage is assumed, we can multiply the above probabilities to get

$$\frac{1}{2}PQ(1-Q)(1-P)(1-Q) \tag{2.2.2}$$

For the three other combinations of states on the interior nodes that we could have obtained for this pattern, we can derive probabilities in similar way to obtain the following probability of 1100.

$$P_{1100} = \frac{1}{2}[PQ(1-Q)^2(1-P)+(1-Q)^2(1-P)^2Q+P^2Q^3+(1-P)(1-Q)^2PQ] \tag{2.2.3}$$

Now that we have the probability for the pattern 1100, there is also the pattern 0011.

Together, these make up the class of combinations *xxyy*, where x and y are any two numbers [6]. By the symmetry of the model, both patterns have equal probabilities, so the total probability of the pattern *xxyy* can be obtained by doubling equation 2.2.3, which will remove the 1/2.

In the same manner, we can do the same for the class of patterns *xyxy* and *xyyx*. The following are total probabilities for all of the class of patterns in this model.

$$\begin{aligned} P_{1100} + P_{0011} &= PQ[(1-Q)^2(1-P) + Q^2P] \\ &\quad + (1-P)(1-Q)[Q(1-Q)(1-P) + Q(1-Q)P] \\ &= Q^3 + 2P^2Q^2 - 2Q^2 - P^2Q + Q, \end{aligned} \tag{2.2.4}$$

$$\begin{aligned} P_{1010} + P_{0101} &= P(1-Q)[Q^2(1-P) + (1-Q)^2P] \\ &\quad + (1-P)Q[Q(1-Q)P + Q(1-Q)(1-P)] \\ &= -Q^3 + 2P^2Q^2 + Q^2 - 3P^2Q + P^2 \end{aligned} \tag{2.2.5}$$

and

$$\begin{aligned} P_{1001} + P_{0110} &= P(1-Q)[Q(1-Q)P + Q(1-Q)(1-P)] \\ &\quad + Q(1-P)[(1-Q)^2P + Q^2(1-P)] \\ &= Q^3 + 2P^2Q^2 - 4PQ^2 - P^2Q + 2PQ \end{aligned} \tag{2.2.6}$$

The total probabilities $P_{1100}+P_{0011}$, $P_{1010}+P_{0101}$, and $P_{1001}+P_{0110}$ represent the configurations (AB)(CD), (AC)(BD), and (AD)(BC) respectively. As more characters are scored, the unrooted tree configuration that is obtained will be determined by which of these three total probabilities is largest. Assuming that Figure 2.1 is the true (but unknown) tree, it is presumed that configuration (AD)(BC) will most likely have the lowest expected frequency of occurring since taxa A and D are separated by an internal branch. Starting with equations 2.2.5 and 2.2.6 we can simplify and compare these total probabilities.

From equations 2.2.5 and 2.2.6 we have that

$$\begin{aligned} -Q^3+2P^2Q^2+Q^2-3P^2Q+P^2 &\geq Q^3+2P^2Q^2-4PQ^2-P^2Q+2PQ \\ -2Q^3+Q^2-2P^2Q+P^2 &\geq -4PQ^2+2PQ \\ -(2Q-1)(Q^2+P^2) &\geq -2PQ(2Q-1) \\ (Q^2+P^2) &\geq 2PQ \end{aligned} \tag{2.2.7}$$

It can be seen that since $Q \leq 1/2$, the following inequality is true,

$$P_{1010}+P_{0101} \geq P_{1001}+P_{0110}. \tag{2.2.8}$$

Thus, given that the true tree phylogeny is as in Figure 2.1, the estimate of the unrooted topology will either converge to configurations (AB)(CD) or (AC)(BD), but will never converge to (AD)(BC) as more characters are collected. Therefore, for Felsenstein to establish a condition for consistency of the estimation of the unrooted tree, he checks whether

$$P_{1100}+P_{0011} \geq P_{1010}+P_{0101}. \tag{2.2.9}$$

Once more, using equations 2.2.4 and 2.2.5, equation 2.2.9 is reduced to

$$\begin{aligned} 2P^2Q - P^2 + 2Q^3 - 3Q^2 + Q &\geq 0 \\ (2Q-1)(P^2 + Q(Q-1)) &\geq 0 \\ P^2 \leq Q(1-Q) \quad \text{since} \quad Q &\leq 1/2. \end{aligned} \tag{2.2.10}$$

This condition defines a Felsenstein zone where parsimony is not consistent in the case of four taxa. Felsenstein notes that this region where parsimony fails to converge to the correct tree phylogeny is due in part to the phenomenon of "long-branches attract" causing unrelated taxa with accelerated rates of evolution to be incorrectly grouped together as sister taxa.

As mentioned in the Introduction, the research done in five-taxa Felsenstein [32] and four-taxa Farris-type [23] trees do not give the analytical conditions for which parsimony is not consistent under a Felsenstein-type or Farris-type tree. Zharkikh and Li (1993) [32] do provide a proof for five taxa not being consistent, however under an equal rates model. Furthermore, Siddall (1998) [23] does not give any analytical conditions for the consistency of parsimony under a Farris-type tree. Thus, we look to compare Felsenstein's method in a Farris zone and look to extend conditions in the case of five taxa.

Chapter 3 will show the same application of total probabilities for a Farris-type tree in the case of four taxa to obtain an analytical condition for consistency.

Chapter 3

Consistency in a Farris-Type Tree in the Case of Four Taxa

3.1 The Farris Zone

In contrast to the Felsenstein zone, the Farris zone is where many [23], [26], [30], [29] have claimed success of parsimony over maximum likelihood in the four taxa case. Siddall [23] claimed that likelihood methods are not consistent in the Farris zone through "long-branch repulsion". The general approach to his method [23] was to construct DNA sequence data from a specific model of DNA substitution, such as the Jukes-Cantor model (1969) [18], through simulation and then evaluate the frequency for which parsimony, likelihood, and distance methods estimate the assumed true tree topology. In doing so, Siddall [23] concluded that in the Farris zone, as more and more characters are collected, maximum likelihood methods were unable to recover the correct tree across the entire parameter space. In contrast, parsimony found recovery rates over 95% as the amount of characters increased [23].

It is likely however that there were errors in Siddall's Method [27]. Nevertheless, Steel and Penny (2000) [26] have also indicated success of parsimony over ML in the Farris zone, however, they do not claim that ML per se is not a consistent method, given that this is a well known mathematical property. They note that the ML inference is based on a model different from that used to generate the data, then ML may appear to be not consistent.

None of these papers have tried to extend their study to the five taxa case, and all use simulations to calculate the frequencies of estimating the correct tree phylogeny to suggest success of parsimony in the Farris zone, rather than showing the underlying conditions for which parsimony is consistent.

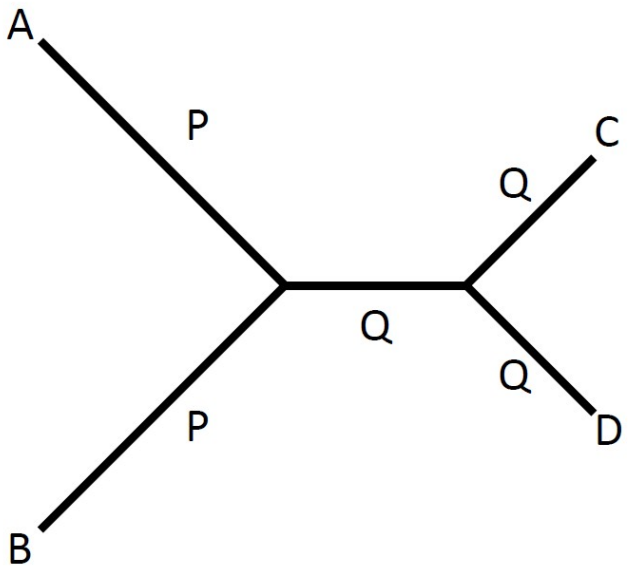

Figure 3.1: The true unknown phylogeny in the case of four taxa in a Farris-type tree. Sister taxa, A and B, have accelerated rates of evolution indicated with longer branches, while shorter branches indicate a slower rate of evolution. Next to each branch is the net probability of character change along the branch, P or Q.

To obtain a condition of consistency for parsimony to be in a Farris zone in the case of four taxa, application of total probabilities for the bifurcating phylogenetic trees will be used, as in Chapter 2 outlining Felsenstein's method [6]. Consider the same characterization of an unrooted phylogenetic tree as outlined in Sections 2.1.2, 2.2, and in Figure 3.1. The following are the total probabilities of the possible bifurcating phylogenies (AB)(CD), (AC)(BD), and (AD)(BC) respectively.

$$\begin{aligned} P_{1100} + P_{0011} &= P^2[(1-Q)^3 + Q^3] + (1-P)^2[Q(1-Q)^2 + Q^2(1-Q)] \\ &= 2P^2Q^2 + 2PQ^2 - Q^2 - 2P^2Q - 2PQ + Q + P^2, \end{aligned} \tag{3.1.1}$$

$$\begin{aligned} P_{1010} + P_{0101} &= 2[P(1-P)[Q^2(1-Q) + Q(1-Q)^2]] \\ &= 2P(1-P)(Q-Q^2), \quad \text{and} \end{aligned} \tag{3.1.2}$$

$$\begin{aligned} P_{1001} + P_{0110} &= 2[P(1-P)[Q^2(1-Q) + Q(1-Q)^2]] \\ &= 2P(1-P)(Q-Q^2). \end{aligned} \tag{3.1.3}$$

Assume that the configuration (AB)(CD) as in Figure 3.1 is the true phylogeny, to obtain a condition for which the parsimony estimate will have the property of consistency, we need to compare total probabilities to see which is largest, indicating the phylogeny that parsimony will converge to as more and more data is collected. Since equations 3.1.2 and 3.1.3 are equal, we only need to inquire if $P_{1100} + P_{0011} \geq P_{1010} + P_{0101}$. Using equations 3.1.1 and 3.1.2, we can rearrange and simplify to get the following inequality

$$4Q^2P - Q^2 - 4QP + Q + P^2 \geq 0. \tag{3.1.4}$$

As the inequality is a polynomial in terms of P and Q, we can find the critical points and determine whether there is a local minimum or maximum. The gradient for the polynomial is as follows,

$$\nabla f(P,Q) = \begin{bmatrix} 4Q^2 - 4Q + 2P \\ 8PQ - 2Q - 4P + 1 \end{bmatrix} \tag{3.1.5}$$

From the gradient, one can see that the critical point is $(\frac{1}{2}, \frac{1}{2})$. With this we get the following Hessian,

$$\mathrm{H}\, f(P,Q) \equiv \frac{\partial^2 f}{\partial P \partial Q} = \begin{bmatrix} 2 & 8Q-4 \\ 8Q-4 & 8P-2 \end{bmatrix} \tag{3.1.6}$$

$$\mathrm{H}\, f(\frac{1}{2}, \frac{1}{2}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \tag{3.1.7}$$

Since the eigenvalues of the Hessian are $2, 2 > 0$, the Hessian is positive-definite and we have a local minimum. This shows that the inequality 3.1.4 is always true for all positive values of P and Q. Additionally, it means that parsimony will converge to the true tree topology (AB)(CD) as more and more characters are collected. Shown here is the condition for parsimony to be successful at recovering the true phylogeny

and is consistent in the case of four taxa. This agrees with the claims by Waddell (1995) [29], Yang (1996) [30], Siddall (1998) [23], and Steel and Penny (2000) [26].

Of course, for the types of trees with the parameterization that Felsenstein used, non-sister taxa and all values of P and Q outside of the Felsenstein zone will also result in consistency in parsimony. So, we now know that parsimony is consistent in the compliment of the Felsenstein zone, and in the Farris zone we know mathematically that parsimony is consistent in those two zones. However, the performance of ML in the Farris zone is to be a matter of simulating models which are irrelevant to the Felsenstein argument. No one of course can prove that ML is not consistent and the work on the Farris zone seems largely a smokescreen thrown up by the cladists to distract from the 'long-branches attract' defect of parsimony.

Chapter 4

Investigating Consistency in Felsenstein and Farris-Type Trees in the Case of Five Taxa

Adding another taxon to this problem means that there are now fifteen possible unrooted bifurcating trees for the five species (see Figure 4.1). In the previous case of four taxa, each pair of taxa constitutes sisters in only one of the three possible tree topologies. However, in the case of five taxa, each tree has two pairs of taxa that are joined by an internal node (Figure 4.1). In other words, each pair of taxa constitutes taxa in three different trees, e.g., the pair (A,B) is found in trees 3, 4, and 5 (Figure 4.1). Consider taxon E to be the known outgroup, i.e., the other four taxa are more closely related to each other than to E, so now the trees can be categorized into two groups; (i) symmetrical trees where the number of taxa on both sides of the outgroup are equal (trees 5, 10, and 15), and (ii) asymmetrical trees (all other trees). In this chapter, we illustrate with symmetrical trees only.

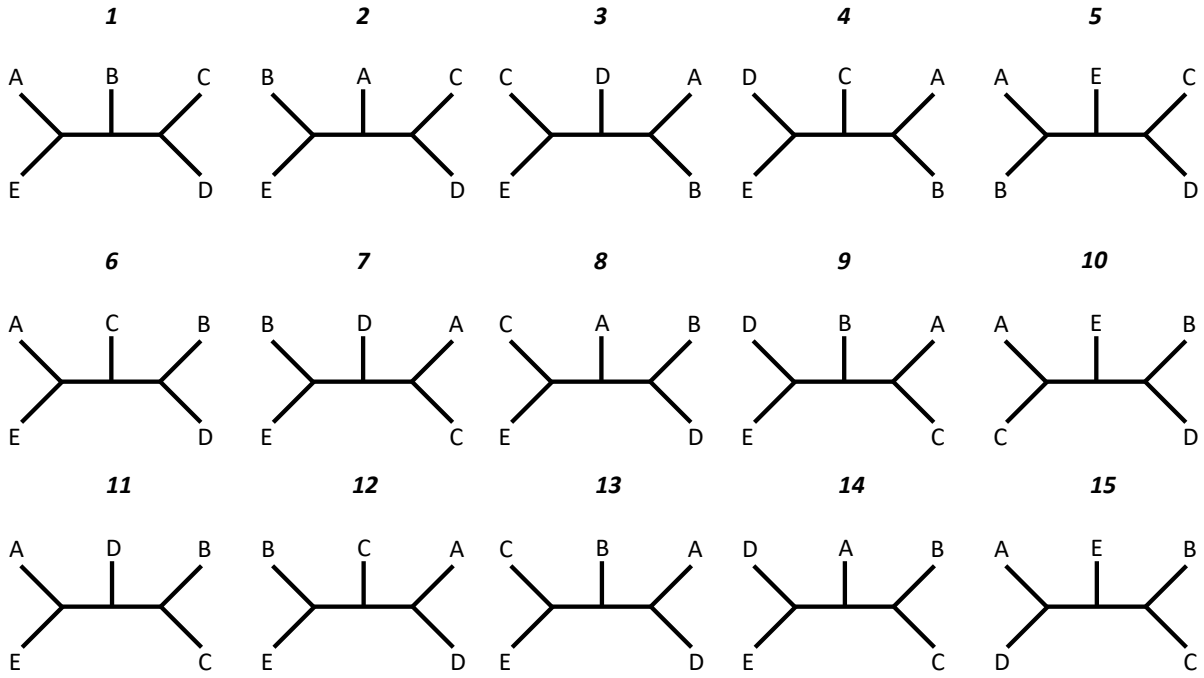

Figure 4.1: The fifteen configurations of unrooted bifurcating trees in the case of five taxa. Trees 5, 10, and 15 are categorized as symmetrical trees.

Hendy and Penny (1989) [14] recalculated Felsenstein's probabilities under the assumption of a molecular clock, i.e., the assumption that the rate of evolution is constant throughout time. With this assumption, they showed that for four taxa, parsimony will always converge to the correct tree [14]. Under a molecular clock, the internal branch lengths on any given tree are all the same length. In contrast, we have shown in Section 3.1 that when two branches in a Farris-type tree have accelerated rates of evolution compared to the other branches, parsimony still always converges to the correct tree. Hendy and Penny (1989) [14] admit they were not able to extend their method of proof to five taxa with equal rates. However, they provided one counter-example and showed that for any small value chosen for the substitution

rate, one can find a symmetrical tree with sufficiently short internal branches that can lead parsimony to converge to the incorrect tree [14]. We will show using Felsenstein's method some mathematical conditions for which parsimony in the case of five taxa fails to converge to the correct tree.

In like manner, Zharkikh and Li (1992, 1993) [31], [32] also studied the case of four and five taxa with a molecular clock using an analytical approach and Monte Carlo simulation . Although they used different models and asked different questions from Felsenstein, they found that when all probabilities like P and Q are low, parsimony is consistent, whereas when certain probabilities were large or some were larger than others, parsimony was not consistent. However, they did not ask the same questions as Felsenstein and others, and they did not establish any formal conditions where parsimony is consistent. The authors acknowledge that they were not able to prove that parsimony is not consistent for five taxa when evolutionary rate is fast [32]. In particular, their study does not focus on the original problem that Felsenstein had set, which was finding the constraints which make parsimony not consistent when two unrelated taxa have accelerated rates of evolution. The following sections will be an investigation of consistency in the case of five taxa with Felsenstein and Farris-type trees respectively.

4.1 Parsimony in the Case of Five Taxa is not Consistent in a Felsenstein-Type Tree

Following the method outlined in Sections 2.2 and 3.1, the total probabilities of symmetrical bifurcating trees will be compared to obtain a condition for the consistency of parsimony in the case of five taxa. Since we will be trying to find simple conditions for the 'long-branches attract' phenomenon to lead to inconsistency, such as those on rates in non-sister taxa for the four-taxa trees, we will focus only on two edges with topologically different roles in the symmetrical tree, namely the branches leading to A and E. The symmetrical bifurcating trees have binary character states, as in Felsenstein [6], and Hendy and Penny [14]. Following the unrooted Wagner trees model of parsimony, just as in the four taxa case, these binary character states are reversible.

Again, assume in Figure 4.2 that Q must be less than P so the phenomenon of 'long-branches attract' seen in Felsenstein-type trees can be studied. Similarly to Felsenstein's study [6], the probabilities in the segments that do not have accelerated evolutionary rate, are assumed to be the same.

Given that the assumed true phylogeny is as in Figure 4.2, the binary-state configurations of interest are (AB)(CD)(E), (AC)(BD)(E), and (AE)(BC)(D). Again, the parentheses with a paired taxa denote neighbours and the parentheses with a singular taxa denote the outgroup. Moreover, similar to the four-taxa case, the following probabilities are given for the ordered set of species P_{ABCDE}. Note that configuration (AD)(BC)(E) has the same probability as (AC)(BD)(E) on a symmetrical bifurcating tree with E as the outgroup. We are only interested in knowing the probabilities for the configurations that pair with taxa A since it has an accelerated branch rate. Thus,

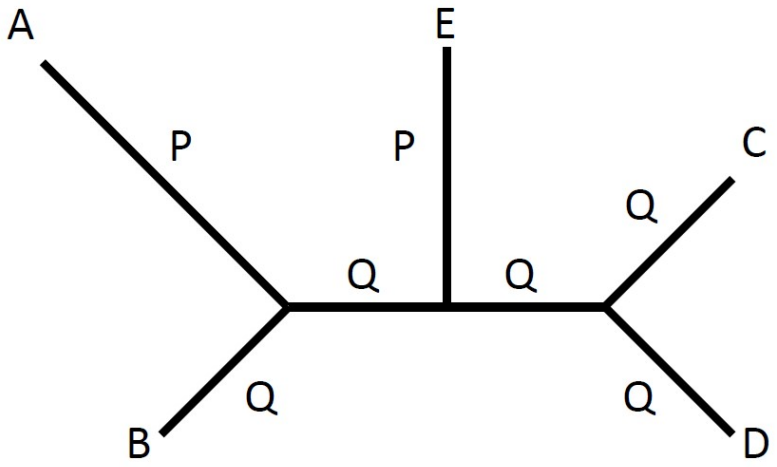

Figure 4.2: True unknown phylogeny in a five-taxa Felsenstein-type tree, where taxa E is the known outgroup. The unrelated taxa, A and E, separated by an internal branch, have accelerated rates of evolution while all other branches have equal rates. The net probability of character change along each branch, P and Q, are labelled.

the total probabilities will depend on data outcomes, $11000 + 00111$, $10100 + 01011$, and $10001 + 01110$. The following total probabilities will be compared, just as in Sections 2.2 and 3.1.

$$\begin{aligned} P_{11000}+P_{00111} &= PQ[PQ^2(1-Q)^2+PQ^3(1-Q)+Q^3(1-P)(1-Q)+(1-P)(1-Q)^4] \\ &+ (1-P)(1-Q)[Q^4(1-P)+Q(1-P)(1-Q)^3+PQ(1-Q)^3+PQ^2(1-Q)^2] \\ &= 2PQ^4-3Q^4-2P^2Q^3-3PQ^3+6Q^3+3P^2Q^2+PQ^2-4Q^2-P^2Q+Q, \end{aligned} \tag{4.1.1}$$

$$P_{10001} + P_{01110} = P(1-Q)[Q^2(1-P)(1-Q)^2 + Q^3(1-P)(1-Q) + PQ^3(1-Q) + P(1-Q)^4] + (1-P)Q[PQ^4 + PQ(1-Q)^3 + Q(1-P)(1-Q)^3 + Q^2(1-P)(1-Q)^2] = 2PQ^4 + Q^4 - 6P^2Q^3 - PQ^3 - 2Q^3 + 9P^2Q^2 + Q^2 - 5P^2Q + P^2, \quad (4.1.2)$$

and

$$P_{10100} + P_{01011} = P(1-Q)[PQ^3(1-Q) + PQ^2(1-Q)^2 + Q^2(1-P)(1-Q)^2 + Q(1-P)(1-Q)^3] + (1-P)Q[Q^3(1-P)(1-Q) + Q^2(1-P)(1-Q)^2 + PQ^2(1-Q)^2 + PQ(1-Q)^3] = 2PQ^4 - Q^4 - 2P^2Q^3 - PQ^3 + Q^3 + 3P^2Q^2 - 2PQ^2 - P^2Q + PQ. \quad (4.1.3)$$

Moreover, it can be shown from equations 4.1.2 and 4.1.3 that provided $Q \leq 1/2$,

$$\begin{aligned} & P_{10001} + P_{01110} \geq P_{10100} + P_{01011} \\ & \iff P^2(1-Q)^5 - P^2Q^2(1-Q)^3 - PQ(1-P)(1-Q)^4 + PQ^5(1-P) \\ & \qquad + Q^2(1-P)^2(1-Q)^3 - Q^4(1-P)^2(1-Q) \geq 0 \\ & \iff 2Q^4 - 4P^2Q^3 - 3Q^3 + 6P^2Q^2 + 2PQ^2 + Q^2 - 4P^2Q - PQ + P^2 \geq 0. \end{aligned} \quad (4.1.4)$$

As in Section 3.1, we can find the gradient of the polynomial in the inequality 4.1.4.

$$\nabla f(P,Q) = \begin{bmatrix} -8PQ^3 + 12PQ^2 + 2Q^2 - 8PQ - Q + 2P \\ -12P^2Q^2 + 8Q^3 - 9Q^2 + 12P^2Q - 4P^2 + 4PQ + 2Q - P \end{bmatrix} \tag{4.1.5}$$

Through the use of Mathematica [17], the real valued solutions (there are imaginary roots of P) found for this gradient are $P = 0, 0.5, 0.188909$ and $Q = 0, 0.5, 0.239992$. We can use the lowest critical point $(P,Q) = (0,0)$ and check to see if equation 4.1.4 has a local minimum or maximum.

With this we get the following Hessian,

$$\mathrm{H}\, f(P,Q) = \begin{bmatrix} -8Q^3 + 12Q^2 - 8Q + 2 & -24PQ^2 + 24PQ + 4Q - 8P - 1 \\ -24PQ^2 + 24PQ + 4Q - 8P - 1 & -24P^2Q + 24Q^2 + 12P^2 - 18Q + 4P + 2 \end{bmatrix} \tag{4.1.6}$$

$$\mathrm{H}\, f(0,0) = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \tag{4.1.7}$$

Since the eigenvalues are positive, the Hessian is positive-definite, meaning there is a local minimum at $(P,Q) = (0,0)$. This shows that the inequality 4.1.4 is always true for all positive P and Q. This implies that when the true tree is as shown in Figure 4.2, the estimate of the unrooted tree topology may converge to either (AB)(CD)(E), which would be correct, or (AE)(BC)(D), which would be wrong, but never to (AC)(BD)(E) or (AD)(BC)(E) as more and more data is collected. Although topology (AE)(BC)(D) would not be the correct phylogeny for parsimony to converge to (see Figure 4.2), with the phenomenon of 'long-branches attract', parsimony will

have a tendency to converge to this erroneous tree, i.e., we are entering a Felsenstein zone. Therefore, the results of this section, narrowed down by the following analytical results will show that there is a region where parsimony is not consistent and fails to converge to the true topology.

Accordingly, to preclude the consistency of the estimation of the unrooted tree topology, the following inequality needs to be satisfied.

$$\begin{aligned} & P_{11000} + P_{00111} \geq P_{10001} + P_{01110} \\ \iff & P_{11000} + P_{00111} - (P_{10001} + P_{01110}) \geq 0. \end{aligned} \tag{4.1.8}$$

Furthermore, with some more manipulation of the inequality 4.1.8, the following equation can be obtained.

$$4P^2Q^3 - 4Q^4 - 2PQ^3 - 6P^2Q^2 + 8Q^3 + PQ^2 + 4P^2Q - 5Q^2 - P^2 + Q \geq 0 \tag{4.1.9}$$

Following Felsenstein's method for obtaining a condition on the consistency of parsimony, equation 4.1.9 can be rewritten as a quadratic in P whose coefficients depend on Q [6].

$$0 \geq (1 - 4Q^3 + 6Q^2 - 4Q)P^2 - (-2Q^3 + Q^2)P + 4Q^4 - 8Q^3 + 5Q^2 - Q \tag{4.1.10}$$

Since the coefficient $(1 - 4Q^3 + 6Q^2 - 4Q) \geq 0 \quad \text{for} \quad Q \leq 1/2$, the quadratic 4.1.10 has a minimum at $P = (-2Q^3+Q^2)/(2(1-4Q^4+6Q^2-4Q))$. As the minimum is always positive for $Q \leq 1/2$, the positive values of P for 4.1.10 to be satisfied are the values of P above the point where the quadratic function is zero. The solution

to inequality 4.1.10, P_1, is the boundary between the values of P and Q for which we have consistency or non-consistency of the estimate of the tree topology by parsimony.

$$P_1 = \frac{2Q^3 - Q^2 + 2Q(Q(16Q^4 - 39Q^3 + 40Q^2 - 20Q + 4))^{(1/2)} - (Q(16Q^4 - 39Q^3 + 40Q^2 - 20Q + 4))^{(1/2)})}{2(4Q^3 - 6Q^2 + 4Q - 1)} \tag{4.1.11}$$

Figure 4.3 shows P_1 plotted in the parameter space of $P \leq 1/2$ and $Q \leq 1/2$. The curve P_1 is the boundary of the regions of where parsimony is consistent (C) and not consistent (NC). Above the curve is the region of values for which $P_{11000} + P_{00111} < P_{10001} + P_{01110}$.

The region above the curve is where the parsimony method is guaranteed to converge to the wrong estimate of the tree topology as more and more data is accumulated. It can be seen that for $0 \leq Q < 1/2$, there is a region where parsimony is not consistent, but the boundary curve ends when $Q = 1/2$. The implication of Equation 4.1.11 is that parsimony will tend to fail when there is large difference between P and Q. This is equivalent to saying that there is a sufficient difference between the elongated branches and the short segments of the tree (Figure 4.2). Another way of looking at what is happening here, as explained by Felsenstein (2003) [8, p. 114], is that with long branches leading to taxa A and E, the probability of parallel changes to arrive at the same state becomes greater than the probability of an informative single change on an interior branch of the tree. This implies that in the case of a Felsenstein-type tree, two changes along a long branch is more probable than a single change on a short interior branch. Taken all together, a condition for parsimony not being consistent in the case of five taxa has been obtained.

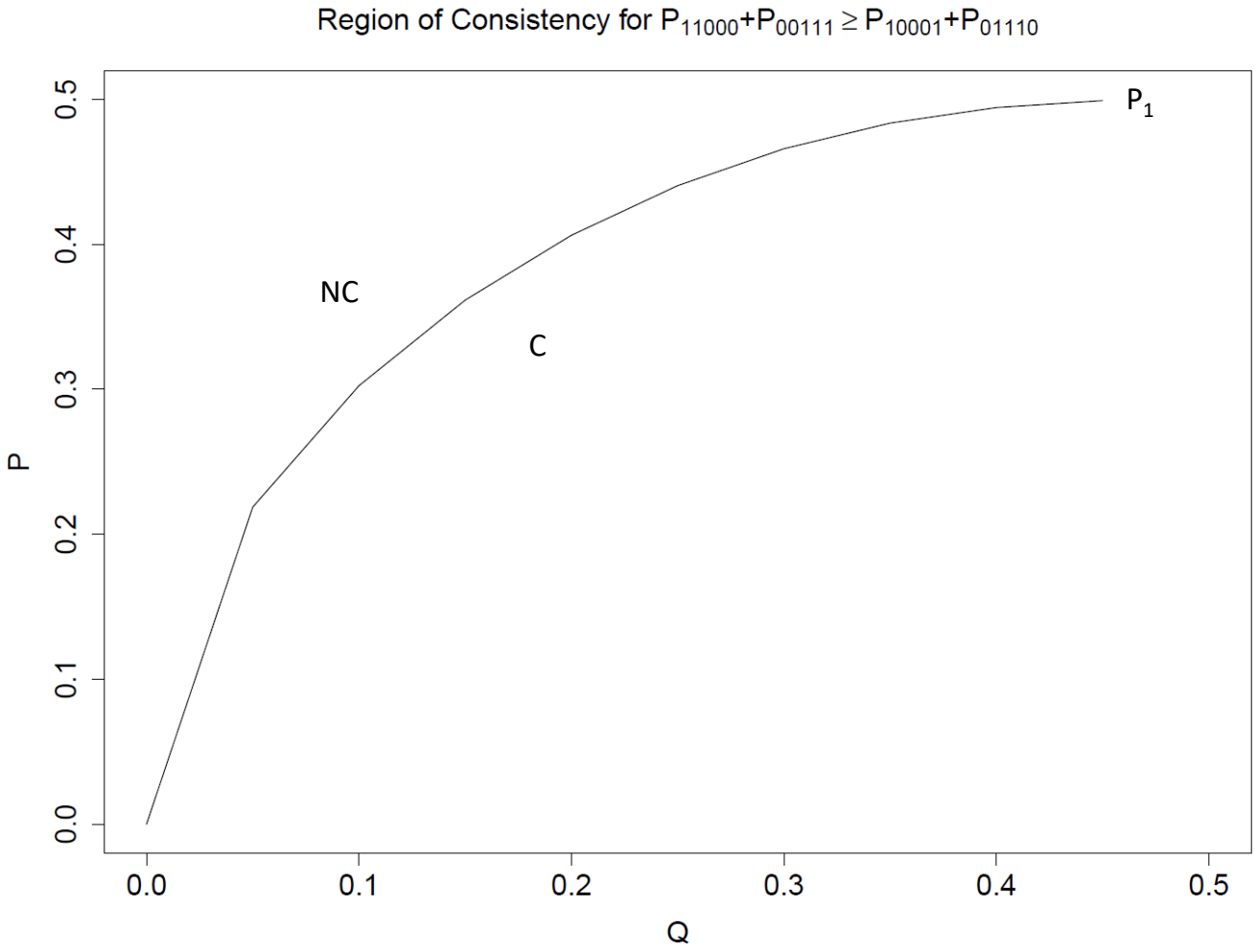

Figure 4.3: The parameter space of the solution for the inequality $P_{11000} + P_{00111} \geq P_{10001} + P_{01110}$. Plotted are the values of P and Q for which parsimony fails to be consistent. C denotes the region of consistency, while NC is the region where parsimony is not consistent. The boundary of these regions is the curve relating P_1 to Q

4.2 Consistency in a Farris-Type Tree in the Case of Five Taxa

In contrast to our study of consistency and lack of consistency in Felsenstein-type trees in the previous section, here, we focus on Farris-type trees, i.e., instead of studying the case where unrelated branches share a probability parameter, now it is sister branches that share a parameter. Authors such as Waddell (1995) [29], Yang (1996) [30], Siddall (1998) [23], and Steel and Penny (2000) [26], had found that parsimony in the case of four taxa in Farris-type trees is consistent and part of a Farris zone. Our results in the four-taxa case from Section 3.1 overlaps with the results (see Figure 4.4) obtained by Waddell (1995) [29]. He [29] demonstrated that given a Farris-type tree (Figure 4.4 (a)) as the true tree model, parsimony applied to the observed data is consistent and will not select an incorrect tree. In his example, the true tree phylogeny is given as T_{13}, where taxa t_1 and t_3 are sister taxa. On the other hand, an incorrect phylogeny is given as T_{12}, where taxa t_1 and t_2 are unrelated. This result overlaps with our four-taxa case in that the true tree, T_{13}, is the tree we gave as (AB)(CD) with the total probability $P_{1100} + P_{0011}$, and we demonstrated in Chapter 3 that parsimony will always converge to this phylogeny under the Farris zone. In contrast, the incorrect tree, T_{12}, is the tree (AC)(BD) with total probability $P_{1010} + P_{0101}$, and we showed that parsimony under the Farris zone will not converge to this incorrect phylogeny.

However, these authors do not provide a mathematical condition for which parsimony is consistent and in a Farris zone. It is important to note that developments based on a four-valued state space have been laid out by authors such as Yang (1996) [30] and Siddall (1998) [23]. Here, we are only looking at the the two-parameter space

Figure adapted from (Waddell, 1995)

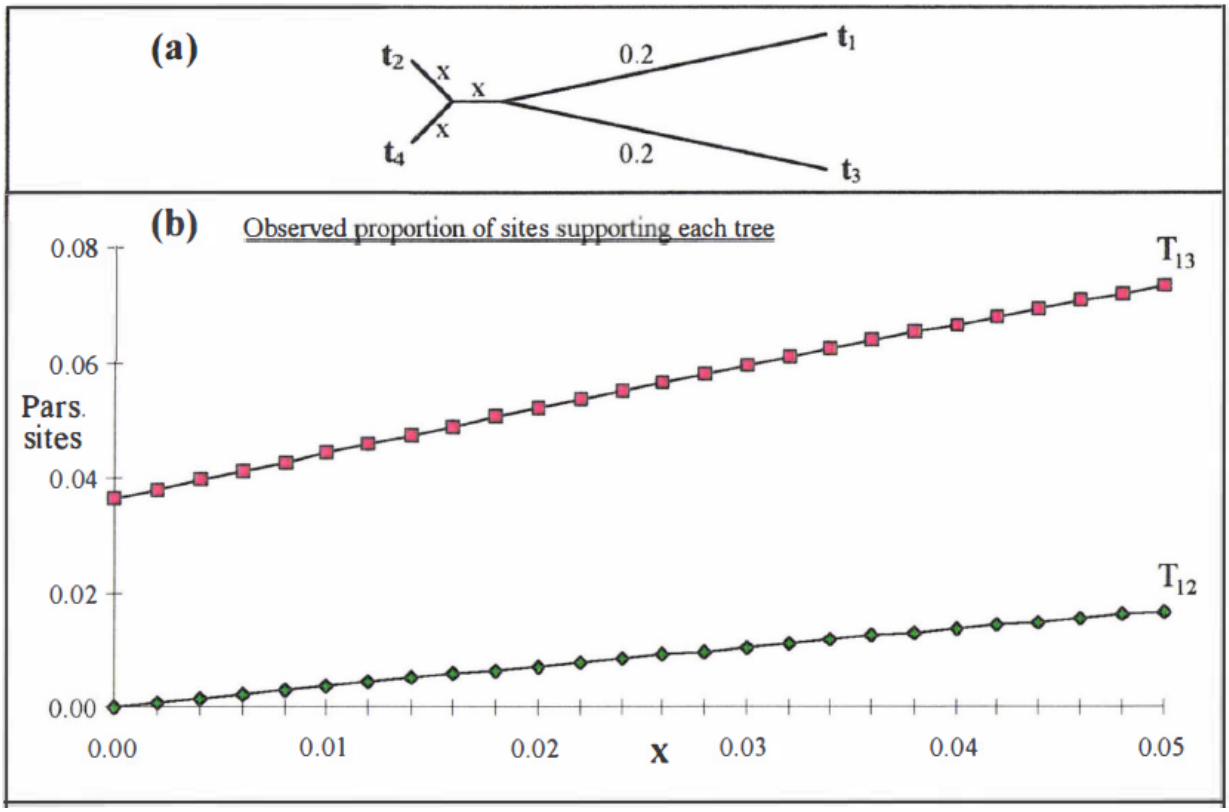

Figure 4.4: A brief simulation study of the consistency of parsimony in a Farris zone done by Waddell (1995) [29]. A Farris-type tree (a) as the model tree is given, where the phylogeny is labelled as $(t_1t_3)(t_2t_4)$. (b) Parsimony is applied to simulated observed data and reveals that in this case, parsimony will not converge to an incorrect tree topology (T_{12}), but will converge to the model tree (T_{13}) and is consistent. The model tree T_{13} relates to the phylogeny given in Chapter 3, (AB)(CD), and we showed that parsimony will always converge to this, while the incorrect tree, T_{12}, is like our tree (AC)(BD), which was demonstrated that parsimony will not converge to.

for the purpose of illustrating the extension of Felsenstein's (1978) [6] method to the Farris zone in the case of five taxa. We aim to provide a condition, in the case of five taxa, that shows parsimony is consistent when given a Farris-type tree. Figure

4.5 is the assumed true topology of a Farris-type tree in the case of five taxa, where taxa A and B have elongated branch lengths, meaning that there is an accelerated rate of evolution, while all other branches are assumed to have a slower rate. A future consideration would be to include a third rate, along with the net probability of character change along the branch, R, for the outgroup, E. The same formulation of a quadratic for the inequality can still be made, and the following quadratics would be in terms of P, Q and R instead.

$$P_{11000} + P_{00111} = (3Q^2 - 4RQ^2 - 3Q + 4RQ + R + 1)P^2 + (-6Q^3 + 4RQ^3 + 6Q^2 - 2RQ^2 - 2Q)P - 2RQ^3 + 3Q^3 + RQ^2 - 3Q^2 + Q \quad (4.2.1)$$

$$P_{10001} + P_{01110} = (-Q^4 + 3Q^3 - 2Q^2 - 6Q^3R + 2Q^2R + 2QR - R + 2Q^4)P^2 (Q^4 - 3Q^3 + 2Q^2 + 6Q^3R - 2Q^2R - 2QR + R - 2Q^4R)P, \quad (4.2.2)$$

and

$$P_{10100} + P_{01011} = (-Q^4 + RQ^3 + Q^3 + Q^2)P^2 + (Q^4 - RQ^3 - Q^3 + RQ^2 - Q^2 + Q)P. \quad (4.2.3)$$

Moreover, it would be even more important to construct a more realistic model which uses independent branch lengths, but this is not without its caveats of increasing the complexity of constructing such a model. This will be examined further in the Discussion. For now, we would like to illustrate the possibility of comparing Felsenstein's (1978) [6] method to the case of five-taxa under a Farris-type tree given only two branch rates and binary character states.

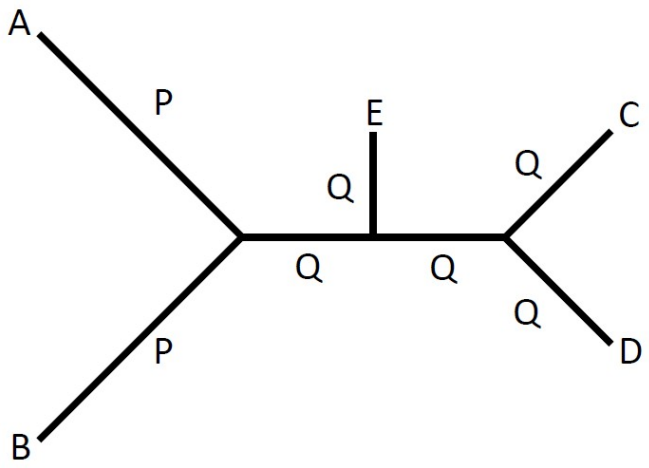

Figure 4.5: The true unknown phylogeny in a five-taxa Farris-type tree, where taxa E is the known outgroup. Sister taxa, A and B, which are joined by an internal node have accelerated rates of evolution, indicated by their longer branch lengths. On the other hand, shorter branches represent slower rates of evolution. Along each branch is the net probability of character change, P or Q.

We are looking to define a condition for consistency by comparing the total probabilities for the configurations of the true tree (Figure 4.5), $P_{11000}+P_{00111}$, $P_{10001}+P_{01110}$, and $P_{10100}+P_{01011}$.

$$\begin{aligned} P_{11000}+P_{00111} &= P^2[Q^3(1-Q)^2+Q^4(1-Q)+Q^3(1-Q)^2+(1-Q)^5] \\ &\quad +(1-P)^2[Q^4(1-Q)+Q(1-Q)^4+Q^2(1-Q)^3+Q^3(1-Q)^2] \\ &= 4PQ^4-2P^2Q^3-2Q^4-8PQ^3+4Q^4+7P^2Q^2+PQ^2-4P^2Q-3Q^2-2PQ+P^2+Q, \end{aligned} \tag{4.2.4}$$

$$\begin{aligned} P_{10001} + P_{01110} &= P(1-P)[Q^2(1-Q)^3 + Q^3(1-Q)^2 + Q^4(1-Q) + Q(1-Q)^4] \\ &\quad + P(1-P)[Q^5 + Q^2(1-Q)^3 + Q(1-Q)^4 + Q^2(1-Q)^3] \\ &= P(1-P)[3Q^2(1-Q)^3 + Q^3(1-Q)^2 + Q^4(1-Q) + 2Q(1-Q)^4 + Q^5], \quad (4.2.5) \end{aligned}$$

and

$$\begin{aligned} P_{10100} + P_{01011} &= P(1-P)[Q^4(1-Q) + Q^3(1-Q)^2 + Q^2(1-Q)^3 + Q(1-Q)^4] \\ &\quad + P(1-P)[2Q^3(1-Q)^2 + 2Q^2(1-Q)^3] \\ &= P(1-P)[3Q^3(1-Q)^2 + 3Q^2(1-Q)^3 + Q(1-Q)^4 + Q^4]. \quad (4.2.6) \end{aligned}$$

Again, we need to show that $P_{10001} + P_{01110} \geq P_{10100} + P_{01011}$, to see what tree topology the estimate will converge to as more data is collected. By using equations 4.2.5 and 4.2.6, the following inequality can be found after some algebra and rearrangement.

$$\begin{aligned} & P_{10001} + P_{01110} \geq P_{10100} + P_{01011} \\ \iff & Q^5 + Q(1-Q)^4 - 2Q^3(1-Q)^2 \geq 0 \end{aligned} \tag{4.2.7}$$

Equation 4.2.7 is always true for $0 \leq Q \leq 1/2$. Hence, when the true tree is given (Figure 4.5), the estimate of the unrooted tree topology may converge to either (AB)(CD)(E), which would be correct, or (AE)(BC)(D), which would be wrong, but never to (AC)(BD)(E) or (AD)(BC)(E) as more data is collected. Thus, to establish a condition for consistency, we need to check the inequality,

$$P_{11000} + P_{00111} \geq P_{10001} + P_{01110} \tag{4.2.8}$$

To obtain a condition of consistency, inequality 4.2.8 can be rearranged and written as a quadratic of P in terms of Q after the following simplification.

$$\begin{aligned} & 4PQ^4 - 2Q^4 - 4P^2Q^3 - 8PQ^3 + 4Q^3 + 7P^2Q^2 + 6PQ^2 - 3Q^2 - 4P^2Q - 2PQ \\ & \geq -4P^2Q^3 + 5P^2Q^2 - 2P^2Q + 4PQ^3 - 5PQ^2 + 2PQ \\ \iff & (2Q^2 - 2Q + 1)P^2 + (4Q^4 - 12Q^3 + 11Q^2 - 4Q)P - 2Q^4 + 4Q^3 - 3Q^2 + Q \geq 0 \end{aligned} \tag{4.2.9}$$

Since the coefficient $(2Q^2 - 2Q + 1) \geq 0 \quad \text{for} \quad Q \leq 1/2$, the quadratic 4.2.9 has a minimum at $P = -4Q^4 + 12Q^3 - 11Q^2 + 4Q/2(2Q^2 - 2Q + 1)$. As the minimum

is always positive when $Q \leq 1/2$, the positive values of P for 4.2.9 to be satisfied are the values of P above the point where the quadratic function is zero. However, since we expect parsimony to be consistent in this case, then it is to be expected that there will be no boundary between the regions where parsimony is consistent and not consistent represented by the following curve of P_1:

$$P_1 = \frac{-4Q^4 + 12Q^3 - 11Q^2 + 4Q \pm (16Q^8 - 96Q^7 + 248Q^6 - 344Q^5 + 281Q^4 - 136Q^3 + 36Q^2 - 4Q)^{(1/2)}}{2(2Q^2 - 2Q + 1)} \tag{4.2.10}$$

Figure 4.6 shows that there is no curve P_1 plotted for values of $0 \leq P \leq 1/2$ and $0 \leq Q \leq 1/2$ since the solution for the quadratic is imaginary, so there is no real solution for the curve. Thus, there is no boundary of the regions where parsimony is consistent (C) and not consistent (NC), implying that parsimony is consistent for all of these values of P and Q. Therefore, parsimony in the case of five taxa always recovers the true topology and is consistent (Figure 4.5). The plot of the five-taxa Farris zone parameter space shown in Figure 4.6 agrees with our results in the four-taxa case from Section 3.1 and overlaps with the results (see Figure 4.4) obtained by Waddell (1995) [29], who claimed that parsimony is consistent in the Farris zone. However, it can be seen that for $Q > 1/2$, there is the beginning of a boundary curve indicating a region where parsimony is not consistent.

Nevertheless, as stated in Section 2.1.2, P and Q cannot be greater than 0.5, since the chance that a state at the end of a branch is different than the state at the beginning is only 1/2, even if we had an arbitrarily long branch. Although we begin to see the boundary between where parsimony is consistent and not consistent for

$Q > 1/2$, this scenario would not be possible.

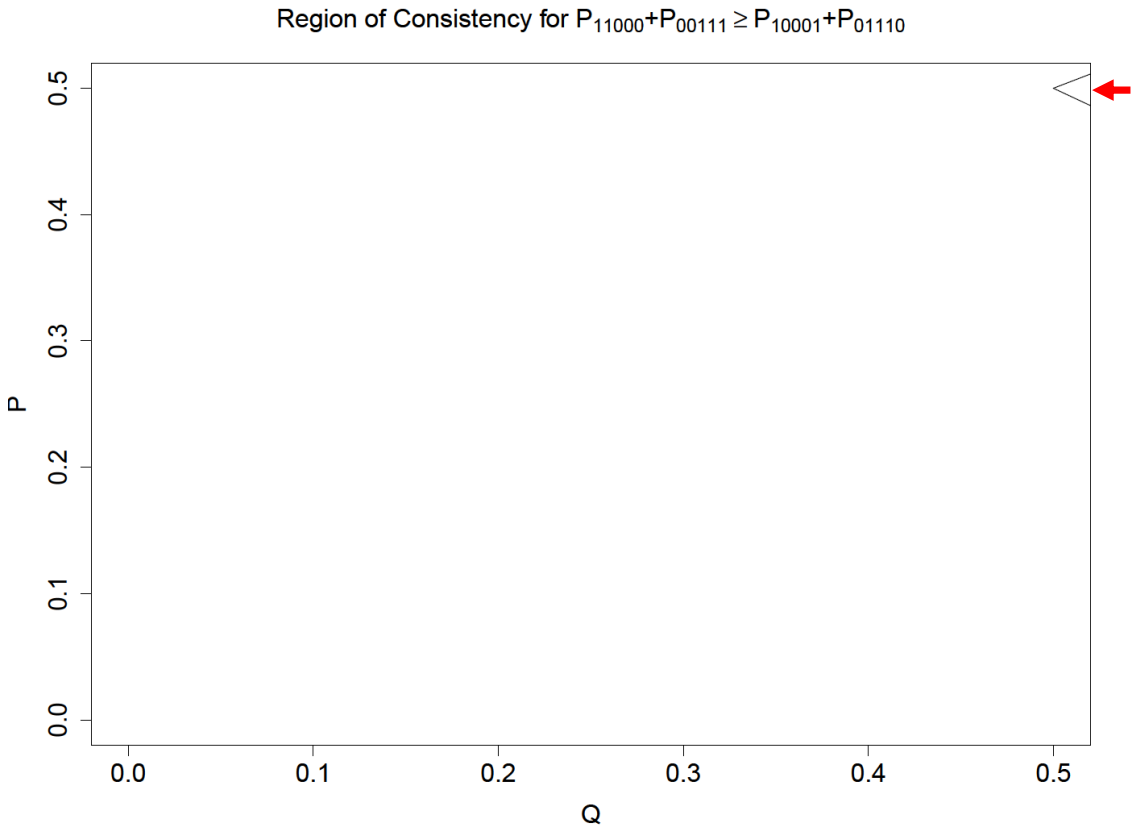

Figure 4.6: The plot of the solutions for the inequality $P_{11000} + P_{00111} \geq P_{10001} + P_{01110}$. These are the values of P and Q for which parsimony is consistent. Note that there is a boundary curve at $Q > 1/2$ indicating where parsimony is consistent and not consistent. When $Q \leq 1/2$, parsimony is consistent and recovers the true tree in the case of five taxa.

Chapter 5

Discussion

What has been shown in Section 4.2 is that parsimony is consistent in Farris-type trees in the case of five taxa. This supports the claims that many, such as Waddell [29], Yang [30], Siddall [23], and Steel and Penny [26], have made in the case of four taxa. This suggests that parsimony may be consistent in Farris-type trees with more than five taxa as well where the phenomenon of 'long-branches attract' will aid in the convergence to the correct tree topology by grouping together sister taxa with extended branches.

On the other hand, it was shown in Section 4.1 that parsimony is not consistent for five taxa in a Felsenstein-type tree. Furthermore, we demonstrated that 'long-branches attract' affects the convergence of parsimony and continues to be a factor in the estimation of tree phylogenies, not only in the case of four taxa, but also with five taxa. This could also suggest that parsimony may not be consistent for Felsenstein-type trees with more than five taxa through 'long-branches attract'.

As mentioned by Felsenstein, the models used here have their limitations as it will hardly be the case that characters will be sampled independently with all char-

acters following the same probability model in real life data. Of course, extending this analysis to more realistic evolutionary models would be the next step to improve and develop more of a concrete and general formulation for consistency, or not, of parsimony. Another aspect of this condition that could be extended, is replacing the binary character states 0 and 1, and instead using the DNA nucleotides, A, C, G, T, as the character states. This would add another layer of complexity as one would need to investigate the effects of using ordered versus unordered characters. Moreover, we would need to employ some evolutionary model for nucleotide substitution rates such as Kimura's K80 model [19], Felsenstein's F81 model [9], or Hasegawa, Kishino and Yano's HKY85 model [13]. Additionally, future endeavours should include independent branch rates rather than just two rates. Since evolution does not proceed at an even pace, it would be more reasonable to include varying branch lengths. The branch rate for the outgroup was assumed to be the same as the other internal branches to find the simple conditions for the 'long-branches attract' problem that leads to inconsistency. We only focused on two edges with topologically different roles in a symmetrical tree, those being branches A and E in a Felsenstein-type tree and branches A and B in a Farris-type tree.

The result of parsimony not being consistent for a Felsenstein-type tree with five taxa, extends the results from Felsenstein's (1978) [6] research. Parsimony can be consistent, but there needs to be specific conditions such as sister taxa having accelerated rates of evolution while restricting all other branch rates to have slower evolving branches. In modern phylogenetic inferrence, many studies often choose maximum likelihood methods over parsimony due to its statistical properties like consistency and asymptotic efficiency as more and more data is collected [8, pp. 269-274].

Another extension of this study would be to apply would be applying a similar treatment to maximum likelihood methods. It would be worth looking into what departures from the simple model we studied here could lead maximum likelihood choosing the wrong tree.

Chapter 6

Conclusion

In this thesis, consistency of parsimony in the four and five taxa case is explored for Felsenstein and Farris-type trees. By extending Felsenstein's methodology for unrooted bifurcating trees, we demonstrated that conditions for consistency can be found for both Felsenstein and Farris-type trees, in the case of four and five taxa. Since the studies [23], [26], [29], [30] suggesting the consistency in four-taxa Farris-type trees used only simulations to estimate the frequency of correct tree topologies for their results, they did not find any analytical conditions or constraints for consistency. We showed that for a Farris-type tree in the case of four taxa, a condition for consistency, i.e., parsimony is in a Farris zone, can be found for the branch lengths, where two sister taxa had accelerated rates of evolution on a given tree phylogeny. We showed in Section 3.1, that the polynomial is always positive for $P \in [0, 1/2]$ and $Q \in [0, 1/2]$, implying that parsimony will always converge to the correct tree, which is thus located in a Farris zone.

We investigated the consistency of a Felsenstein-type tree in the case of five taxa. By comparing the total probabilities of the symmetric tree topologies for five taxa

as an inequality, we established the conditions for which the unrooted tree topology may converge to the parsimony estimate. Additionally, we established a condition for consistency based on these inequalities. As in the case with four taxa, we plotted the boundary between the regions where parsimony is consistent and not consistent. This corroborates with Felsenstein's [6] original finding that 'long-branch attraction' heavily influences the convergence to the incorrect tree by grouping non-related taxa as a sister group. The phenomenon of 'long-branch attraction' may also be true for phylogenies with more species.

We extended the consistency of parsimony in a Farris-type tree in the case of five taxa. In contrast to a Felsenstein-type tree, where two unrelated taxa have accelerated evolutionary rates and are grouped together by 'long-branch attraction', a Farris-type tree can take advantage of 'long-branches attract' by having a pair of sister taxa with the accelerated rates, and guaranteeing their arrangement. Once more, we found an analytical condition in the form of inequalities for the symmetrical tree phylogenies and demonstrated that parsimony is consistent for the five-taxa case. The condition for consistency was represented by a plot of the parameter space.

Nevertheless, these are not the be-all and end-all set of conditions for the consistency of parsimony, as this study uses Felsenstein's [6] assumptions which are not entirely biologically reasonable, such as characters being independently sampled and characters all following the same evolutionary probability model. Additionally, we are only looking at a specific subset of tree topologies with symmetrical trees, so we only have conditions for a few examples under Felsenstein and Farris-type trees.

We have shown how to extend the formulation, that Felsenstein first introduced, to the Farris zone and to five taxa. The newly formulated results in this thesis pave a way for more realistic evolutionary models, and for multiple character states. We

obtained conditions for consistency in the case of five taxa for a Felsenstein and a Farris-type tree respectively, which has not been shown analytically before, and further work could make it possible to extend to six or more taxa. However, the complexity also increases as the number of possible unrooted phylogenies for n taxa is $(2n-5)!/2^{n-3}(n-3)!$. The number of different branchings that could occur in just six taxa significantly increases the dimensions on a given phylogeny and makes it difficult to determine what regions will be in a Felsenstein or Farris zone. This thesis and the many papers that precede it, serves as a basis for any future exploration of conditions for consistency in parsimony. Although parsimony not being consistent may stem from 'long-branches attract', the results from these studies will give a more solid understanding of why parsimony may be positively misleading in certain cases.

Bibliography

[1] Joseph H. Camin and Robert R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, 19(3):311–326, 1965.

[2] L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis: Models and estimation procedures. *Evolution*, 21(3):550–570, 1967.

[3] William H.E. Day, David S. Johnson, and David Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81(1):33 – 42, 1986.

[4] A.W.F. Edwards and L.L. Cavalli-Sforza. *Reconstruction of evolutionary trees*, page 67–76. Cambridge University Press, 1964.

[5] James S. Farris. Methods for computing Wagner trees. *Systematic Zoology*, 19(1):83–92, 1970.

[6] Joseph Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1978.

[7] Joseph Felsenstein. Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution*, 35(6):1229–1242, 1981.

[8] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer, 2003.

[9] Joseph Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 2005.

[10] Walter M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[11] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[12] J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29(1):53–65, 1973.

[13] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 10 1985.

[14] Michael Hendy and David Penny. A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38, 12 1989.

[15] John P. Huelsenbeck. Is the Felsenstein Zone a Fly Trap? *Systematic Biology*, 46(1):69–74, 03 1997.

[16] John P. Huelsenbeck and David M. Hillis. Success of Phylogenetic Methods in the Four-Taxon Case. *Systematic Biology*, 42(3):247–264, 09 1993.

[17] Wolfram Research, Inc. Mathematica, Version 12.2.

[18] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. *Munro, H.N., Ed., Mammalian Protein Metabolism, Academic Press, New York*, pages 21–132, 1969.

[19] Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mole Evol*, 16:111–120, 1980.

[20] Naruya Saitou and Masatoshi Nei. The Neighbor-Joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 07 1987.

[21] David Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

[22] Susanne Schulmeister. Inconsistency of maximum parsimony revisited. *Systematic Biology*, 53:521–8, 09 2004.

[23] Mark E Siddall. Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris zone. *Cladistics*, 14(3):209 – 220, 1998.

[24] Mark E Siddall and Michael F Whiting. Long-branch abstractions. *Cladistics*, 15(1):9–24, 1999.

[25] Elliott Sober. The contest between parsimony and likelihood. *Systematic Biology*, 53(4):644–653, 2004.

[26] Mike Steel and David Penny. Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. *Molecular Biology and Evolution*, 17(6):839–850, 06 2000.

[27] David Swofford, Peter Waddell, John Huelsenbeck, Peter Foster, Paul Lewis, and James Rogers. Bias in phylogenetic estimation and its relevance to the choice

between parsimony and likelihood methods. *Systematic Biology*, 50:525–39, 09 2001.

[28] Chris Tuffley and Mike Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Journal of The Neurological Sciences - J NEUROL SCI*, 59, 05 1997.

[29] Peter J. Waddell. *Statistical methods of phylogenetic analysis : including Hadamard conjugations, LogDet transforms and maximum likelihood : a thesis presented in partial fulfilment of the requirements for the degree of Ph.D. in Biology at Massey University.* PhD thesis, 1995.

[30] Ziheng Yang. Phylogenetic analysis using parsimony and likelihood methods. *J Mol Evol*, 42:294–307, 2 1996.

[31] Andrey Zharkikh and Wen-Hsiung Li. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. four taxa with a molecular clock. *Molecular Biology and Evolution*, 9:1119–47, 12 1992.

[32] Andrey Zharkikh and Wen-Hsiung Li. Inconsistency of the maximum-parsimony method: The case of five taxa with a molecular clock. *Systematic Biology*, 42:113–125, 06 1993.