GHOST: A time-reversible mixture model for recovering phylogenetic signal from heterotachously-evolved sequence alignments.

Stephen Crotty

Thesis submitted for the degree of

Doctor of Philosophy

in

Applied Mathematics

at

The University of Adelaide

(Faculty of Mathematical and Computer Sciences)

Department of Applied Mathematics

January 13, 2017

Contents

List of Tables

List of Figures

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: ..... DATE: .....

Acknowledgements

Much of the credit for this Degree must go to my wife Jennifer, who has contributed as much indirectly as I have directly. There are several sound arguments against commencing a PhD at the same time as bringing twins into the world. It places a ceiling on earning ability at a time when expenses rise dramatically. It requires attention when free time has become a faded memory. It drains mental and emotional resources that are already in short supply. Over the last four years we have had to overcome all of these challenges, and more. At no stage was it easy, but we have survived. Thank you for your support at every step of the way. I hope that all of our efforts bear ample fruit for our children and for ourselves.

My Mum and Dad, to whom I owe everything I have. Your love, support and generosity has been a constant throughout my life, regardless of any decisions I have made. Your perpetual hard work and sacrifice is the only reason that I was in a position to consider commencing a PhD at the age of 33 with a young family on the way. You have taught me to stand on my own two feet, all the while making sure that I never had to.

To my robust supervisory panel: Professor Nigel Bean, Doctor Jonathan Tuke, Associate Professor Barbara Holland and Doctor Lars Jermiin. As in the movies, the team of superheroes all have their own particular skill set. Nigel, I could not have asked for a better principal supervisor. You were always patient and understanding of the external pressures that often took precedence over my PhD. Without your compassion in this area I doubt that I would have been able to successfully complete. Your generosity in funding my trip to Vienna last year provided the opportunity

to form a collaboration that strengthened my PhD and resulted in my ongoing employment. Your mathematical assistance and mentorship also exceeded anything I could have hoped for. Our weekly meetings were essential to me, helping me to organise my thoughts and give focus to the project. Many times I called past your office unannounced, seeking enlightenment on some trivial concept or result. I know you must often have been very busy on far more important matters, yet you never made me feel like I was interrupting. Jono, in contrast, you always made me feel like I was interrupting even when I had an appointment. When I returned to tertiary study seven years ago, entering a formula in the cell of an Excel spreadsheet was about the extent of my coding expertise. Any progress I have made in this area has been greatly facilitated by your advice, guidance and example. You often made great improvements to my code in a matter of seconds ('select all + delete' usually sufficed), you taught my son a valuable lesson about trusting Mancunians and you never let maths spoil a fun meeting. Barbara, there have been a handful of times over the last four years that my understanding of the subject matter has undertaken a quantum leap forward. These occasions coincided with the all too rare times where I was able to sit down with you at a conference and discuss the material in depth. I left every conference with renewed clarity of thought and enthusiasm for the project. There is no doubt in my mind that had I been lucky enough to have you as a local supervisor, I would have found the project easier and the end result stronger. Lars, your enthusiasm was contagious and always provided me with motivational boosts when I needed them most. Often I would enter a supervisor meeting to report what I believed to be mundane results, only to have you convince me that they were exciting, interesting and novel. This encouragement was vital, particularly in the early stages of the Degree when it was very easy to feel that the task was too great. To all of my supervisors, thank you for your efforts over the past four years. You have all made significant contributions that have shaped not only the project but myself as a researcher.

To Professor Arndt von Haeseler and Doctor Bui Quang Minh, the collaboration

that you facilitated marked a major turning point in my project. Without your help the results would have been primarily theoretical, and the thesis significantly weaker. The implementation of the GHOST model in IQ-TREE has enabled its application to a wide variety of biological problems, hopefully ensuring its relevance long after my PhD is complete. I also must thank you for the faith you have shown in me by your offer of employment when I was only half way through my PhD. As a student with a young family, I had anticipated the uncertain transition from PhD to Post Doc employment as a very stressful time. The security provided by your employment offer has minimised this stress for myself and my family.

To Ben, history will show that the first couple of phylogeneticists to come out of the School of Mathematics at the University of Adelaide were of outstanding quality, on average. It is no coincidence that over the years I have requested your assistance far more regularly than you have requested mine. The times we have shared at conferences have been a highlight, some more memorable than others. I look forward to catching up with you at these events long into the future.

Dedication

For Emily and Daniel, may you find happiness wherever life takes you.

Abstract

The accuracy and reliability of phylogenetic inference is compromised by the adoption of models of sequence evolution that don't adequately reflect the dynamic nature of evolution by natural selection. Heterotachy refers to variation in the rate of evolution of a particular site across lineages on a tree. We carry out simulations, showing that phylogenetic inference using popular methods and models is unreliable when the data evolved under the influence of heterotachy. We carry out a theoretical analysis of these methods and models, concluding that their failure was inevitable given the nature of the data.

To remedy this we introduce the General Heterogeneous evolution On a Single Topology (GHOST) model. We implement the GHOST model under a maximum-likelihood (ML) framework in the phylogenetic inference program IQ-TREE. We perform extensive simulation studies, showing that the GHOST model can successfully recover the tree topology, branch lengths and substitution model parameters from heterotachously-evolved sequences. We apply our model to a real dataset and identify a subtle phylogenetic signal linked to the convergent evolution of the electric organ in two geographically distinct lineages of electric fish. Furthermore, we use the model to successfully identify specific sites in the alignment that are pivotal to the effective function of the electric organ.

The GHOST model and its implementation in IQ-TREE provide the most flexible mixture model currently available for performing phylogenetic inference in a ML framework. This increased flexibility better equips the GHOST model to represent the process of evolution by natural selection. We show that the GHOST model is

able to highlight subtleties in evolutionary relationships that coarser models cannot. We foresee the GHOST model having potential uses in a variety of applications: helping to resolve disputed topologies; focusing the efforts of biologists by identifying alignment sites of functional importance; bringing to light evidence of convergent evolution; and investigating the coevolution that occurs between disease and immune cells, or hosts and parasites. As computing resources continue to grow and phylogenetic algorithms are revised and improved, the GHOST model will be applicable to ever larger MSAs, ultimately assisting in illuminating the history of life on earth.

Chapter 1

Introduction

Human beings have existed on the planet in our current form for at least the 100,000 years. For most of that time our species existed in comparative ignorance of the fundamental workings of the natural world. We are fortunate to live in a time that could be described as the dawn of enlightenment. The scientific revolution is generally considered to have commenced in Europe in the 16th century. The establishment of the scientific method paved the way for ground breaking advancements. One such advancement was the development of the theory of evolution by natural selection. The theory was developed independently in the 19th century by British naturalists Alfred Wallace and Charles Darwin, the latter publishing it in his book On the Origin of Species (Darwin, 1859). The theory of evolution was truly revolutionary, it contradicted the commonly accepted view that species were created in their current form by a Creator. The advancement in the biological sciences of the last 150 years has by far exceeded that of the preceding 100 millenia. The importance of the theory of evolution cannot be overstated. It is the foundation on which modern medicine is built and it underpins the entire field of biology. It is also fundamental in modern agriculture: the development of genetically modified crops can increase resistance against drought and pests, thereby increasing crop yields.

As naturalists, Wallace and Darwin both traveled the world, observing and categorising different morphological traits in species, depending on their location and

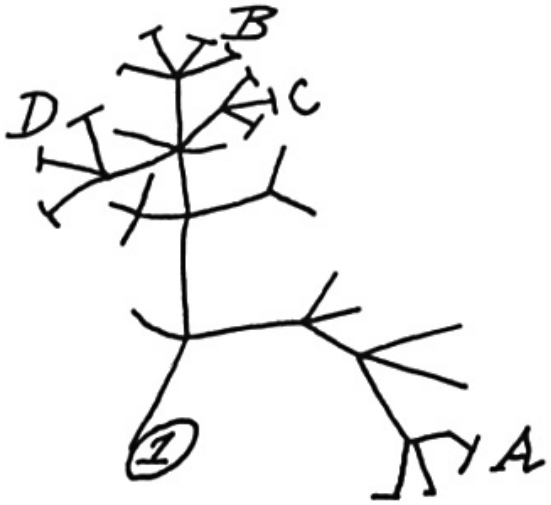

Figure 1.1: Darwin's original sketch from On the Origin of Species (Darwin, 1859), showing him formulating the concept of a phylogenetic tree.

environment. Darwin famously categorised the subtle differences in the shape of the beak of the finches he found on each of the Galapagos Archipelago. He hypothesised that an ancestral species had established isolated populations on different islands and natural selection then acted upon these populations, independently evolving the shape of their beaks to best suit their diets.

The evolutionary history of a set of species was first characterised by Darwin on a tree diagram (Figure 1.1), known as a phylogenetic tree. We can think of phylogenetic trees in the same way as we might be familiar with our own family tree. Closely related species appear close to each other on the phylogenetic tree just as our immediate family appears close to us on our family tree. In the time following the discoveries of Darwin and Wallace, morphological data was relied upon to construct phylogenetic trees. Species were positioned on a tree depending on shared common features, *e.g.* presence or absence of vertebra, hair, scales or feathers. Species sharing many morphological features are placed (such as Darwin's finches) close to each other on the tree, while those with little in common are placed further apart.

The construction of phylogenetic trees that provide an accurate picture of evo-

lutionary history is the core goal of the field of phylogenetics. Advancements in molecular biology, stochastic modeling techniques and computer science provide fertile ground for progress.

We begin in Chapter 2 by providing a foundation for understanding the thesis. We familiarise the reader with the relevant fundamental definitions, concepts and ideas of molecular biology. We introduce deoxyribose nucleic acid (DNA) and amino acids, explaining how the process of natural selection operates at the molecular level, ultimately driving biodiversity. We then explain phylogenetic inference and some of the more popular mathematical methods employed to carry it out: maximum parsimony (MP), neighbour joining (NJ) and maximum likelihood (ML). These methods are the process by which we can take DNA from a number of subject species and produce trees which estimate the evolutionary history of those subjects. Most sophisticated methods of phylogenetic inference require an evolutionary model to be adopted. These models define a set of assumptions about the DNA of the subject species and their ancestors. The assumptions generally relate to the composition of the DNA and how each base evolves over time. We provide a brief history of models of evolution, from their simple beginnings through to the more complex models that are still being proposed and developed today.

Chapter 3 focuses on the phenomena of heterotachy and provides a survey of the related literature. The word heterotachy is derived from the Greek *heteros* (different) and *tachos* (speed), meaning in essence variable speed. In the phylogenetic sense it is used to describe the scenario where a particular part of the DNA evolves at different speeds in different species. We then explore heterotachy in detail. We simulate data under varying degrees of heterotachous influence. We apply some popular inference methods and models of evolution to the simulated data, and empirically show that they are not well equipped for heterotachously-evolved data. We also explain our results theoretically using stochastic modeling techniques. We conclude from Chapter 3 that the root cause of the difficulties is the misspecification of the evolutionary model as none of the models allowed for heterotachous evolution. Thus

when presented with heterotachously-evolved data, the models cannot be relied upon to recover the true evolutionary history.

In Chapter 4 we propose a solution to this problem, in the form of a model of evolution that allows DNA to evolve heterotachously. We implement a simplified version of this model, restricted to only four species, in a ML framework. This is sufficient to show that using the model we are able to successfully recover tree topology, branch lengths and parameters of the substitution model for the heterotachously-evolved data simulated in Chapter 3. Encouraged by these results we then sought to implement the model in a more general sense, so that it may be used with meaningful biological data. A collaboration with the Centre of Integrative Bioinformatics Vienna (CIBIV) enabled the model to be implemented in their phylogenetic inference software package: IQ-TREE. The remainder of Chapter 4 provides a background to IQ-TREE and an explanation of the algorithms and structure it employs to carry out phylogenetic inference. We then discuss the necessary extension of these algorithms to incorporate our proposed model, as well as the addition of a new algorithm specifically for our model.

Chapter 5 focuses on the validation of the implementation of the model in IQ-TREE. We carry out three rigorous simulation studies, each focusing on a different aspect of the phylogenetic inference problem. The first simulation study is a reproduction of an experiment from a seminal paper on heterotachy by Kolaczkowski and Thornton (2004). The original experiment demonstrated that MP and ML could not be relied upon to infer the correct tree topology when the data evolved heterotachously. Furthermore, they found that MP outperformed ML for such data, a result that shocked many phylogeneticists at the time. Our simulations demonstrate that the ML implementation of our model in IQ-TREE is able to successfully infer the tree topology for this type of data. Recovery of tree topology is often of primary interest in phylogenetics, but two trees may share a common topology while still having vastly different implications for the species displayed on them, depending on their respective branch lengths and model parameters. Correctly inferring these

quantities is equally important. Our second simulation study focuses on this aspect of the inference. We simulate several datasets on a single random tree topology, with random branch length and model parameters. We find that IQ-TREE is able to use our model to successfully recover all of these parameters to a high level of accuracy. We also demonstrate that the results of the analysis could be used to identify areas of a subject's DNA that may have evolved under different evolutionary pressures. This is an unexpected benefit of the model. The third simulation study focuses on establishing that the results of the second test can be replicated. We generate many random tree topologies and a random set of branch lengths and model parameters on each of them. From each of these we simulate one dataset and use IQ-TREE to infer the tree topology, branch lengths and model parameters. Some simple metrics are constructed to quantify the difference between the true and inferred quantities, allowing the results from different datasets to be directly comparable. Again the results are encouraging, showing that IQ-TREE consistently infers the topology, branch lengths and model parameters to a high level of accuracy from heterotachously-evolved data.

The successful validation of our model's implementation in IQ-TREE effectively licenses it for use on real biological datasets. Chapter 6 demonstrates the performance of our model when applied to a dataset containing the DNA of 11 species of fish Zakon *et al.* (2006). The 11 species consist of five electric fishes and six non-electric fishes. The electric fishes originate from either South America or Africa and an earlier paper had established that these two distinct lineages evolved their electric organ convergently. That is to say, the most recent common ancestor of the South American and African electric fishes was not an electric fish. The results of the analysis with IQ-TREE using our model concur with this finding. We are able to highlight a subtle phylogenetic signal that clearly indicates evolutionary pressure acting on the electric fishes to the exclusion of the non-electric fishes. Further, we are able to highlight specific parts of the DNA that are responsible for generating this signal. We find that these parts correlate strongly with findings of biologists that

have isolated specific amino acids that are critical to the development and function of the electric organ.

We conclude the thesis by summarizing the work and suggesting future refinements of the model as well as other potential applications. Heterotachy is a critical issue within the field of phylogenetics. Failure to adequately model heterotachous evolution casts doubt over any phylogenetic analysis where the data is suspected of evolving under the influence of heterotachy. Given the complex nature of the evolutionary process it is reasonable to assume this applies to the vast majority of biological data. Our model, and its implementation in IQ-TREE, represent a significant advancement in the field that may lead to new insights about the evolutionary history of life on earth.

Chapter 2

Background

Phylogenetics is the study of the underlying evolutionary relationships that exist between all life on earth. In the 19th century when Darwin and Wallace were independently developing their theories on evolution they needed to rely primarily on morphological data: cataloguing species and observing their similarities and differences. Morphological data is still relied upon to some extent but its influence has diminished since Watson and Crick (1953) proposed the structure of deoxyribose nucleic acid (DNA), the fundamental molecule for all known life. Their discovery provides the foundation for the entire field of molecular biology. In the 1970's efforts were focused on the development of techniques to sequence the DNA of organisms (Maxam and Gilbert, 1977; Sanger *et al.*, 1977). Mullis and Faloona (1987) developed a fast method for sequencing the DNA of organisms via the polymerase chain reaction. Ongoing development in high throughput DNA sequencing techniques has provided phylogenetic researchers with a wealth of sequence data for analysis (Goodwin *et al.*, 2016).

2.1 DNA

DNA is found in every cell of all living organisms. It contains the molecular instructions for the construction of the cells and controls their function within the

organism. It is composed of two strands of nucleotides, coiled around each other in a double helix structure. There are four different types of nucleotides that make up these strands, also known as bases: adenine (A), guanine (G), cytosine (C) and thymine (T). The arrangement of nucleotides on the two strands of the double helix is not independent. The nucleotide on one strand determines the nucleotide in the corresponding position of the other strand. An A on one strand always bonds with a T on the other, and vice versa. Likewise, a C on one strand bonds with a G on the other, and vice versa. Consequently, it is only necessary to consider one strand of DNA as this fully defines the base pairs.

2.1.1 Sequence structure

Ultimately, the function of DNA is to build amino acids in a specific order, such that they form the proteins necessary for the growth and survival of the organism. With this in mind we restrict our focus to coding regions of the DNA sequence. Coding regions of DNA directly determine the amino acids that are constructed. The nucleotides in a coding sequence of DNA are grouped into sets of three bases each, known as codons. Each codon encodes for a specific amino acid. There are four distinct bases so therefore there are $4^3 = 64$ unique codons. Since there are only 20 unique amino acids, there is significant redundancy present in this system. Most amino acids can be constructed by multiple different codons. Additionally, the codon ATG signifies the start of a coding sequence of DNA whereas the codons TAA, TGA and TAG all signify the end of a coding sequence. Table 2.1 displays the 64 codons and the amino acids for which they encode.

2.1.2 Mutation

The cells of living organisms, along with the DNA molecules contained within those cells, are self replicating. However, the replication process is not perfect and errors will occur on a small fraction of the nucleotides. These errors are known as mu-

	T		C		A		G		
	Codon	AA	Codon	AA	Codon	AA	Codon	AA	
T	TTT	F	TCT	S	TAT	Y	TGT	C	T
	TTC	F	TCC	S	TAC	Y	TGC	C	C
	TTA	L	TCA	S	TAA	*STOP*	TGA	*STOP*	A
	TTG	L	TCG	S	TAG	*STOP*	TGG	W	G
C	CTT	L	CCT	P	CAT	H	CGT	R	T
	CTC	L	CCC	P	CAC	H	CGC	R	C
	CTA	L	CCA	P	CAA	Q	CGA	R	A
	CTG	L	CCG	P	CAG	Q	CGG	R	G
A	ATT	I	ACT	T	AAT	N	AGT	S	T
	ATC	I	ACC	T	AAC	N	AGC	S	C
	ATA	I	ACA	T	AAA	K	AGA	R	A
	ATG	M	ACG	T	AAG	K	AGG	R	G
G	GTT	V	GCT	A	GAT	D	GGT	G	T
	GTC	V	GCC	A	GAC	D	GGC	G	C
	GTA	V	GCA	A	GAA	E	GGA	G	A
	GTG	V	GCG	A	GAG	E	GGG	G	G

Table 2.1: The Genetic Code: The 64 codons and the amino acids (AA) they encode. Some amino acids correspond to as many as six codons. Consequently not all nucleotide substitutions will result in an amino acid replacement.The process by which the codons produce the stated amino acid is more complex, involving transcription of the DNA into ribonucleic acid (RNA), followed by the translation of RNA into the amino acids. The details of these processes are beyond the scope of the thesis.

Type	Original DNA	New DNA
Substitution	GGACGAAT	GTACGAAT
Insertion	GGACGAAT	GGTACGAAT
Deletion	GGACGAAT	GACGAAT

Table 2.2: Common types of mutation. Substitution: the nucleotide G at the second site in the sequence has been substituted to a T. Insertion: an additional nucleotide, T, has been inserted after the second site in the sequence. Deletion: the nucleotide, G, at the second site in the sequence has been deleted.

tations. There are several different types of mutations, for example substitutions, insertions and deletions. A substitution occurs when a nucleotide is copied incorrectly and is replaced with a different nucleotide. As the names suggest, insertions and deletions occur when a nucleotide is inserted into, or deleted from, the sequence. Examples of these mutations can be found in Table 2.2. In the coding region of a sequence, insertions and deletions are highly likely to have a detrimental effect on the organism. This is because their effect is not limited to the codon where the insertion or deletion takes place. All nucleotides in the codons that follow will be shifted by one position (either forward or backward). However a substitution affects only the codon in which it takes place.

2.1.3 Substitutions and natural selection

Substitutions occurring in the DNA of an organism may or may not lead to an amino acid replacement. Referring to Table 2.1, consider the codon CCC which encodes the amino acid proline (P). A C↔T substitution at the third codon position does not result in an amino acid replacement, as codon CCT also encodes proline. However, a C↔T substitution at the first codon position does result in an amino acid replacement, as codon TCC encodes the amino acid serine (S). Substitutions which do not result in an amino acid replacement are called synonymous. Since synonymous substitutions have no immediate effect on the amino acid produced

they are of no functional importance to the organism. As such, synonymous substitutions have limited influence in evolution by natural selection. Substitutions that do result in an amino acid replacement are referred to as non-synonymous. Therefore non-synonymous substitutions have the potential to have a positive, negative or neutral effect on the organism. According to the theory of evolution by natural selection, amino acid replacements that negatively impact on an individual's fitness will decrease the likelihood of that individual reproducing. Consequently, that particular amino acid replacement is less likely to be passed on to future generations. This is known as purifying selection pressure: it acts to prevent deleterious non-synonymous substitutions proliferating throughout the population. Conversely, amino acid replacements that positively impact on an individual's fitness will increase the likelihood of that individual reproducing and passing on the mutation. Consequently, that particular amino acid replacement is more likely to be passed on to future generations. This is known as positive selection pressure: it promotes beneficial non-synonymous substitutions proliferating throughout the population. In this way, non-synonymous substitutions can be thought of as the raw material that powers the process of evolution.

2.1.4 Multiple sequence alignments

There are certain biological functions and processes that are critical to many species, *e.g.* carbohydrate metabolism. As such the proteins required for these functions are highly conserved across species that rely on them. Consequently, taxa that can appear vastly different morphologically will still have significant regions of identical DNA in their respective genomes, known as homologous regions. Large sections of highly conserved DNA over multiple taxa allow their sequences to be aligned. A set of two or more DNA sequences that have been aligned is referred to as a multiple sequence alignment (MSA). MSAs can be created at the nucleotide, codon or amino acid level. Table 2.3 displays a short, fictional example of a MSA. The pattern of

Taxa	1	2	3	4	5	6	7	8	9	10
Dog	A	A	T	G	G	T	T	T	C	T
Cat	A	T	T	T	G	A	G	T	C	G
Mouse	C	G	T	T	G	T	T	A	C	T
Horse	G	A	T	C	G	T	G	A	A	C

Table 2.3: An example of a multiple sequence alignment (MSA) for nucleotide data. The pattern of nucleotides at a particular site is known as a site pattern. For example, the site pattern at site four is GTTC.

nucleotides found across the taxa at a particular site is referred to as a site pattern. Phylogenetic inference is the process of analysing the site patterns of a MSA with the aim of inferring the evolutionary relationships that exist between the taxa.

The challenge of creating a MSA from a set of nucleotide sequences is non-trivial. A wide variety of computational algorithms are available to perform sequence alignment. Dynamic programming-based methods are computationally expensive but are more likely to produce the optimal sequence alignment (Smith and Waterman, 1981; Needleman, 1970). Alternatively, heuristic algorithms can be a more practical option, especially for large datasets, and are commonly employed in database alignment tools such as BLAST (Madden, 2013).

2.2 Inferring phylogenetic trees

The recent rapid advancement in DNA sequencing technology has provided phylogeneticists with a wealth of data to help inform evolutionary relationships. However drawing meaningful inferences from this data is no trivial task. An exhaustive search among potential tree topologies is not practical as the number of possible topologies increases rapidly as taxa are added. Semple and Steel (2003) show that the number

of unique, unrooted topologies for n taxa, $T(n)$, is given by:

$$T(n) = \frac{(2n-4)!}{(n-2)!2^{n-2}}. \tag{2.1}$$

To gain an appreciation of the implications of equation (2.1), the number of unique trees is given for up to 15 taxa in Table 2.4. Different methods of phylogenetic inference arrive at the best tree in different ways. Some methods build the tree from the data deterministically, whereas others employ innovative heuristics to search the tree space.

The general approach taken to infer phylogenetic trees is to select a statistical method for tree inference, assume an underlying model of sequence evolution (if required by the method), and produce a phylogenetic tree based on the method and model. Problems arise when we realise that different methods of tree inference may yield significantly different trees from the same MSA. There are a number of methods of tree inference used in phylogenetics, and in this thesis we focus on maximum parsimony (MP), neighbour joining (NJ) and maximum likelihood (ML).

2.2.1 Maximum parsimony

The application of the parsimony principle to the tree reconstruction problem was first presented by Edwards and Sforza (1963). The basic concept of MP is that the phylogenetic tree that explains the observed data with the minimum number of evolutionary changes (nucleotide substitutions) is the most parsimonious tree. It is effectively the application of Ockham's razor to the tree inference problem: the simplest tree that adequately explains the data is the best. MP grew in popularity as a method of tree reconstruction on the back of the Fitch algorithm (Fitch, 1971), a fast method for calculating the minimum number of state changes for a MSA on a given phylogeny. MP is non-parametric since no model of sequence evolution is required in order to construct the tree.

2.2.2 Neighbour joining

The NJ method of Saitou and Nei (1987) is a distance based clustering algorithm. The NJ method can be either non-parametric or parametric. It requires the selection of a distance metric in order to calculate pairwise distances between taxa. The chosen distance metric can be based on a model of sequence evolution (making NJ parametric), or not (making NJ non-parametric). Having calculated pairwise distances, the pair of taxa separated by the shortest distance are joined, creating a node on the tree. Distances between the new node and all remaining taxa and nodes are then calculated and the shortest distance defines which pair of taxa (or nodes) are joined next. The process is repeated until the tree is fully resolved.

2.2.3 Maximum likelihood

Felsenstein (Felsenstein, 1973) first applied the statistical principle of ML to the phylogenetic tree inference problem, detailing how to calculate the likelihood of any tree given its topology and branch lengths. However, due to computational limitations, no suitable method was available to use this approach in practice. This was remedied when Felsenstein introduced his pruning algorithm (Felsenstein, 1981), a systematic and computationally efficient method of tree inference based on ML. The pruning algorithm relies on two key assumptions: that changes in the sequence at different sites are independent probabilistic events; and that the model of sequence evolution is time reversible (the likelihood of observing a substitution from state i to j is the same as the likelihood of observing a substitution from state j to i).

The probabilities of nucleotide substitution are modelled by a Markov process, so that the probability of transitioning from state i to state j in time period t depends only on states i and j and the length of time t, not on any historical states. The time reversibility of the Markov process can be used to formulate the Pulley Principle (Felsenstein, 1981). This principle establishes that the likelihood of a given tree is independent of the placement of the root on the tree. The root of the tree can

be moved to any location on any edge on the tree and the likelihood will remain unchanged.

As Table 2.4 emphasises, unless considering only a small number of taxa an exhaustive search of all possible tree topologies is not practical. An alternative method is therefore required and Felsenstein proposed starting with a three species tree and adding to it, one species at a time. As each species is added, it is placed in all possible positions (*i.e.* attached to all nodes) and a likelihood calculated. The position of the new species which results in the maximum likelihood for the tree is the one chosen. When the tree has more than four species, local rearrangement is carried out to try to improve the likelihood. There are several methods of local rearrangement described in the literature, *e.g.* nearest neighbour interchange (NNI), subtree pruning and regrafting (SPR) and tree bisection and reconnection (TBR). For a detailed explanation of these methods see Felsenstein (2004). When no further improvement is possible, the next species is added to the tree and the process is repeated. For each tree topology the edge lengths must be optimised in order to find the true likelihood score for that tree. Felsenstein (1981) outlines an iterative method for achieving this optimisation. This technique ensures that a thorough search of plausible topologies is carried out, without wasting computational resources evaluating topologies that lack plausibility.

2.2.4 Bayesian Markov-chain Monte-Carlo

Bayesian Markov-chain Monte-Carlo (BMCMC) offers a fundamentally different approach to phylogenetic inference than that of ML. BMCMC makes no effort to optimize tree topology analytically. Rather, it defines a prior distribution for the topology and samples from the distribution. From the sample, the likelihood of the data is computed and, based on some predefined criterion, the tree is either accepted or rejected. If it is accepted the tree is added to the posterior distribution. The process is then repeated thousands or perhaps millions of times. In theory the

Number of taxa	Number of unique unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
11	3.446×10^7
12	6.547×10^8
13	1.375×10^{10}
14	3.162×10^{11}
15	7.906×10^{12}

Table 2.4: Number of unique unrooted phylogenetic trees for a given number of taxa

proportion of time a tree appears in the posterior distribution can be thought of as the posterior probability of that tree being correct. This process is most commonly implemented via the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970).

The advantage of BMCMC is that it is able to be applied to complex models that are not well suited to ML. For example, if a model of evolution is not time reversible then traditional ML optimization is not possible, due to the reliance on Felsenstein's pruning algorithm (Felsenstein, 1981). The primary disadvantage of BMCMC is the computation time required, particularly for large datasets or when many iterations are required to establish the posterior distribution.

2.3 Models of sequence evolution

The assumption of an evolutionary model is a necessary prerequisite for performing phylogenetic inference with a parametric method. Models of sequence evolution summarise the evolutionary process in the form of a continuous time Markov chain, defined by a rate matrix Q and the consequential probability transition function $P(t)$. Additionally, the models include a vector of base frequencies, $\boldsymbol{\pi} = \{\pi_A, \pi_G, \pi_C, \pi_T\}$. Ideally the model chosen should closely approximate the actual process of evolution that gave rise to the MSA. However, developing a model of evolution that accurately represents the actual evolutionary processes that have taken place in the natural world is a significant challenge. Furthermore, even if such a model was achievable it is likely to be prohibitively complex and over parameterised. In practice, assumptions need to be made to simplify models so that they are practical, but this can also come at a cost in that a poor model may lead to inferring the incorrect tree. The development of models of sequence evolution has been ongoing for more than four decades. The earliest models were simplistic, assuming equal base frequencies and homogeneity of substitution rates across sites (*i.e.* substitutions occur at the same rate on all sites in the alignment). These oversimplifications of the evolutionary

process were necessary in their day, however the increase of computational capacity has resulted in the proposal and implementation of more complex heterogeneous models. A detailed review of these models has been provided by Lio and Goldman (1998).

Homogeneous models

The simplest model of sequence evolution is the Jukes Cantor (JC) model (Jukes and Cantor, 1969). The JC model assumes equal base frequencies (that the historical sequence was composed of equal proportions of all four nucleotides) and that the rate at which nucleotide i is substituted and replaced by nucleotide j is independent of nucleotides i and j. In other words the overall substitution rate is given by μ, and the substitutions are split evenly between the 4 nucleotides. The rate matrix then looks like:

$$Q = \begin{bmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{bmatrix}. \tag{2.2}$$

The diagonals of the rate matrix are set such that the column sums are 0. So in this case each diagonal entry would be $-\frac{3\mu}{4}$. The rate matrix is then used to derive the probability transition function:

$$P(\nu) = e^{Q\nu} \tag{2.3}$$

$$= \begin{bmatrix} P_{AA}(\nu) & P_{GA}(\nu) & P_{CA}(\nu) & P_{TA}(\nu) \\ P_{AG}(\nu) & P_{GG}(\nu) & P_{CG}(\nu) & P_{TG}(\nu) \\ P_{AC}(\nu) & P_{GC}(\nu) & P_{CC}(\nu) & P_{TC}(\nu) \\ P_{AT}(\nu) & P_{GT}(\nu) & P_{CT}(\nu) & P_{TT}(\nu) \end{bmatrix}, \tag{2.4}$$

where $P_{xy}(\nu)$ is the probability of a site transitioning from state x to state y over time period (expressed in terms of branch length) ν, where x and y are taken from

the four nucleotide state space $\{A, G, C, T\}$. The requirement of time reversibility reduces to the constraint:

$$\pi_i P_{ij}(\nu) = \pi_j P_{ji}(\nu),$$

for all states i and j. Essentially this means that the probability of starting in state i and transitioning to state j over a branch of length ν, is equal to the probability of starting in state j and transitioning to state i over the same branch.

The JC model assumption of equal substitution rates among all pairs of nucleotides is not very realistic. Molecular biologists have defined two types of nucleotide substitution, known as transitions and transversions. A transition involves an A↔G or C↔T substitution whereas a transversion involves either an A↔C, A↔T, G↔C or G↔T substitution. The distinction is necessary because of the molecular structure of the nucleotides: A and G (known as purines) have a similar structure as do C and T (known as pyrimidines). Consequently transitions occur more frequently than transversions, even though there are twice as many possible transversions as transitions. Kimura (1980) recognised this deficiency in the JC model and proposed the K80 model, which allowed two different substitution rates, one for transitions and the other for transversions. He later extended this model to the K81 (Kimura, 1981), which introduced a third rate parameter. The allocation of rates to substitution types can be seen in Figure 2.2.

The JC, K80 and K81 all assume equal base frequencies in the ancestral sequence. A different approach was proposed by Felsenstein (1981) in which this assumption was relaxed. The F81 model assumes that the rate of substitution to a particular nucleotide depends only on that nucleotides equilibrium frequency which can be estimated by analysing the base frequencies of the available data. The HKY85 model Hasegawa *et al.* (1985) combined the features of the K80 and F81, introducing a transition/transversion bias to Felsenstein's model. This model was extended by Tamura and Nei (1993), introducing a model (TrN) which incorporated unequal base frequencies and three substitution rates. All of the models mentioned thus

far are special cases of the General Time-Reversible (GTR) model (Lanave *et al.*, 1984). As the name suggests, the GTR model is the most general model possible while maintaining the desirable quality of reversibility. It contains nine parameters in total: six substitution rate parameters and three base frequency parameters. Typically, most phylogenetic inference programs fix one of the substitution rate parameters and infer the other five relative to the fixed one. Branch lengths are parameterised as the mean number of substitutions per site so that the total tree length is indicative of the evolutionary rate. This reduces the number of substitution model parameters to be inferred to eight.

2.3.1 Heterogeneous models

The homogeneous models discussed above do facilitate computation but they are not supported by empirical evidence. The assumption that each site evolves at the same rate is not reasonable, most obviously because it relies on all sites being variable. Fitch and Margoliash (1967) pointed out that it is likely that many nucleotide sites are necessarily fixed because of functional constraints. To arrive at this realisation one can imagine a codon that encodes a particular amino acid that is absolutely critical for the fitness of the individual. Any non-synonymous substitution occurring within this codon will not proliferate throughout the population as the individual will not survive long enough to reproduce. Hence this codon will become fixed in the population. Even taking into account the redundancy inherent in the Genetic Code (Table 2.1), in effect this will mean that at least one nucleotide in that codon is invariable. As such, the presence of any amino acid that is essential to reproduction (either directly or indirectly) is sufficient to show that some sites in an alignment are invariable. To accommodate this fact the incorporation of a certain proportion of invariable sites into substitution models (Churchill *et al.*, 1992; Adachi and Hasegawa, 1995; Reeves, 1992; Swofford *et al.*, 1996) is now commonplace.

The inclusion of invariable sites only addresses the most obvious flaw in rate

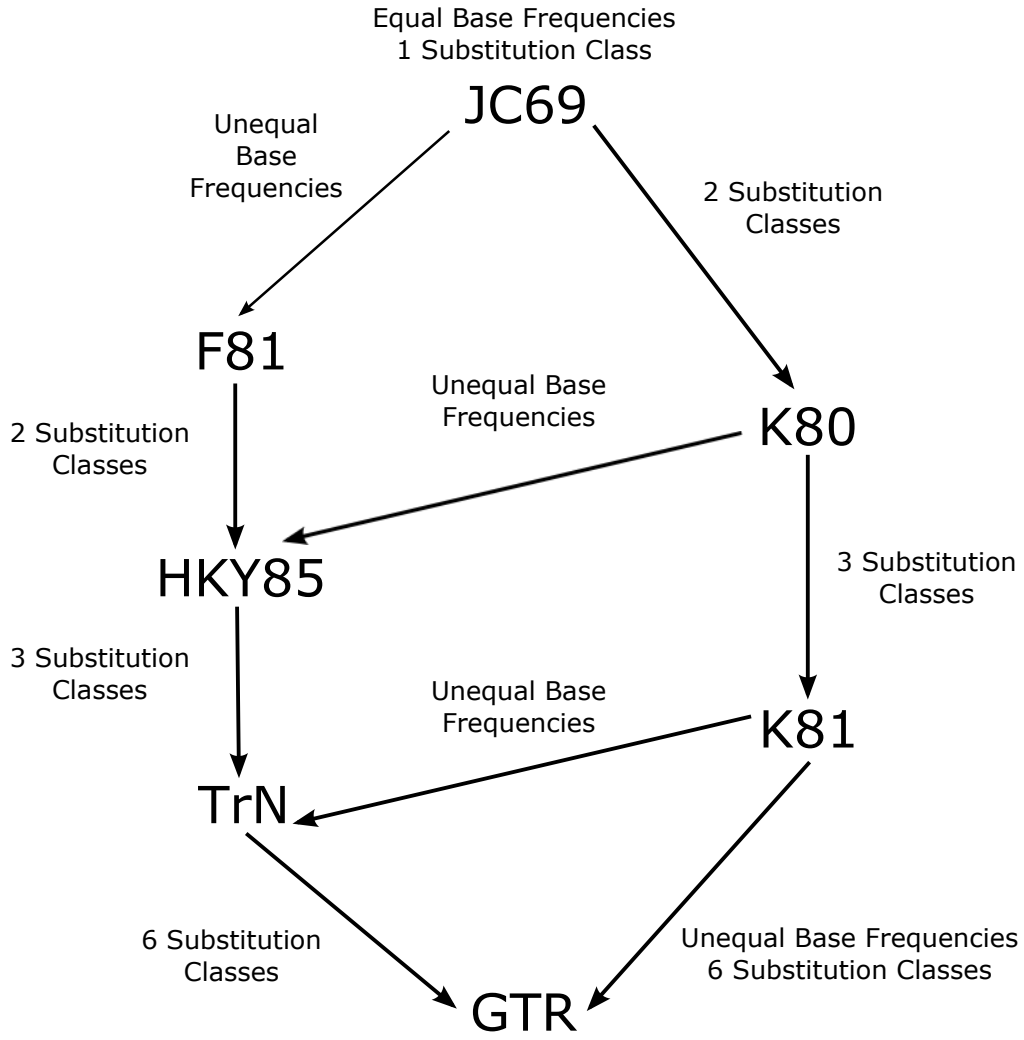

Figure 2.1: Flowchart depicting the relationships of the models of sequence evolution to each other. Models on the right hand side are those that maintain equal base frequencies while those on the left allow for unequal base frequencies.

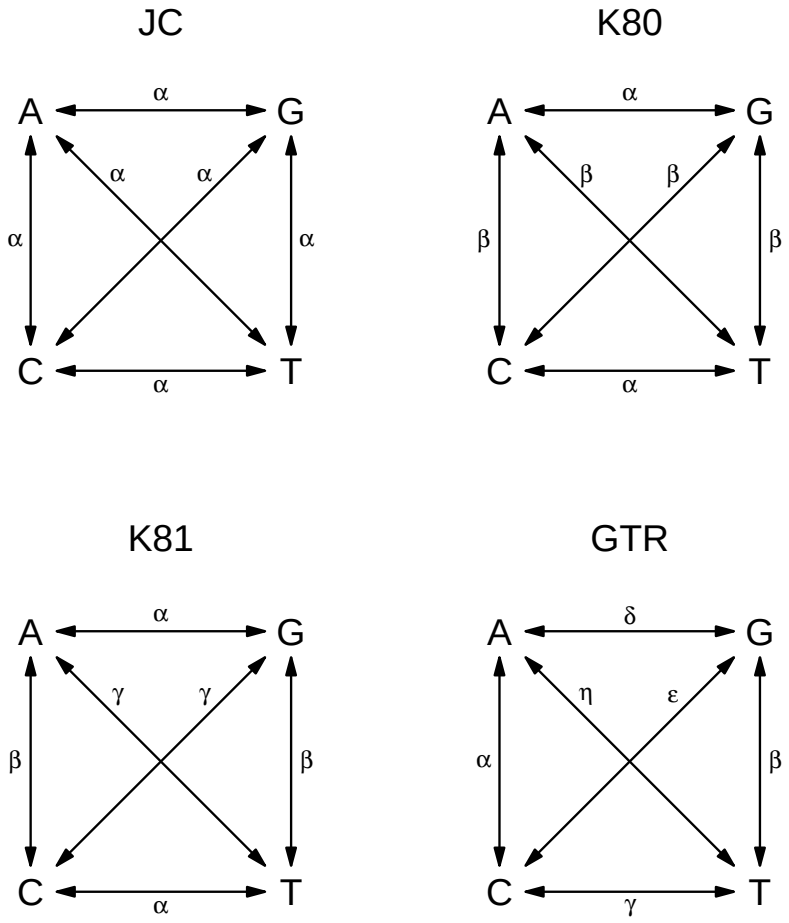

Figure 2.2: Schematic displaying the increasing complexity of the arrangement of substitution rate parameters for a selection of models of nucleotide sequence evolution. The JC model is the simplest, all substitutions share the same rate parameter α. The K80 introduces a second rate parameter, so that transitions occur at rate α and transversions at rate β. The K81 introduces a third parameter so that transversions now occur either at rate β or rate γ. Finally the GTR model defines a separate rate parameter for each of the six unique pairs of nucleotides. This is the most general model possible while still maintaining the desirable property of time reversibility.

homogeneous models. In reality we would expect that the substitution rates of different sites exist on a continuum between invariable and fast evolving. Yang (1993) introduced the gamma model, which assumes the evolutionary rates of sites in the MSA follow a gamma distribution. This model was difficult to implement computationally, due to the requirement to integrate over the range of the gamma distribution when evaluating the likelihood function. Yang overcame this problem by proposing a discrete approximation of his gamma model (Yang, 1994), which includes a number of different rate classes. The rates of the classes are defined relative to each other (maintaining a mean rate equal to one), with the scalar multiples drawn from a gamma distribution. The discrete gamma model has often been combined with the practice of inferring a proportion of invariable sites (Gu *et al.*, 1995; Waddell and Penny, 1996). This has proven to be an economical way of achieving significant improvements in the fit between model and data. Such a model requires the addition of only two parameters: a proportion of invariable sites; and the shape parameter, α, of the gamma distribution. For this relatively cheap price one gains tremendous flexibility in the establishment of a class of invariable sites in addition to several classes of sites evolving at different speeds.

The discrete gamma model only addresses rate heterogeneity across sites. It does not account for compositional heterogeneity: when the sequences of different taxa are composed of different base frequencies. This has been shown to mislead phylogenetic inference when it is not adequately modelled (Hasegawa and Hashimoto, 1993; Lockhart *et al.*, 1994; Chang and Campbell, 2000; Tarrío *et al.*, 2001; Ho and Jermiin, 2004; Jermiin *et al.*, 2004). Additionally, it does not account for rate heterogeneity across lineages: when a specific site has different rates on different branches of the tree. This is known as heterotachy and will be addressed in more detail in Chapter 3. This feature of sequence data is much harder to model, there are no economical solutions such as that offered by the discrete gamma model. As such heterotachy is often ignored when conducting phylogenetic inference, resulting in significant model misspecification. Studies have demonstrated the significant

detrimental effect imposed by model misspecification on the reliability of phylogenetic inference. Ho and Jermiin (2004) provide a thorough example. They generated seven data sets, each with a different substitution model to simulate evolution under a variety of conditions incorporating compositional heterogeneity, rate heterogeneity across sites, rate heterogeneity across lineages and combinations thereof. They included a null case of compositional and rate homogeneity across all sites and lineages. They then applied a number of different phylogenetic inference methods to the data, including MP, NJ and ML. They concluded that if the model of evolution assumed in the phylogenetic analysis does not reasonably approximate the actual evolutionary conditions which gave rise to the sequences, then the probability of inferring the correct tree is significantly reduced. A further example is provided by Shavit Grievink *et al.* (2010). They aimed to evaluate the effect of changes in the proportion and positions of variable sites on model fit and tree estimation. Data was simulated on a 16 taxon tree. Incorporated in the tree were two events at which a certain proportion of sites that were classified as invariable were switched to become variable. Generally speaking, they found that as the proportion of sites that switch from invariable to variable increases, the less accurate the phylogenetic tree inference becomes. They also found that when this proportion was greater than 20% the incorrect tree is inferred most often.

Heterotachy presents itself as a significant obstacle to accurate and reliable tree reconstruction. The challenge is twofold. The development of complex models that can accommodate heterotachously-evolved sequence data is necessary, but this is only the first step. Once developed these models must drive advancements in computational phylogenetics so that models that would once have been considered prohibitively complex can be efficiently implemented within phylogenetic software. These challenges are the focus of the remainder of the thesis.

Chapter 3

Heterotachy

The success and reliability of phylogenetic inference methods is limited to a large extent by the accuracy of the evolutionary models that are assumed in the inference process. Homogeneous evolutionary models have long been recognised as inadequate and ever more complex models have been introduced in the literature to try to account for the heterogeneous nature of sequence evolution. Popular models such as the gamma model (Yang, 1993) account for heterogeneity of evolutionary rate across sites in an alignment, but they still assume a constant substitution rate for each site across all lineages. This is too restrictive. There is a significant body of research suggesting variability in substitution rates across lineages, known as heterotachy (Lopez *et al.*, 2002), is the norm rather than the exception (Lopez *et al.*, 2002; Baele *et al.*, 2006; Wu and Susko, 2011). Biologically speaking it seems apparent that heterotachy is more plausible than homotachy, given the rate of mutation of a site is correlated to the amount of selective pressure acting on that site.

The effect of heterotachy on phylogenetic inference using traditional models and methods has been a topic of significant debate. Kolaczkowski and Thornton (2004) found that parametric tree fitting methods such as maximum likelihood (ML) and Bayesian Monte-Carlo Markov Chain (BMCMC) were outperformed by maximum parsimony (MP) when the data evolved heterotachously. This study was widely challenged on the grounds that the simulations captured only a special case of het-

erotachy (Spencer *et al.*, 2005; Steel, 2005; Gadagkar and Kumar, 2005; Philippe *et al.*, 2005) and more general studies of heterotachy concluded that ML performed at least as well and in most cases better than MP (Spencer *et al.*, 2005; Gadagkar and Kumar, 2005). Lockhart *et al.* (2006) attempted to differentiate between types of heterotachy, pointing out that the term heterotachy has been used to describe the empirically inferred variation in p_{var} among orthologues and paralogues. This is distinct from the somewhat more theoretical form of heterotachy simulated by Kolaczkowski and Thornton (2004). Lockhart *et al.* (2006) found that for an empirical case with plastid and eubacterial polymerase sequences, p_{var} could differ between orthologues in anciently diverged evolutionary lineages. The authors found parsimony and a least squares method to be susceptible to long branch attraction (Felsenstein, 1978), whereas ML was robust to model misspecification in this particular case.

The development of models that can handle heterotachous data has been fertile ground. One approach has been covarion (COV) models (Fitch and Markowitz, 1970). Tuffley and Steel (1998) described a model in which sites could switch between variable and invariable states in different lineages. All variable sites in the model shared a common substitution model and rate. This model was gradually extended (Galtier, 2001; Huelsenbeck, 2002), eventually reaching its most complex form in which sites can switch along lineages between a number of different rates as well as an invariable state (Wang *et al.*, 2007).

Partition models can be used when the data can be sensibly partitioned before analysis (Yang, 1996; Pupko *et al.*, 2002). Datasets might naturally lend themselves to being partitioned, based on coding/regulatory regions, codon position or gene boundaries. The edge unlinked partition model is the most complex version which allows for inference of separate branch lengths, substitution models and evolutionary rates for each partition.

If a sensible partitioning of the dataset is not possible *a priori* then a mixture model can be used. Mixture models are apt for handling heterotachous data as they provide flexibility, allowing the model to capture much more of the variability in

the sequences. Matsen and Steel (2007) asserted that caution should be exercised when applying mixture models to the tree reconstruction problem. They showed that data evolved on a mixture of branch lengths on one topology could be indistinguishable from unmixed data evolved on a different topology. Further work in this area from Allman *et al.* (2011) showed that this result was limited to 2-state models of character change on 4-taxon trees. They went on to show algebraically that when using 4-state models of character change (such as DNA) tree and model parameters were in fact identifiable for data evolved on two tree mixtures. Several mixture models have been proposed in the literature to combat the problem of heterotachous evolution, differentiated by the parameters that are free to vary across mixture components. Mixed branch length (MBL) models allow branch lengths to vary on a fixed topology across mixture components (Kolaczkowski and Thornton, 2004; Spencer *et al.*, 2005; Meade and Pagel, 2008). Another approach is mixed rate matrix (MRM) models, which allow the substitution rate matrix to vary between mixture components (Lartillot and Philippe, 2004; Pagel and Meade, 2004). Wu and Susko (2009) proposed a general framework for heterotachy, at the time encompassing all of the models discussed so far as special cases. The HAL-HAS model (Jayaswal *et al.*, 2014) is the most complex mixture model proposed in the literature to date. It addresses heterogeneity across lineages (HAL) and sites (HAS) in a two stage process. A consistent drawback of these models is the computational expense. Mixture models are parameter rich and, as such, it is difficult to come up with suitably efficient heuristics to search both the tree and parameter space. For this reason these types of models have often been implemented within a Bayesian framework (Pagel and Meade, 2004; Meade and Pagel, 2008) and without tree search capability (Jayaswal *et al.*, 2014).

The high dimensional nature of mixture models is beneficial as it allows more freedom in modelling the infinitely complex process of evolution. Steel (2005) cautions against this approach, warning that the quest for biological realism should not drive the development of models with an ever increasing number of parameters.

Parameter estimation comes at a computational cost, as well as a loss of predictive power. However one could argue that the cost of simple models is a bigger problem for phylogeneticists. The vast model misspecification that exists between simple evolutionary models and natural evolutionary processes is of greater concern. One reason for parameter restriction is that over-parameterised models compromise the predictive capacity of the model. This may be true but how many parameters is too many remains unclear. Another argument for simpler models is the amount of data and computational resources required to resolve many parameters. This is becoming less of a concern with next generation sequencing techniques leading to a rapid expansion in the quality and quantity of biological sequences. At the same time computer scientists are developing faster, more efficient algorithms in order to handle them (Stamatakis *et al.*, 2005; Stamatakis, 2006; Nguyen *et al.*, 2015; Guindon and Gascuel, 2003).

3.1 Simulating heterotachously-evolved alignments

It is well known that model misspecification due to heterotachously-evolved sequence data reduces the accuracy of phylogenetic inference using current models and methods. What is not so well understood is that the effect of heterotachy on phylogenetic inference can be subtly different, depending on the model and method chosen. We carried out a simulation study to investigate in detail the effects of a simple form of heterotachy on three different phylogenetic inference methods, MP, Neighbour Joining (NJ) and ML.

We simulated heterotachously-evolved multiple sequence alignments (MSAs) on a four taxon tree, following the same method as Shavit Grievink *et al.* (2008). The tree is shown in Figure 3.1. Branch lengths used for the simulations were: $a = 0.4, b_1 = 0.1, b_2 = 0.3, c_1 = 0.1, c_2 = 0.3, d = 0.4$ and $k = 0.1$. We refer to sites that are free to vary across all lineages of the tree as variable sites, and those that are not free to vary on any lineages as invariable. The proportion of

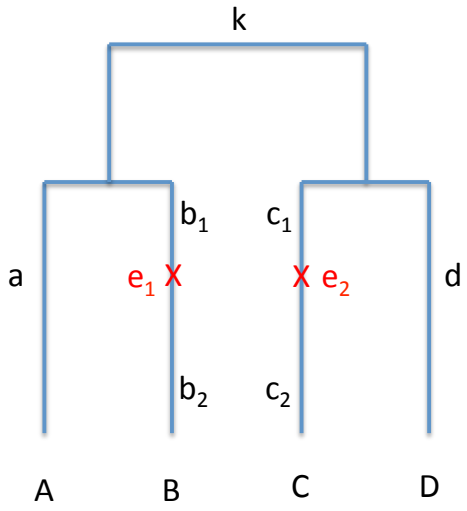

Figure 3.1: Four taxon tree with two events, e_1 and e_2 at which a certain percentage of invariable sites are switched to become variable. Branch lengths used for the simulations were: $a = 0.4, b_1 = 0.1, b_2 = 0.3, c_1 = 0.1, c_2 = 0.3, d = 0.4$ and $k = 0.1$.

variable sites, p_{var}, was fixed at 0.2 for all simulations. The variable sites evolved according to the Jukes Cantor (JC) (Jukes and Cantor, 1969) model of evolution. Heterotachy was introduced into the simulation procedure by defining the remaining sites as either invariable or heterotachous. The heterotachous sites were invariable on the majority of the tree, but switched from invariable to variable at the two switching events (labelled e_1 and e_2 on Figure 3.1) located on lineage B and C. The proportion of heterotachous sites, p_{het} was varied from 0 to 0.8 at increments of 0.008 with the remainder being the proportion of invariable sites, p_{inv}. At each value of p_{het} 100 MSAs were simulated using LineageSpecificSeqgen (Shavit Grievink *et al.*, 2008). The sequence length was fixed at 100,000 base pairs. In keeping with the technique adopted by Shavit Grievink *et al.* (2008), the branch length was defined as the average number of substitutions per variable site, as opposed to the more common definition: average number of substitutions per site. We used Phylip v3.69 (Felsenstein, 2002) to infer trees for each of the MSAs using MP; NJ with the JC distance metric; and ML. We measured the success of each method by calculating

the fraction of MSAs for which the correct topology (AB|CD) was inferred, at each value of p_{het}.

3.2 Maximum parsimony

The results for MP can be seen in Figure 3.2. MP performs well for $p_{het} < 0.2$, inferring the correct topology for all MSAs. However for $p_{het} \geq 0.21$ the success rate of MP decreases rapidly until the incorrect AD|BC topology is inferred in 100% of the MSAs for $p_{het} \geq 0.35$. As discussed in Section 2.2.1 maximum parsimony (MP) is a non-parameteric method of phylogenetic inference. This means that we do not need to assume a model of sequence evolution and therefore model misspecification plays no role in the poor performance of MP. As the name suggests, the tree ultimately inferred by MP is the one which requires the fewest substitutions to produce the sequences in the MSA. We can analyse the parsimony scores on the correct and incorrect topologies for every site in an alignment, thereby establishing which site patterns support the correct topology and which support the incorrect topologies. We can then calculate the probability of observing favourable or unfavourable site patterns as a function of p_{het} and find the critical value of p_{het} beyond which MP will infer the incorrect tree given sufficient sequence length.

3.2.1 Definitions

1. Let F be the event that a particular site pattern provides evidence for the correct AB|CD tree.

2. Let A be the event that a particular site pattern contributes evidence for the incorrect AD|BC tree.

3. Let I be the event that a particular site is invariable over the entire tree.

4. Let V be the event that a particular site is variable over the entire tree.

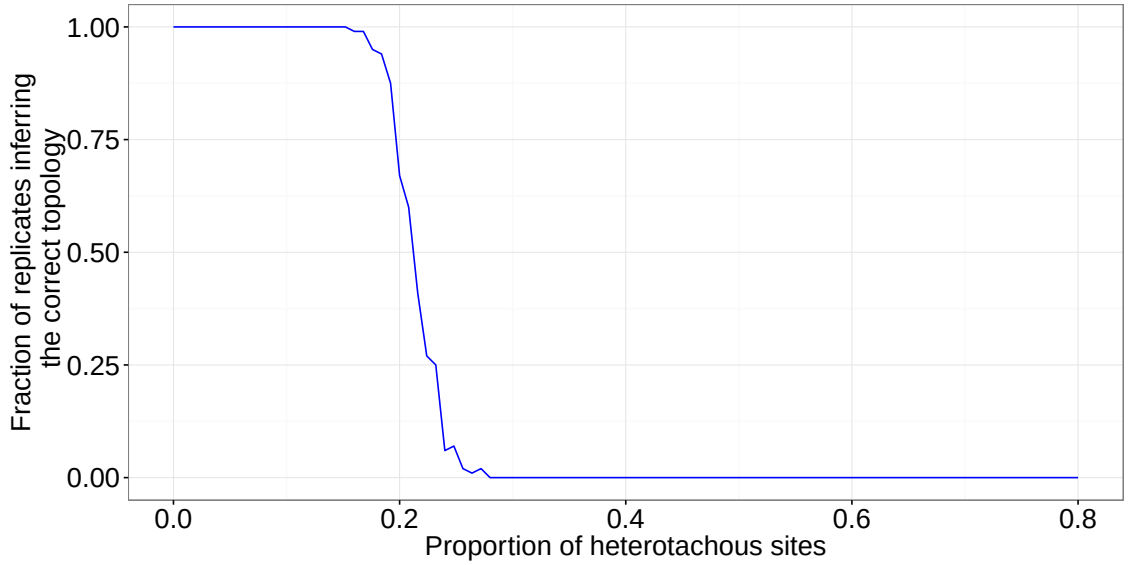

Figure 3.2: MP results for the four taxa simulation study on heterotachously-evolved 100,000bp MSAs. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 at increments of 0.008. At each value, 100 MSAs were simulated. The y-axis reports the fraction of MSAs for which MP inferred the correct tree topology.

5. Let H be the event that a site is heterotachous, that is it switches from invariable to variable at both events e_1 and e_2.

Applying the law of total probability yields:

$$P(F) = p_{inv}P(F|I) + p_{var}P(F|V) + p_{het}P(F|H), \quad (3.1)$$

$$P(A) = p_{inv}P(A|I) + p_{var}P(A|V) + p_{het}P(A|H). \quad (3.2)$$

Finally, we introduce the notation $P_{ii}(\nu)$ to define the probability that a particular site will exhibit the same nucleotide at two points on the tree separated by branch length ν, and $P_{ij}(\nu)$ to define the probability that a particular site will exhibit a specific pair of differing nucleotides at two points on the tree separated by branch length ν. Under the JC model of evolution, we know that

$$P_{ii}(\nu) = \frac{1}{4} + \frac{3}{4}e^{-4\nu/3}, \quad (3.3)$$

$$P_{ij}(\nu) = \frac{1}{4} - \frac{1}{4}e^{-4\nu/3}. \quad (3.4)$$

Essentially, $P_{ii}(\nu)$ is equal to the diagonal elements of the probability transition matrix for the JC model and $P_{ij}(\nu)$ is equal to its off diagonal elements.

3.2.2 Elucidation of events F and A

In order to progress further, it is necessary to understand exactly what constitutes evidence for (event F) and against (event A) the correct tree under the MP method. Given our four taxa example, there are three possible bifurcating trees that could be inferred. They are the AB|CD tree (correct), the AD|BC tree and the AC|BD tree (both incorrect). However it was found during experimentation that the AC|BD tree was never inferred for long sequence lengths. It is easy to see why by looking at Figure 3.1. As the proportion of heterotachous sites increases, the evolutionary distance between taxa A and B increases, due to the greater opportunity for nucleotide substitution. At some point this could create the situation where the expected distance (by some appropriate metric) between taxa A and B is actually greater than the expected distance between taxa A and D. However the situation can never occur whereby the expected distance between taxa A and B exceeds that between taxa A and C, because the switching sites have the same diverging effect on the AC sequences as the AB sequences, and the AC sequences have the additional substitution opportunity of the inner branch length. For this reason we shall only consider the AD|BC topology as the incorrect tree, ignoring the very small probability that MP will infer the AC|BD tree.

We now need to understand what specific site patterns constitute evidence for the correct tree, and what site patterns constitute evidence for the incorrect tree. Given our MSAs contain four taxa and there are four character states for nucleotide data, there are $4^4 = 256$ different site patterns that we can potentially observe in an alignment. However, not all of these site patterns are unique from a MP perspective. MP is interested only in minimizing substitutions, it does not favour or penalise one type of substitution over another. So observing the site pattern $AAAG$

is identical to observing the site patterns $AAAT$, $GGGA$ or $CCCT$ for example. Each of these patterns can be represented by the 'generic' site pattern $xxxy$, where x and y can be any of the four character states, subject to the condition $x \neq y$. A generic site pattern will always be written corresponding to the taxon order of the correct topology. So the generic site pattern $wxyz$ is always interpreted as: w as the nucleotide present in taxon A, x is the nucleotide present in taxon B, y is the nucleotide present in taxon C and z is the nucleotide present in taxon D for one specific site. This can be confusing when referring to the incorrect topology so care should be taken. Site patterns of the form $xxxy$ require a minimum of one substitution (regardless of topology) so they would have a parsimony score of one. The 256 unique site patterns simplify to 15 different generic site patterns. These are listed in Table 3.1 along with their associated parsimony scores for the AB|CD and AD|BC topologies. Table 3.1 indicates that 13 of the 15 generic site patterns have the same parsimony score on both topologies. We refer to these as uninformative sites as they do not favour either topology over the other. Under the AB|CD topology, the site pattern $xxyy$ can be produced by a single substitution on the internal edge, resulting in a parsimony score of one. However, under the AD|BC topology the site pattern $xxyy$ requires a minimum of two substitutions (one on either the B and C or the A and D terminal branches), resulting in a parsimony score of two. This is the only site pattern for which the correct topology has a lower parsimony score than the incorrect topology. Consequently the event F is synonymous with observing an $xxyy$ site pattern. Analagously, the $xyyx$ is the only generic site pattern for which the correct topology has a higher parsimony score than the incorrect topology. Consequently the event A is synonymous with observing the site pattern $xyyx$.

Site Pattern	Parsimony score AB|CD	AD|BC
xxxx	0	0
xxxy	1	1
xxyx	1	1
xyxx	1	1
yxxx	1	1
xxyy	1	2
xyyx	2	1
xyxy	2	2
xxyz	2	2
xyzx	2	2
xyxz	2	2
yxxz	2	2
yxzx	2	2
yzxx	2	2
wxyz	3	3

Table 3.1: Parsimony scores of the 15 generic site patterns for the correct AB|CD topology and the incorrect AD|BC topology. Highlighted in blue, $xxyy$ is the only site pattern for which the correct tree is more parsimonious than the incorrect tree, whereas highlighted in red, $xyyx$ is the only site pattern for which the incorrect tree is more parsimonious than the correct tree. All other site patterns are uninformative.

3.2.3 Evidence for and against the correct tree

To determine the evidence for and against the correct tree we must evaluate equations (3.1) and (3.2). To do this we must establish expressions for $P(F|I)$, $P(A|I)$, $P(F|V)$, $P(A|V)$, $P(F|H)$ and $P(A|H)$.

Recall firstly that the event I indicates that a particular site is invariable over the entire tree. This means that for such a site the only possible site pattern is $xxxx$. There is no chance for any nucleotide substitution to occur at this site in any of the taxa. Therefore we can conclude that:

$$\begin{aligned} P(F|I) &= 0, \\ P(A|I) &= 0. \end{aligned}$$

Next recall that the event H indicates that a site is a switching site. This means that the site is invariable everywhere on the tree until being switched to variable at the two events (see Figure 3.1). Therefore the only opportunity for nucleotide substitution is along both edges b_2 and c_2. It then follows with certainty that the site will contain the same nucleotide at taxa A and D, leading us to conclude that $P(F|H) = 0$. It is however possible to obtain the site pattern $xyyx$, thereby supporting the incorrect tree. We are already guaranteed that taxa A and D will be in the same state, all that is required is for taxa B and C to also be in the same state and for that state to differ from taxa A and D. Put simply, we need an identical nucleotide substitution along edges b_2 and c_2 (Figure 3.1). Firstly, we consider taxon B and edge b_2. There are 3 possible target states for taxon B, each of them equally valid. All we require is that there is a change in state along edge b_2. Finally, we consider taxon C and edge c_2. Again there are 3 possible target states for taxon C however only one of them is valid, since we require taxa B and C to end up in the same state to establish our $xyyx$ site pattern. The probability of this occurring can be calculated by applying equation (3.4):

$$P(A|H) = 3P_{ij}(b_2)P_{ij}(c_2).$$

Finally, recall that the event V indicates that a site is variable over the entire tree. This means that for variable sites any site pattern is possible. Therefore the evaluation of $P(F|V)$ and $P(A|V)$ will be quite complex and will need to be approached methodically. We begin by defining another event, let S be the event that there is a nucleotide substitution at a particular site along the inner edge k. Put simply, S is the event that at a particular site the most recent common ancestor of taxa A and B has a different nucleotide from the most recent common ancestor of taxa C and D. It then follows that $\bar{S}$ is the event that these two most recent common ancestors share the same nucleotide at a particular site. Conditioning on S we can again apply the law of total probability to write:

$$\begin{aligned} P(F|V) &= P(S|V)P(F|V,S) + P(\bar{S}|V)P(F|V,\bar{S}), \\ P(A|V) &= P(S|V)P(A|V,S) + P(\bar{S}|V)P(A|V,\bar{S}). \end{aligned}$$

We can simplify the task by noting that in the event of $\bar{S}$ both informative site patterns $xxyy$ and $xyyx$ require at least 2 mutations to occur, and by symmetry both patterns are equally likely. Therefore we will abstain from deriving the expression and simply state that:

$$\begin{aligned} P(\bar{S}|V)P(F|V,\bar{S}) &= P(\bar{S}|V)P(A|V,\bar{S}) \\ &= C, \end{aligned}$$

which then yields:

$$\begin{aligned} P(F|V) &= P(S|V)P(F|V,S) + C, \\ P(A|V) &= P(S|V)P(A|V,S) + C. \end{aligned}$$

We also note that by again applying equation (3.4) we can establish that:

$$P(S|V) = 3P_{ij}(f).$$

It now remains to derive expressions for $P(F|V,S)$ and $P(A|V,S)$. There are unfortunately no short cuts to be found in doing this, the only way is to systematically

Description	Expression	W
No substitution on any lineage	$P_{ii}(a)P_{ii}(b_1+b_2)P_{ii}(c_1+c_2)P_{ii}(d)$	1
Substitutions on lineage A and B only	$3P_{ij}(a)3P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)P_{ii}(d)$	$\frac{2}{9}$
No substitution on lineage C and D only	$P_{ii}(a)P_{ii}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d)$	$\frac{2}{9}$
Substitutions on lineage A, B, C and D	$3P_{ij}(a)3P_{ij}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d)$	$\frac{7}{81}$

Table 3.2: Possible substitution combinations that will result in the site pattern $xxyy$, given a substitution has occurred along the internal edge, k. The weight, W, indicates the proportion of time that the described substitution combination will result in the $xxyy$ site pattern.

sum all the combinations of substitutions that result in the desired site pattern. We will commence with $P(F|V,S)$. There are four distinct ways that the site pattern $xxyy$ can be generated given the event S has occurred. These, along with their expressions, are listed in Table 3.2.

The weight given in the third column of Table 3.2 refers to the proportion of time that the given combination of substitutions will result in the $xxyy$ site pattern. By way of example, consider the situation involving identical substitutions on lineages A and B but no substitution on lineages C and D. There are three possible nucleotides that can end up in taxon A, however one of these nucleotides is the same as that in taxa C and D. If taxon A substitutes to this nucleotide, then the desired $xxyy$ site pattern is impossible. So given that there is a substitution in taxon A, there is a $\frac{2}{3}$ probability that this substitution will be valid in terms of creating the $xxyy$ site pattern. We then require the identical substitution in taxon B. So given that there is a substitution in taxon B, there is a $\frac{1}{3}$ probability that the $xxyy$ site pattern results. Since substitutions on lineages A and B are independent (under the JC model), the probability of these events occurring simultaneously is given by their product, $\frac{2}{9}$. The details will not be shown here but the same line of reasoning results in the $\frac{7}{81}$

Description	Expression	W
Substitutions on lineages A and C	$3P_{ij}(a)P_{ii}(b_1+b_2)3P_{ij}(c_1+c_2)P_{ii}(d)$	$\frac{1}{9}$
Substitutions on lineages B and D	$P_{ii}(a)3P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)3P_{ij}(d)$	$\frac{1}{9}$
Substitutions on lineages A, B and C	$3P_{ij}(a)3P_{ij}(b_1+b_2)3P_{ij}(c_1+c_2)P_{ii}(d)$	$\frac{2}{27}$
Substitutions on lineages A, B and D	$3P_{ij}(a)3P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)3P_{ij}(d)$	$\frac{2}{27}$
Substitutions on lineages A, C and D	$3P_{ij}(a)P_{ii}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d)$	$\frac{2}{27}$
Substitutions on lineages B, C and D	$P_{ii}(a)3P_{ij}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d)$	$\frac{2}{27}$
Substitutions on all four lineages	$3P_{ij}(a)3P_{ij}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d)$	$\frac{2}{81}$

Table 3.3: Possible substitution combinations that will result in the site pattern $xyyx$, given a substitution has not occurred along the internal edge, k. The weight, W, indicates the proportion of time that the described substitution combination will result in the $xyyx$ site pattern.

weighting on the last term.

The expressions given in Table 3.2 can then be combined to give:

$$\begin{aligned}
P(F|V) &= P(S|V)P(F|V,S)+C \\
&= 3P_{ij}(k)[P_{ii}(a)P_{ii}(b_1+b_2)P_{ii}(c_1+c_2)P_{ii}(d) \\
&\quad +\frac{2}{9}3P_{ij}(a)3P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)P_{ii}(d) \\
&\quad +\frac{2}{9}P_{ii}(a)P_{ii}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d) \\
&\quad +\frac{7}{81}3P_{ij}(a)3P_{ij}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d)]+C \\
&= 3P_{ij}(k)[P_{ii}(a)P_{ii}(b_1+b_2)P_{ii}(c_1+c_2)P_{ii}(d) \\
&\quad +2P_{ij}(a)P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)P_{ii}(d) \\
&\quad +2P_{ii}(a)P_{ii}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d) \\
&\quad +7P_{ij}(a)P_{ij}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d)]+C.
\end{aligned}$$

It now remains to carry out a similar process to obtain an expression for $P(A|V,S)$. This situation is unfortunately even more complex, the possible substitution combinations that can result in the $xyyx$ site pattern are detailed in Table 3.3.

Again the weight column refers to the proportion of time that the specified

substitution combination will result in the desired $xyyx$ site pattern. Although not inherently difficult, derivation of these weight values is quite involved and it is left to the keen reader to confirm their accuracy.

The expressions given in Table 3.3 can then be combined to give:

$$\begin{aligned}
P(A|V) &= P(S|V)P(A|V,S) + C \\
&= 3P_{ij}(k)[\frac{1}{9}3P_{ij}(a)P_{ii}(b_1+b_2)3P_{ij}(c_1+c_2)P_{ii}(d) \\
&\quad +\frac{1}{9}P_{ii}(a)3P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)3P_{ij}(d) \\
&\quad +\frac{2}{27}3P_{ij}(a)3P_{ij}(b_1+b_2)3P_{ij}(c_1+c_2)P_{ii}(d) \\
&\quad +\frac{2}{27}3P_{ij}(a)3P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)3P_{ij}(d) \\
&\quad +\frac{2}{27}3P_{ij}(a)P_{ii}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d) \\
&\quad +\frac{2}{27}P_{ii}(a)3P_{ij}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d) \\
&\quad +\frac{2}{81}3P_{ij}(a)3P_{ij}(b_1+b_2)3P_{ij}(c_1+c_2)3P_{ij}(d)] + C \\
&= 3P_{ij}(k)[P_{ij}(a)P_{ii}(b_1+b_2)P_{ij}(c_1+c_2)P_{ii}(d) \\
&\quad +P_{ii}(a)P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)P_{ij}(d) \\
&\quad +2P_{ij}(a)P_{ij}(b_1+b_2)P_{ij}(c_1+c_2)P_{ii}(d) \\
&\quad +2P_{ij}(a)P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)P_{ij}(d) \\
&\quad +2P_{ij}(a)P_{ii}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d) \\
&\quad +2P_{ii}(a)P_{ij}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d) \\
&\quad +2P_{ij}(a)P_{ij}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d)] + C.
\end{aligned}$$

We have now established expressions for all the ingredients of equations (3.1) and (3.2). Recalling that $P(F|I) = 0$, $P(A|I) = 0$ and $P(F|H) = 0$ substituting

the remaining terms yields:

$$\begin{aligned} P(F) &= p_{inv}P(F|I) + p_{var}P(F|V) + p_{het}P(F|H) \\ &= p_{var}\{3P_{ij}(k)[P_{ii}(a)P_{ii}(b_1+b_2)P_{ii}(c_1+c_2)P_{ii}(d) \\ &\quad +2P_{ij}(a)P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)P_{ii}(d) \\ &\quad +2P_{ii}(a)P_{ii}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d) \\ &\quad +7P_{ij}(a)P_{ij}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d)] + C\}, \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} P(A) &= p_{inv}P(A|I) + p_{var}P(A|V) + p_{het}P(A|H) \\ &= p_{var}\{3P_{ij}(k)[P_{ij}(a)P_{ii}(b_1+b_2)P_{ij}(c_1+c_2)P_{ii}(d) \\ &\quad +P_{ii}(a)P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)P_{ij}(d) \\ &\quad +2P_{ij}(a)P_{ij}(b_1+b_2)P_{ij}(c_1+c_2)P_{ii}(d) \\ &\quad +2P_{ij}(a)P_{ij}(b_1+b_2)P_{ii}(c_1+c_2)P_{ij}(d) \\ &\quad +2P_{ij}(a)P_{ii}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d) \\ &\quad +2P_{ii}(a)P_{ij}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d) \\ &\quad +2P_{ij}(a)P_{ij}(b_1+b_2)P_{ij}(c_1+c_2)P_{ij}(d)] + C\} \\ &\quad +3P_{ij}(g)P_{ij}(h)p_{het}. \end{aligned} \tag{3.6}$$

Derivation of Critical p_{het}

Equations (3.5) and (3.6) give expressions for $P(F)$ and $P(A)$ in their most general form, however they are very cumbersome and a great deal of simplification is possible if we apply the conditions of our simulated data, namely the branch lengths. Recalling from Figure 3.1 the branch lengths used were $a = 0.4, b_1 = 0.1, b_2 = 0.3, c_1 = 0.1, c_2 = 0.3, d = 0.4$ and $k = 0.1$. Notice that $a = b_1 + b_2 = c_1 + c_2 = d = 0.4$. Using this information the expressions for $P(F)$ and $P(A)$ simplify to become:

$$P(F) = p_{var}\{3P_{ij}(0.1)[P_{ii}(0.4)^4 + 4P_{ii}(0.4)^2P_{ij}(0.4)^2 + 7P_{ij}(0.4)^4] + C\},$$

and

$$\begin{aligned} P(A) &= p_{var}\{3P_{ij}(0.1)[2P_{ii}(0.4)^2P_{ij}(0.4)^2 + 8P_{ii}(0.4)P_{ij}(0.4)^3 + 2P_{ij}(0.4)^4] + C\} \\ &\quad +3P_{ij}(0.3)^2p_{het}. \end{aligned}$$

With these expressions the theoretical critical value of p_{het} should be readily attainable. It stands to reason that when the value of $P(F)$ exceeds that of $P(A)$ then the correct tree should be inferred for long sequence lengths. Similarly when the value of $P(A)$ exceeds that of $P(F)$ then the incorrect tree should be inferred for long sequence lengths. So we should be able to obtain the critical p_{het} value by finding when $P(F) - P(A) = 0$:

$$\begin{aligned} P(F) - P(A) &= p_{var}\{3P_{ij}(0.1)[P_{ii}(0.4)^2 + 4P_{ii}(0.4)^2P_{ij}(0.4)^2 + 7P_{ij}(0.4)^4] + C\} \\ &\quad -p_{var}\{3P_{ij}(0.1)[2P_{ii}(0.4)^2P_{ij}(0.4)^2 + 8P_{ii}(0.4)P_{ij}(0.4)^3 \\ &\quad +2P_{ij}(0.4)^4] + C\} + 3P_{ij}(0.3)^2p_{het} \\ &= p_{var}3P_{ij}(0.1)[P_{ii}(0.4)^2 + 2P_{ii}(0.4)^2P_{ij}(0.4)^2 - 8P_{ii}(0.4)P_{ij}(0.4)^3 \\ &\quad +5P_{ij}(0.4)^4] - 3P_{ij}(0.3)^2p_{het} \\ &= 0 \end{aligned} \tag{3.7}$$

when

$$\begin{aligned} p_{het} &= \frac{p_{var}P_{ij}(0.1)}{P_{ij}(0.3)^2}[P_{ii}(0.4)^2 + 2P_{ii}(0.4)^2P_{ij}(0.4)^2 \\ &\quad -8P_{ii}(0.4)P_{ij}(0.4)^3 + 5P_{ij}(0.4)^4]. \end{aligned}$$

We can now calculate the critical value of p_{het} by substituting in $p_{var} = 0.2$. This yields $p_{het} = 0.2125$. Figure 3.3 displays this critical value in relation to the results of the simulation study. There is clear agreement between the calculated theoretical value and the empirical results obtained in the simulation study.

3.3 Neighbour joining

The results of the NJ method are shown in Figure 3.4. For the most part NJ performs in a similar way to MP, consistently inferring the correct tree for $p_{het} <$

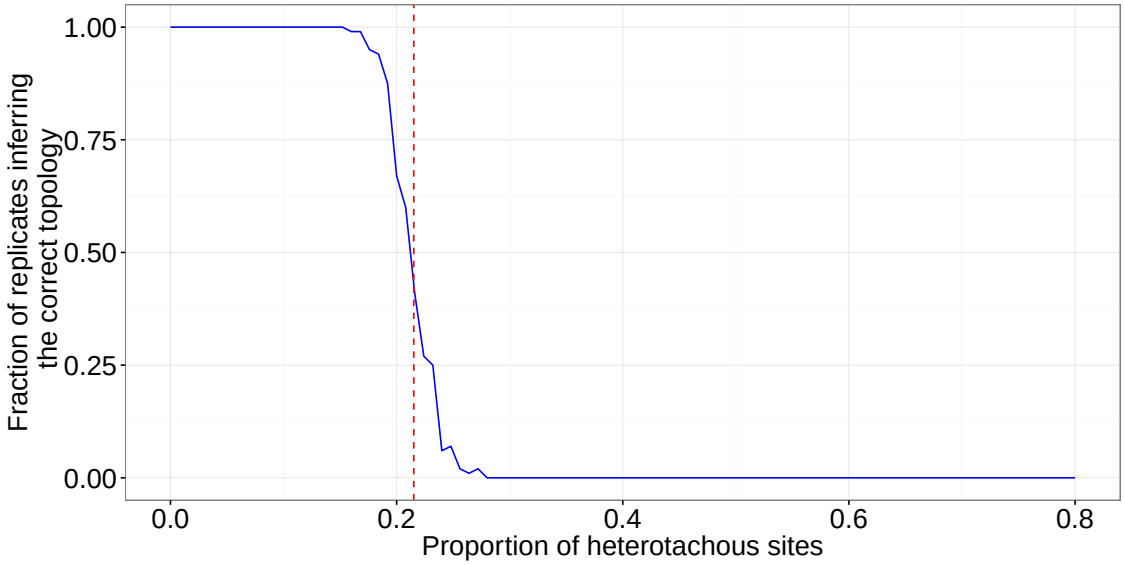

Figure 3.3: MP results for the 4-taxa simulation study on heterotchosly-evolved sequence data. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value, 100 MSAs were simulated. The y-axis reports the fraction of MSAs for which MP inferred the correct tree topology. The theoretical proportion of heterotachous sites at which MP should fail to recover the correct topology is shown by the dashed red line. Clearly the calculations concur with the empirical data.

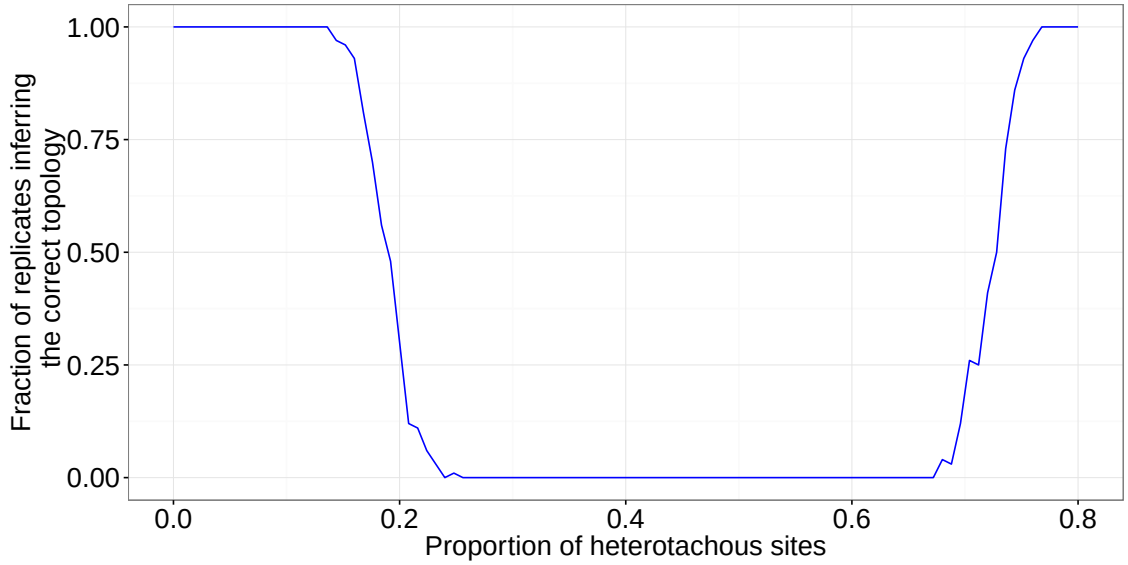

Figure 3.4: NJ results for the 4-taxa simulation study on heterotchosly-evolved sequence data. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value 100 replicate MSAs were simulated. The y-axis reports the fraction of replicates for which NJ inferred the correct tree topology.

0.15, before the success rate drops sharply and NJ consistently infers the incorrect tree for $p_{het} > 0.25$. However, at high values of p_{het} the performance of NJ diverges from that of MP. When $p_{het} > 0.67$ the NJ method starts to recover and infer the correct tree again. The recovery continues and for $p_{het} > 0.76$ NJ infers the correct tree for all MSAs. To find out why this occurs we need to look in more detail at the NJ method.

The NJ method is a distance-based clustering algorithm. It is fairly simple to implement, particularly when dealing with only four taxa. The first step is to select a distance metric, defining D_{ij} as the distance between taxa i and j. The JC distance (Jukes and Cantor, 1969) is the obvious choice in our case: where P_{ij} refers to the proportion of differing nucleotides between taxa i and j.

$$D_{ij} = -\frac{3}{4} ln(1 - \frac{4}{3} P_{ij}), \tag{3.8}$$

where P_{ij} refers to the proportion of differing nucleotides between taxa i and j. After calculating pairwise distances between all taxa we then calculate the pairwise

Q scores for all taxa:

$$Q_{ij} = (r-2)D_{ij} - \sum_{k=1}^{r} D_{ik} - \sum_{k=1}^{r} D_{jk}. \tag{3.9}$$

The pair of taxa with the lowest Q score are joined together, creating a new node. Distances and Q scores are then recalculated from each taxa to the new node and and the process is repeated until the tree is constructed. With four taxa we need only one iteration, and with only 3 possible binary trees (AB|CD, AD|BC and AC|BD, hereafter referred to as the AB tree, AD tree and AC tree respectively) we only need to find $min(Q_{AB}, Q_{AD}, Q_{AC})$ for any given dataset. Applying Equation (3.9) and simplifying yields:

$$\begin{aligned} Q_{AB} &= -(D_{AC} + D_{AD} + D_{BC} + D_{BD}), \\ Q_{AD} &= -(D_{AB} + D_{AC} + D_{BD} + D_{CD}), \\ Q_{AC} &= -(D_{AB} + D_{AD} + D_{BC} + D_{CD}). \end{aligned}$$

These equations are simple reformulations of the 4-point condition (Zaretskii, 1965; Buneman, 1971), which must be satisfied in order for a distance metric to be representable on a tree. The Q scores are all functions of D_{ij}, which in turn are all functions of P_{ij}. It is possible to derive the $E[P_{ij}]$, making use of Equation (3.4) and applying the law of total probability. Define P_{ij}^I, P_{ij}^V and P_{ij}^H to be the respective probabilities of an invariable, variable or heterotachous site changing state between taxa i and j. For any given value of p_{het}, $E[P_{ij}]$ can then be determined as follows:

$$E[P_{ij}] = p_{var}E[P_{ij}^V] + p_{inv}E[P_{ij}^I] + p_{het}E[P_{ij}^H]. \tag{3.10}$$

We can simplify Equation (3.10) by noting that $P_{ij}^I = 0$, leaving:

$$E[P_{ij}] = p_{var}E[P_{ij}^V] + p_{het}E[P_{ij}^H]. \tag{3.11}$$

$E[P_{ij}^V]$ and $E[P_{ij}^H]$ are calculated using Equation (3.4) using the appropriate branch length depending on whether the site is variable or heterotachous. Therefore we can easily calculate $E[P_{ij}]$ for all values of p_{het}. Since Q_{ij} is a function of D_{ij},

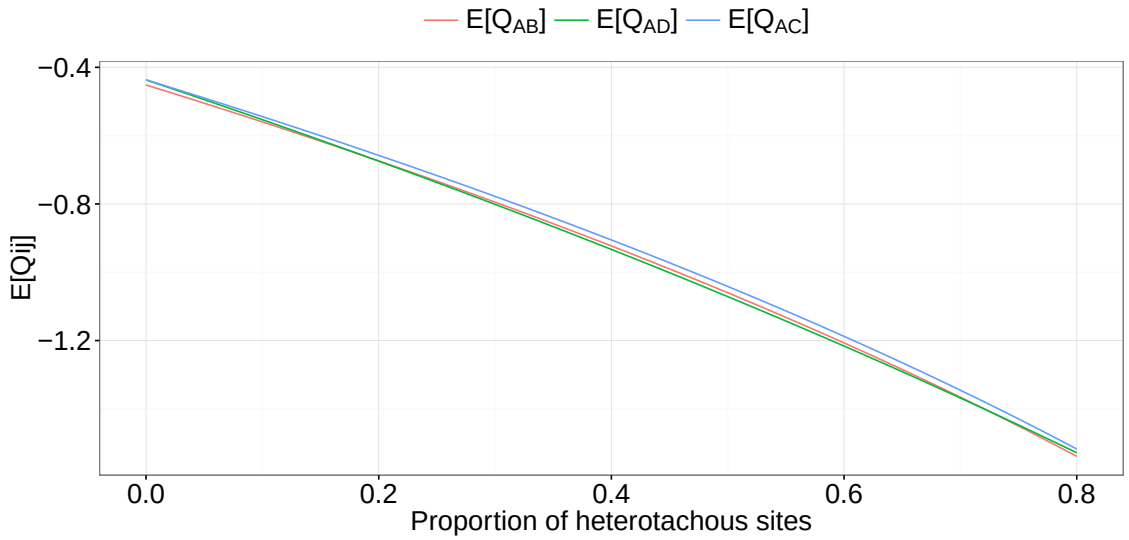

Figure 3.5: $E[Q_{ij}]$ scores, as a function of p_{het}. The x-axis displays proportion of heterotachous sites in the alignment. The y-axis displays the $E[Q_{ij}]$ scores for each topology. The minimum Q_{ij} dictates which pair of taxa will be clustered together, which fully resolves the topology in the four taxa case. It is difficult to see in detail which topology is the minimum for some values of p_{het}. To further investigate the plot is reproduced in Figure 3.6 with the trend removed.

which is in turn a function of P_{ij}, $E[Q_{ij}]$ can be calculated and plotted as a function of p_{het}, as shown in Figure 3.5. The detail on which topology is preferred is difficult to make out in Figure 3.5 without removing the trend. This is a simple matter of subtracting the mean $E[Q_{ij}]$. The detrended plot of the $E[Q_{ij}]$ scores can be found in Figure 3.6. Recalling that the minimum of our three Q scores indicates which tree will be inferred, we can see that the AC|BD tree should never be inferred, as $E[Q_{AC}] > E[Q_{AB}]$ for the entire range of p_{het}. For $p_{het} < 0.19$ we should expect to infer the correct tree (AB|CD); for $0.19 < p_{het} < 0.72$ we should expect to infer the incorrect tree(AD|BC); and for $p_{het} > 0.72$ we should expect to infer the correct tree again. Figure 3.7 displays these values superimposed on the results of the simulation study. Both the steep failure and recovery are centred around the two theoretical transition points, with the uncertainty immediately either side of the these points attributable to stochastic variation in the simulation process.

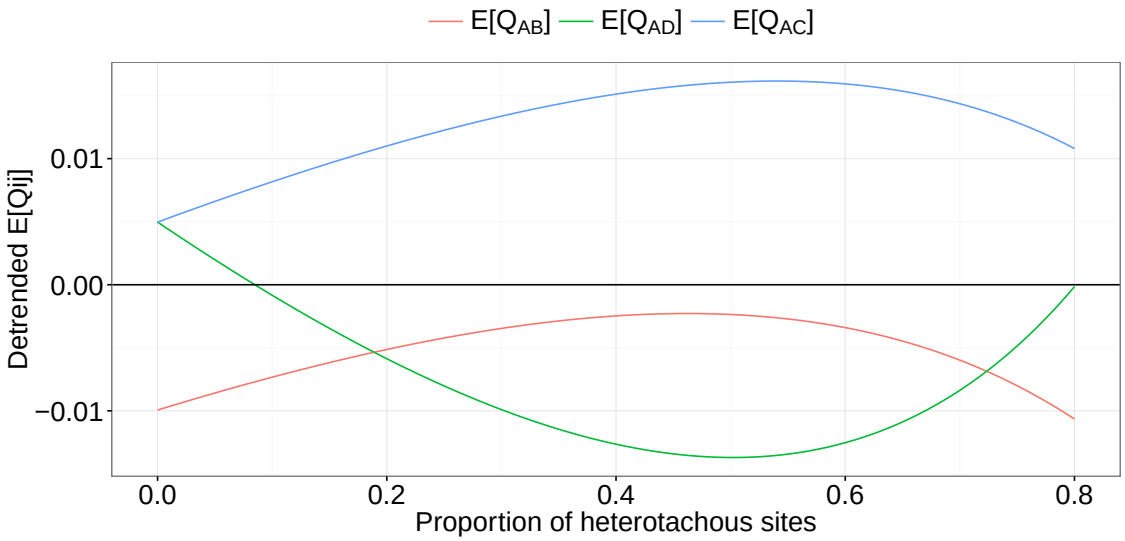

Figure 3.6: Detrended $E[Q_{ij}]$ scores, as a function of p_{het}. The x-axis displays proportion of heterotachous sites in the alignment. The y-axis displays the detrended $E[Q_{ij}]$ scores, that is $E[Q_{ij}] - \frac{1}{3}(E[Q_{AB}] + E[Q_{AD}] + E[Q_{AC}])$. The minimum Q_{ij} dictates which pair of taxa will be clustered together, which fully resolves the topology in the four taxa case. We can see that, consistent with the empirical results seen in Figure 3.4, we expect to infer the AB|CD tree for low and high values of p_{het}, and the AD|BC tree in between.

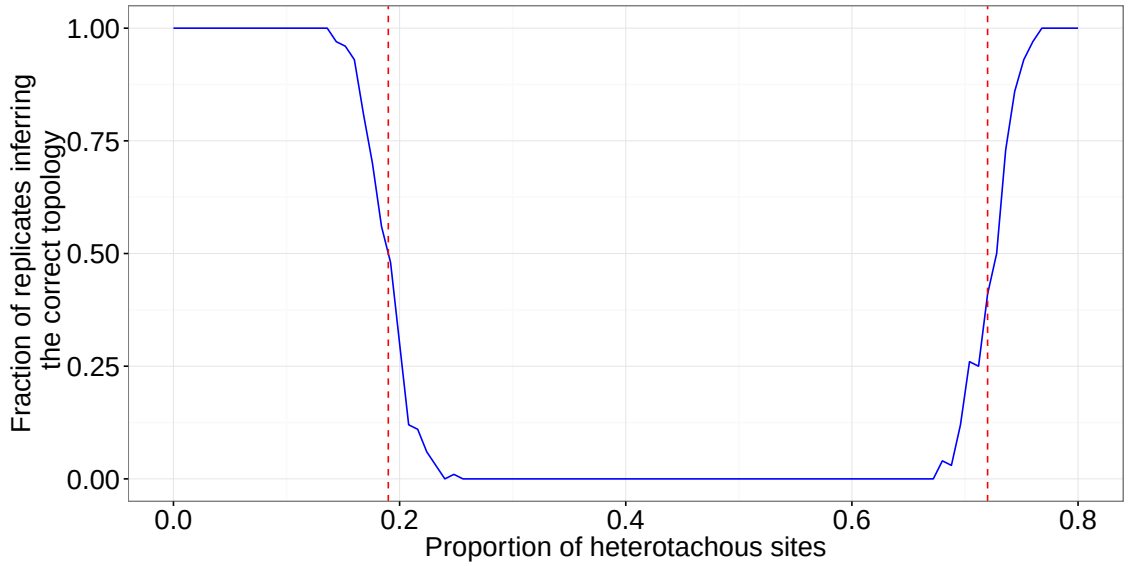

Figure 3.7: NJ results for the four taxa simulation study on heterotachously-evolved sequence data. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value 100 MSAs were simulated. The y-axis reports the fraction of MSAs for which NJ inferred the correct tree topology. The dashed red lines indicate the theoretical transition points at which the NJ method should switch from inferring the correct tree to the incorrect tree, and then from the incorrect tree back to the correct tree. Clearly the calculated transition points concur with the empirical data.

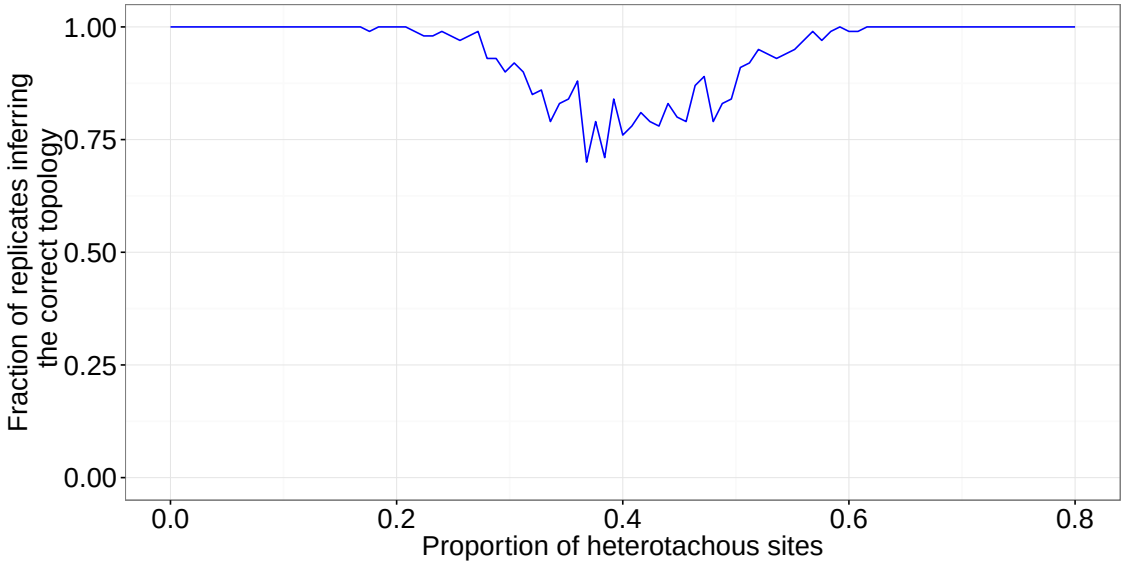

Figure 3.8: ML results for the four taxa simulation study on heterotachously-evolved sequence data. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value 100 MSAs were simulated. The y-axis reports the fraction of MSAs for which ML inferred the correct tree topology.

3.4 Maximum likelihood

The results for the ML method can be seen in Figure 3.8. There are three distinguishing features that are immediately apparent in the ML results. First, it is more successful than MP and NJ, inferring the correct tree more often and never dropping below a 70% success rate over the entire range of p_{het}. Second, similar to NJ it exhibits the tendency to recover successful inference at higher values of p_{het}. Finally, there is a significant amount of noise present in the results, a feature not observed in the MP or NJ methods.

The approach taken to reconciling these features was to analyse likelihood scores of the data, conditional on each of the three possible trees. For $j \in \{A, B, C\}$ we define LLp_{Aj} to be the mean maximum log likelihood of the 100 MSAs for $p_{het} = p\%$, constrained to the Aj tree. For example $LL90_{AD}$ refers to the mean maximum log likelihood of the 100 MSAs for $p_{het} = 90\%$, conditional on the AD|BC topology being chosen. We can then compare the difference between the correct tree, LLp_{AB},

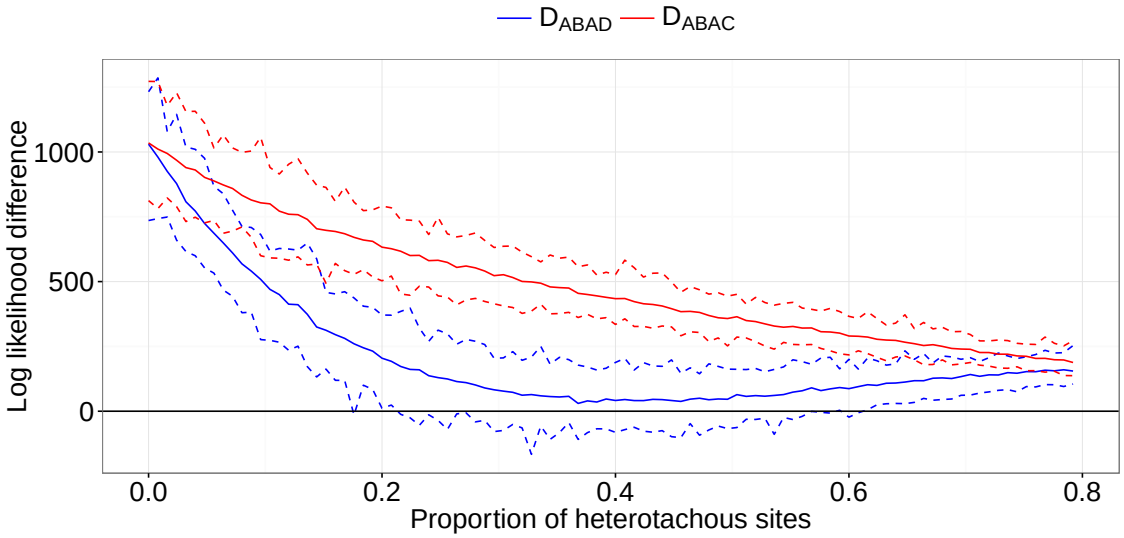

Figure 3.9: Comparison of conditional likelihoods of the simulated datasets. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value 100 replicate MSAs were simulated. The y-axis reports the difference in log likelihood. The solid lines show the mean difference in conditional likelihoods between the correct topology and the two incorrect topologies. The dashed lines indicate minimum and maximum difference over the 100 MSAs.

and the two incorrect trees, LLp_{AD} and LLp_{AC}, across the full range of p_{het}. We denote these differences by:

$$D_{ABAD} = LLp_{AB} - LLp_{AD}, \tag{3.12}$$

$$D_{ABAC} = LLp_{AB} - LLp_{AC}. \tag{3.13}$$

These results are shown in Figure 3.9. The upper and lower limits shown in Figure 3.9 are indicative of 95% prediction intervals for the two quantities, obtained from the simulated data.

The first thing we notice when looking at Figure 3.9 is that D_{ABAC} remains well above 0 for the entire range of p_{het}, particularly given the width of the prediction interval. From this we can reasonably conclude that the ML method should never choose the AC|BD tree. This was indeed the case for all of the simulated data. We

can now restrict our consideration to either the correct AB|CD tree or the incorrect AD|BC tree. From Figure 3.9 we can observe that D_{ABAD} also never falls below 0. This indicates that, in expectation at least, ML should be inferring the correct tree for all values of p_{het}. However, Figure 3.9 shows the minimum D_{ABAD} does fall below 0 for a significant range of p_{het}. This implies that the stochastic variation in the simulation process is sufficient to ensure that ML prefers the AD|BC tree in some of the MSAs .

This evidence suggests that the three distinguishing features of ML discussed at the start of this section can be explained by considering the ML results as a series of multinomial distributions. At each value of p_{het} there is some probability that a random simulation will yield AB|CD as the ML tree, some probability that the ML tree will be AD|BC, and some probability ($\simeq 0$) that the ML tree will be AC|BD. The 100 MSAs at each value of p_{het} can then be thought of as independent, random observations from the underlying multinomial distribution. This explains the noise seen in Figure 3.8.

Detailed analysis of results for the simulated MSAs assists in understanding Figure 3.8 but it does not inform us more generally. We seek to show that the results we observed for ML were predictable, in the same way that we have shown that the results of MP and NJ were predictable. The concept of expectation was critical to the derivation of the the transition points for MP and NJ. We relied on the expected site pattern frequencies in the MP calculations and the expected JC distance between taxa in the NJ calculations. We approach the ML problem in a similar way, although now we must construct expected datasets.

3.4.1 The Expected Dataset

We have shown in Section 3.2 that given a tree and model of evolution (in our case JC) it is possible to calculate the probability of observing a particular site pattern. Therefore we can calculate the probability distribution over all possible

site patterns and use this information to construct the expected MSA. Conceptually the task of constructing the expected dataset is simple. Computing the site pattern probability distribution is an onerous task, despite the economies offered by the generic site pattern notation of Section 3.2.2. Hendy and Penny (1989) outlined a linear algebra-based method for computing the probability distributions, employing Hadamard matrices. The attraction of this method to the current application was that it could be coded in a general sense, allowing the site pattern probability distributions of any 4 taxa tree to be computed in seconds.

For a set of DNA sequences that evolved under homogeneous conditions, on a given tree and model of evolution, we calculate the probability of observing a particular site pattern. Recall in the case of a 4-taxa tree there are 256 unique site patterns. We define $\mathbf{P}$ as the vector of site pattern probabilities, one for each of the 256 site patterns. We can then make use of $\mathbf{P}$ to construct the expected dataset for some desired sequence length, N. All that is required is to multiply $\mathbf{P}$ by N and round appropriately so that the sum of the resultant vector is N. We define this vector, $\mathbf{S}$, as the vector of site pattern counts in the expected dataset. It is then a simple task to write this dataset to a text file in the required format.

The constraint of homogeneous evolution is too restrictive for our application, but the method is easily adapted to create heterotachously-evolved expected datasets. We simply construct site pattern probability distributions for each of the variable, invariable and heterotachous classes. These will be referred to as $\mathbf{P}^{[\mathbf{V}]}$, $\mathbf{P}^{[\mathbf{I}]}$ and $\mathbf{P}^{[\mathbf{H}]}$ respectively. Calculating the site pattern frequencies for the expected dataset is then simply a matter of taking the weighted average of the three constituent probability distributions:

$$\mathbf{S} = N(p_{var}\mathbf{P}^{[\mathbf{V}]} + p_{inv}\mathbf{P}^{[\mathbf{I}]} + p_{het}\mathbf{P}^{[\mathbf{H}]}).$$

A separate expected dataset was constructed for all values of p_{het} considered in the simulations. For each of these expected datasets we used Phylip v3.69 (Felsenstein, 2002) to find the maximum likelihood, conditional on each of the three possi-

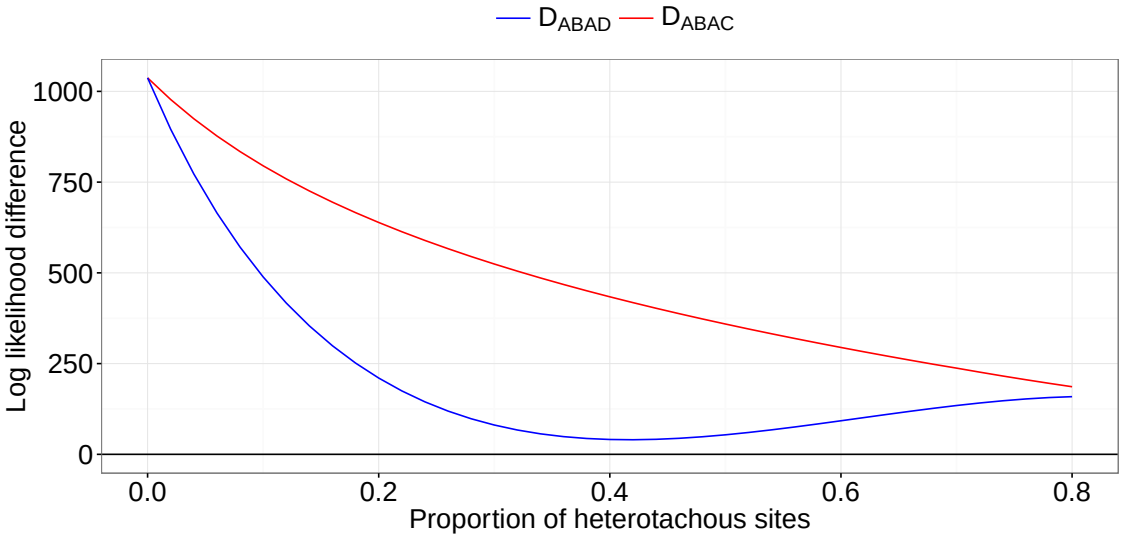

Figure 3.10: Comparison of conditional likelihoods of the expected datasets. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was constructed. The y-axis shows the difference in conditional likelihoods between the correct topology and the two incorrect topologies. Both curves correspond closely with the empirical evidence shown Figure 3.9.

ble tree topologies. In this way we established values for $E[LLp_{AB}]$, $E[LLp_{AC}]$ and $E[LLp_{AD}]$ which enables $E[D_{ABAD}]$ and $E[D_{ABAC}]$ to be calculated. Figure 3.10 displays these expected differences for the range of p_{het}. The only noticeable difference between the expected differences and the mean differences from the simulated MSAs (Figure 3.9) is the removal of the noise that was present in the simulation results.

3.5 Model misspecification

Our simulations show that heterotachously-evolved data presents a problem for all three inference methods. But each is affected in subtly different ways. We have shown that the results of each method were theoretically predictable, but that does not mean they are intuitive. The MP results (Figure 3.2) were not surprising, but

the recovery of the NJ and ML methods (Figures 3.4 and 3.8) initially seems counter intuitive. Why do these methods recover at high values of p_{het} and why does MP not recover also? To understand this we must think carefully about the types of model misspecification we have introduced. Both NJ and ML assume a JC model: NJ through the use of the JC distance metric; and ML directly. This means we are assuming the JC substitution model accurately reflects the process governing the evolution at each site in the alignment. This is true for the variable sites (20% of the alignment) but not for the invariable or heterotachous sites (80% of the alignment combined). Thus we have introduced two separate forms of model misspecification: that due to invariable sites, MM_{inv}; and that due to heterotachous sites, MM_{het}. For $p_{het} = 0$ ($p_{inv} = 0.8$) MM_{het} is not present but MM_{inv} is maximized. As p_{het} increases there is a trade-off between the two misspecifications: MM_{het} increases while MM_{inv} decreases. When $p_{het} = 0.8$ ($p_{inv} = 0$) MM_{inv} is not present but MM_{het} is maximized. The NJ and ML results indicate that both methods are able to accommodate one of these types of misspecifcation and still return the correct tree. However when the two misspecifications are simultaneously present in the data in sufficient quantities then performance is negatively affected. If one considers the simplicity of the form of heterotachy that was simulated compared to that resulting from countless changes in the evolutionary process operating on biological sequence data then it is clear that more realistic models of sequence evolution are urgently needed.

Chapter 4

Modeling heterotachous evolution

The simulation study of the previous chapter reinforces what had already been well established in the literature: that heterotachy is a critical source of model misspecification that negatively impacts phylogenetic inference using current methods. The best way to remove this model misspecifciation is to assume a model of sequence evolution that precisely matches the evolutionary processes that gave rise to the data. This is not possible with biological data, as the historical evolutionary process can not be known, but it is for simulated multiple sequence alignments (MSAs). In this chapter, we explore whether the tree topology, branch lengths and weights of heterotachously-evolved MSAs are recoverable using maximum likelihood (ML) based phylogenetic inference - provided the model of sequence evolution assumed is also heterotachous.

4.1 Inference with heterotachously-evolved data

4.1.1 Sequence alignments

We used the MSAs from the simulation study of Chapter 3, since we had already shown that these MSAs can result in topological errors when using common models and methods. In order to remove the effects of stochastic noise in the simulation pro-

cess we used the expected MSAs that were constructed in Section 3.4.1. Recall that the relative site pattern frequencies of these MSAs accurately reflect the expected site pattern frequency for a hypothetical sequence of infinite length. Specifically, we constructed expected MSAs with the following properties:

- Each MSA consists of a combination of variable, heterotachous and invariable sites.
- The variable sites evolved according to the Jukes Cantor (JC) (Jukes and Cantor, 1969) model of evolution on the four taxa tree shown in Figure 3.1 with branch lengths $a = d = 0.4$, $b_1 = c_1 = 0.1$, $b_2 = c_2 = 0.3$ and $k = 0.1$.
- The heterotachous sites evolved according to the JC model of evolution on the four taxa tree shown in Figure 3.1 with branch lengths $a = d = b_1 = c_1 = k = 0$ and $b_2 = c_2 = 0.3$.
- The invariable sites were held constant across the four taxa.
- The proportion of variable sites, p_{var}, was fixed at 0.2 for all MSAs.
- The proportion of heterotachous sites, p_{het}, varied from 0 to 0.8 at increments of 0.008.
- The proportion of invariable sites, p_{inv}, was determined by $p_{inv} = 1 - p_{var} - p_{het}$.
- The Hadamard conjugation method proposed by Hendy and Penny (1989) was used to calculate expected site pattern frequencies for the variable and heterotachous classes of each MSA.

4.1.2 The proposed model: JC+I+H2

In order for the model misspecification to be completely removed, the assumed model of sequence evolution had to precisely match the model used to generate the MSAs. As we have three distinct classes of sites (variable, heterotachous and invariable)

the only way to accurately reflect this is by using a three-class model. We use the notation JC+I+H2 to reflect this: JC indicates that sites evolve according to the JC model of evolution; +I indicates that we include a class of sites that is invariable; and we introduce the notation +H2 to indicate that there are two classes of sites that can be considered variable: each free to vary on the same tree topology but with different branch lengths. The JC+I+H2 model has 13 parameters to be estimated:

- The tree topology, T.
- The five branch lengths of the first variable class, $\boldsymbol{\lambda_1} = (\lambda_{1a}, \lambda_{1b}, \lambda_{1c}, \lambda_{1d}, \lambda_{1k})$, where the subscripts a, b, c and d refer to the four terminal edges and k refers to the inner edge.
- The five branch lengths of the second variable class, $\boldsymbol{\lambda_2} = (\lambda_{2a}, \lambda_{2b}, \lambda_{2c}, \lambda_{2d}, \lambda_{2k})$.
- The weight of the two variable classes, w_1 and w_2.

Since the weights must sum to one, the weight of the invariable class, w_I, is given by $w_I = 1 - (w_1 + w_2)$.

We define D_i as the site pattern observed at the i^{th} site in the alignment. We define L_i as the likelihood of observing the data given a tree T and an estimate of $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$. Recall from Section 3.2.1 that I is the event that a particular site in the alignment is invariable. Therefore, L_i is given by:

$$L_i = P(D_i|T, \boldsymbol{\lambda_1})w_1 + P(D_i|T, \boldsymbol{\lambda_2})w_2 + P(D_i|I)w_I.$$

For a sequence of length N, the log-likelihood, ℓ, of observing the MSA given T, $\boldsymbol{\lambda_1}$, $\boldsymbol{\lambda_2}$ and class weights is given by the likelihood function:

$$\ell = \sum_{i=1}^{N} log[P(D_i|T, \boldsymbol{\lambda_1})w_1 + P(D_i|T, \boldsymbol{\lambda_2})w_2 + P(D_i|I)w_I]. \tag{4.1}$$

4.1.3 Implementation in R

Neither the JC+I+H2 model, nor any model that could be coerced to behave as such, was implemented in any freely available phylogenetic inference software package.

Software programs for performing ML phylogenetic inference are extremely complex and the time and resources required to write our own were not available. It was therefore decided to utilise the `optim` function in the statistical computing software program `R` to construct a rudimentary ML-based phylogenetic inference algorithm. We simplified the task considerably by considering only four taxa, and the three-class case of heterotachy that we had constructed. For example, the restriction to four taxa allowed an exhaustive search of all topologies at little computational expense. Therefore it was not necessary to implement any kind of tree search heuristic.

Parameter initialization

Initial branch length parameters were drawn randomly from a uniform distribution on $[0, 1]$. The w_1 and w_2 had to be initialized so that they were contained in $[0, 1]$. Furthermore, it was necessary to ensure that $w_I = 1 - (w_1 + w_2)$ was also contained in $[0, 1]$. This was achieved as follows:

1. Choose w_1 randomly from the uniform distribution on $[0, 1]$.
2. Choose x randomly from the uniform distribution on $[0, 1]$.
3. Set $w_2 = x(1 - w_1)$.
4. Set $w_I = 1 - (w_1 - w_2)$.

Maximizing the likelihood function

The likelihood function given in Equation (4.1) was passed to `optim` as the function to be maximized. The optimiztion method used was the L-BFGS-B (Byrd *et al.*, 1995): a bounded version of the BFGS method (Fletcher, 2013) which allows upper and lower bounds to be chosen for each parameter. This was necessary as we had obvious parameter constraints: branch lengths must be non-negative and class weights must lie in $[0, 1]$. For each of the three topologies the branch lengths and class weights that maximized the likelihood function were found.

4.1.4 Performance

For all 101 expected datasets the correct AB|CD topology was found to have the highest likelihood. Figure 4.1 shows the difference in maximum likelihood scores between the correct topology and the two incorrect topologies. Figure 4.1 is the JC+I+H2 equivalent to Figure 3.10. There are some obvious differences between the two plots. Figure 4.1 is not as smooth as Figure 3.10, despite using the same expected MSAs. This is likely to be the result of the difference in inference programs. Phylogenetic inference programs, such as Phylip (Felsenstein, 2002), contain sophisticated optimization algorithms designed specifically for phylogenetic inference. The JC+I+H2 model was implemented using standard optimization routines in `R` and so the accuracy of the inference can be expected to suffer. Furthermore the JC+I+H2 model contains 13 parameters (tree topology, ten branch lengths and two class weights) as opposed to the JC model which contains just six (tree topology and five branch lengths). The increase of dimension may have a flattening effect on the likelihood space making the JC+I+H2 model more likely to terminate early. We also notice that the difference in likelihoods between the correct topology and both incorrect topologies increases as p_{het} increases. This suggests that as the influence of heterotachy increases the JC+I+H2 model is more likely to infer the correct topology.

Figure 4.2 shows the inferred branch lengths for the variable class (categorised as class 1). The inferred branch lengths are not precisely accurate but they are reasonably close to the true branch lengths (indicated by dashed lines on Figure 4.2). All branch lengths appear to be slightly overestimated, particularly the terminal branches leading to taxa A and D. The magnitude of the overestimation appears to increase as p_{het} increases. Figure 4.3 shows the inferred branch lengths for the heterotachous class (categorised as class 2). The inferred branch lengths for the heterotachous class appear more accurate than those of the variable class, particularly terminal branches B and C. Figure 4.4 shows the inferred class weights. The weight

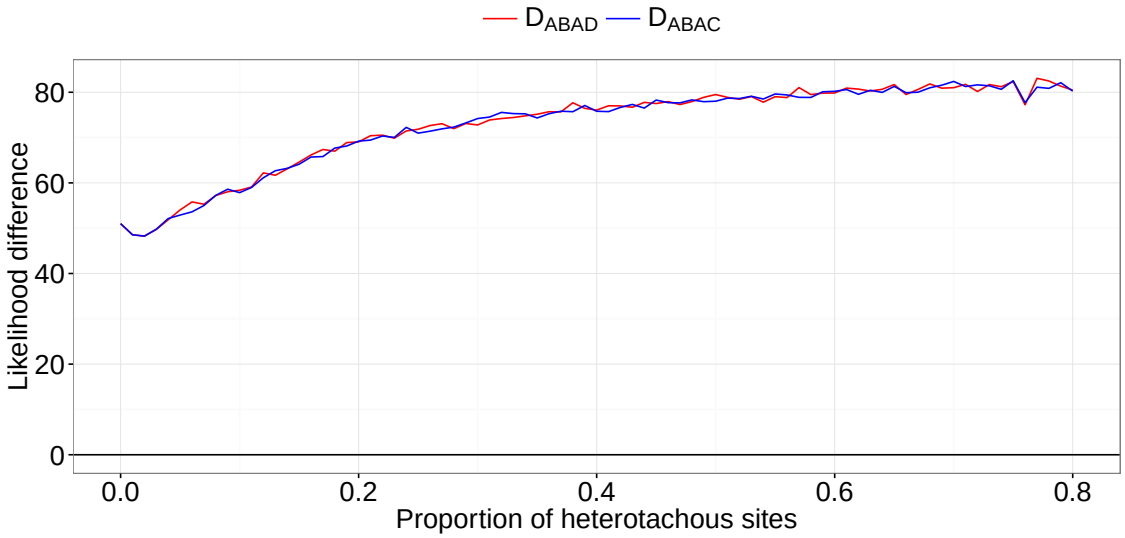

Figure 4.1: The difference in conditional maximum likelihood scores between the correct topology and the two incorrect topologies. D_{ABAD} refers to the difference between the maximum likelihood conditional on the AB|CD topology and the maximum likelihood conditional on the AD|BC topology. D_{ABAC} refers to the difference between the maximum likelihood conditional on the AB|CD topology and the maximum likelihood conditional on the AC|BD topology. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was generated as described in Section 3.4.1. The fact that the differences increase as p_{het} increases suggests that as the influence of heterotachy increases the JC+I+H2 model becomes more likely to infer the correct topology.

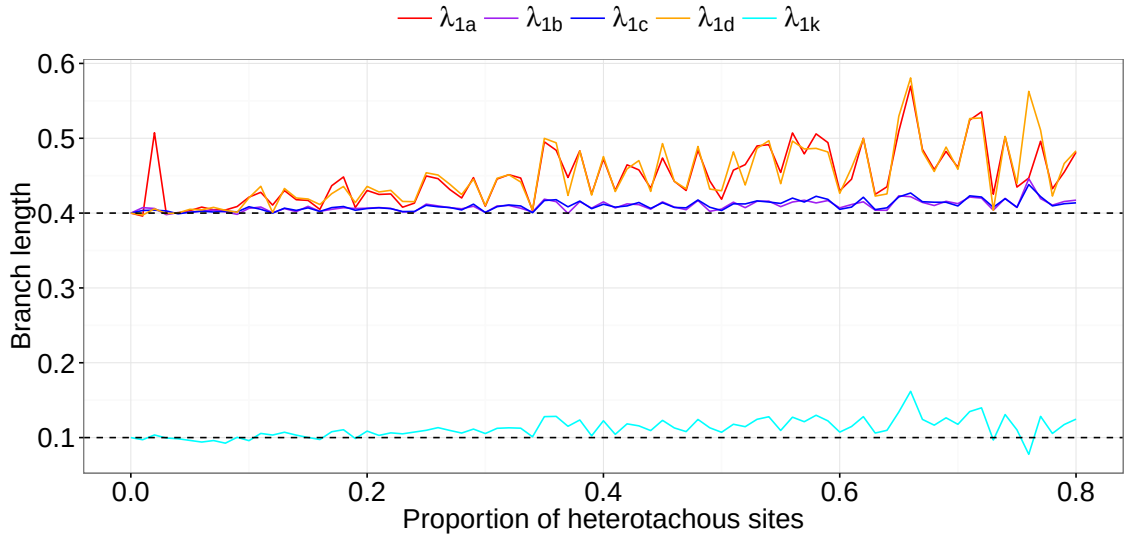

Figure 4.2: Branch lengths inferred by R under the JC+I+H2 model for the variable class of the expected MSAs. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was generated as described in Section 3.4.1. All branch lengths appear to be slightly overestimated, particularly the branches leading to taxa A and D. The magnitude of the overestimation appears to increase as p_{het} increases.

of the invariable class is recovered with a high degree of precision. However, the weight of the variable and invariable classes appear to be systematically underestimated and overestimated respectively. Once again, the magnitude of the inference error appears to increase as p_{het} increases. This is consistent with the overestimation of branch lengths in Class 1 and supports the hypothesis of a flat likelihood surface. It appears that, within a neighbourhood of the true values, a variety of parameter combinations can result in likelihoods very close to the maximum.

While not perfect, the performance of the JC+I+H2 model suggests that recovering model parameters from heterotachously-evolved data is achievable, provided one can minimize model misspecification. However the limitations of the `R` implementation of the model meant that further development of the approach would require a collaboration with the developers of an already established phylogenetic inference software program. Furthermore, the JC+I+H2 is too specific to be of much use with

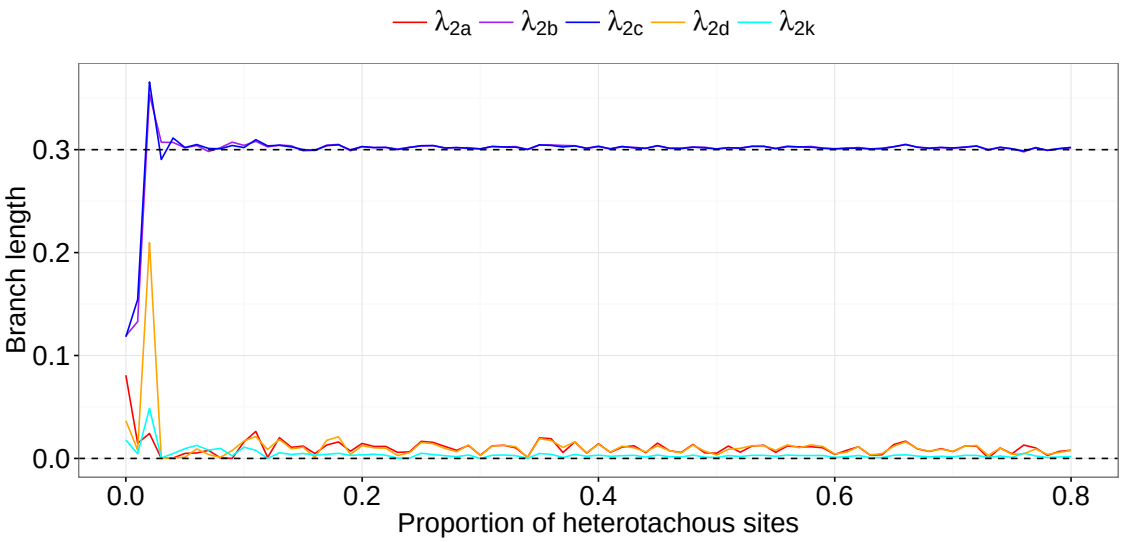

Figure 4.3: Branch lengths inferred by R under the JC+I+H2 model for the heterotachous class of the expected MSAs. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was generated as described in Section 3.4.1. All branch lengths appear to be recovered reasonably accurately, particularly the branches leading to taxa B and C.

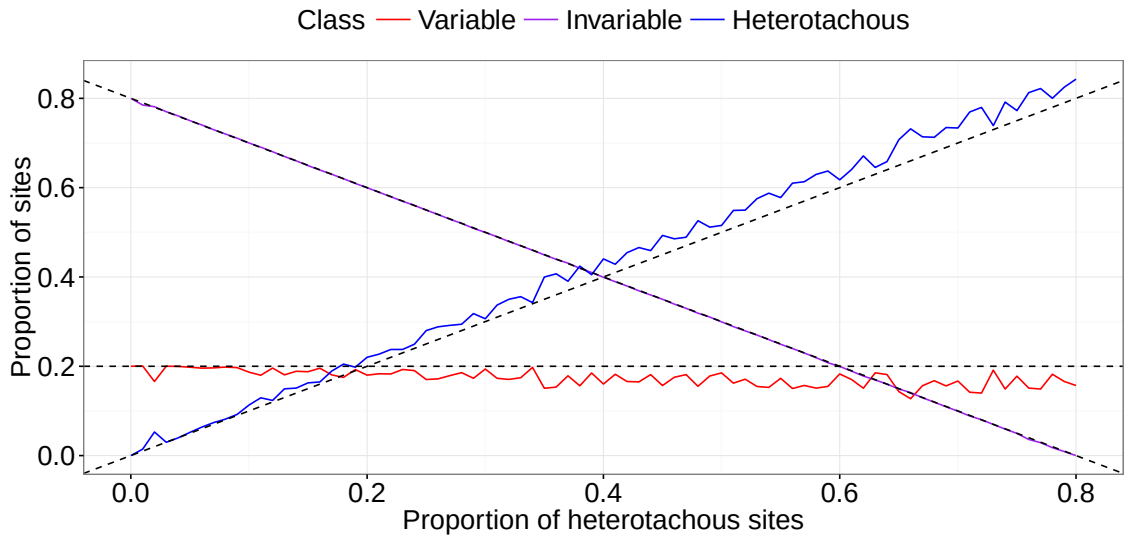

Figure 4.4: Class weights inferred by R under the JC+I+H2 model for the expected MSAs. The proportion of heterotachous sites, shown on the x-axis, was varied from 0 to 0.8 an increments of 0.008. At each value the expected MSA was generated as described in Section 3.4.1. The proportion of invariable sites is recovered with a high degree of accuracy. The proportion of variable sites appears to be slightly underestimated while the proportion of heterotachous sites appears to be slightly overestimated. The magnitude of these errors appears to increase as p_{het} increases.

biological data. There is no reason biologically that we should limit the model to three classes, or fix one of these classes to be invariable. The best way to minimize model misspecification is with a model that imposes as few constraints as possible, so that phylogenetic inference is predominantly driven by the data rather than the selected model.

4.2 The GHOST model

It is apparent that in order to adequately account for heterotachy when performing phylogenetic inference we require a model with a great deal more flexibility than those currently proposed in the literature. To this end we introduce the General Heterogeneous evolution On a Single Topology (GHOST) model. The GHOST model consists of m classes and one tree topology, T, common to all classes. Apart from the tree topology, all parameters are inferred separately for each class. For the j^{th} class we define λ_j as the set of branch lengths on T, Q_j, the substitution rate matrix, F_j, the set of nucleotide or amino acid frequencies and w_j, the class weight ($w_j > 0, \sum w_j = 1$). Given a multiple sequence alignment (MSA), S, we define L_{ij} as the likelihood of the data observed at the i^{th} site in S under the j^{th} class of the GHOST model. L_{ij} is computed using Felsenstein's pruning algorithm (Felsenstein, 1981). The likelihood of the i^{th} site, L_i, is then given by the weighted sum of the L_{ij} over all j:

$$L_i = \sum_{j=1}^{m} w_j L_{ij}(T, \lambda_j, Q_j, F_j).$$

So if S contains N sites (length of the alignment), the full log-likelihood, ℓ, is given by:

$$\ell = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{m} w_j L_{ij}(T, \lambda_j, Q_j, F_j) \right).$$

4.3 IQ-TREE Development

IQ-TREE is a phylogenetic inference software package developed in 2010 at the Centre for Integrative Bioinformatics Vienna (CIBIV). The IQ-TREE software was created as the successor of two earlier programs: IQPNNI (Vinh and von Haeseler, 2004) and TREE-PUZZLE (Schmidt *et al.*, 2002) (thus the name IQ-TREE). IQ-TREE was motivated by the rapid accumulation of phylogenetic data, leading to a need for efficient phylogenetic software that can handle a large amount of data and provide more complex models of sequence evolution. Among the stated aims of the IQ-TREE developers is to provide open source software that:

1. Provides novel computational methods that perform better than existing approaches.

2. Facilitates the inclusion of new maximum likelihood models that adequately capture more realistic aspects of sequence evolution.

The implementation of the GHOST model in IQ-TREE is clearly in keeping with these aims. Thus we instigated a collaboration in May 2015 with Professor Arndt von Haeseler and Dr Minh Bui of CIBIV. This collaboration culminated in the August 2016 release of IQ-TREE Heterotachy: a version of the software incorporating the GHOST model.

The tasks involved in implementing the GHOST model in IQ-TREE are attributed as follows:

1. Model specification - Stephen Crotty

2. Algorithm design - Bui Quang Minh and Stephen Crotty

3. Implementation in IQ-TREE - Bui Quang Minh

4. Validation - Stephen Crotty

4.4 Inferring phylogenies with IQ-TREE

Given a multiple sequence alignment (MSA) and a model of sequence evolution, IQ-TREE attempts to find the tree topology, branch lengths and substitution model parameters that maximize the likelihood of observing the MSA. Even for traditional, single class models this is not a trivial task. Branch length and model parameters can be optimized for a given tree topology but there are too many unique topologies to carry out an exhaustive search. IQ-TREE accomplishes the task by integrating the branch length and substitution model parameter optimization within an efficient tree search heuristic. The core algorithm is displayed in a flow chart in Figure 4.5.

4.4.1 Parameter optimization in IQ-TREE

Given an MSA, a tree topology T and a substitution rate matrix Q, one wants to optimise the length λ (Figure 4.6) of branch (a, b) connecting node a and node b, keeping all other branch lengths fixed. We define:

1. π_x to be the equilibrium frequency of (nucleotide or amino acid) character x.

2. $p_{xy}(\lambda)$ to be the probability of change from character x to y on branch of length λ. $p_{xy}(\lambda)$ are the entries of the transition probability matrix $P(\lambda) = e^{Q\lambda}$, as per the current parameter estimates.

3. $L_a^h(x)$ to be the partial likelihood of the sub-tree rooted at node a having the character x at node a for alignment site h.

4. $f_h(\lambda)$ to be the site likelihood of T at alignment site h, given branch length λ between a and b.

$L_a^h(x)$ is computed in a dynamic programming manner using Felsenstein's pruning algorithm. If a is an internal node, let j and k be the two child nodes of a.

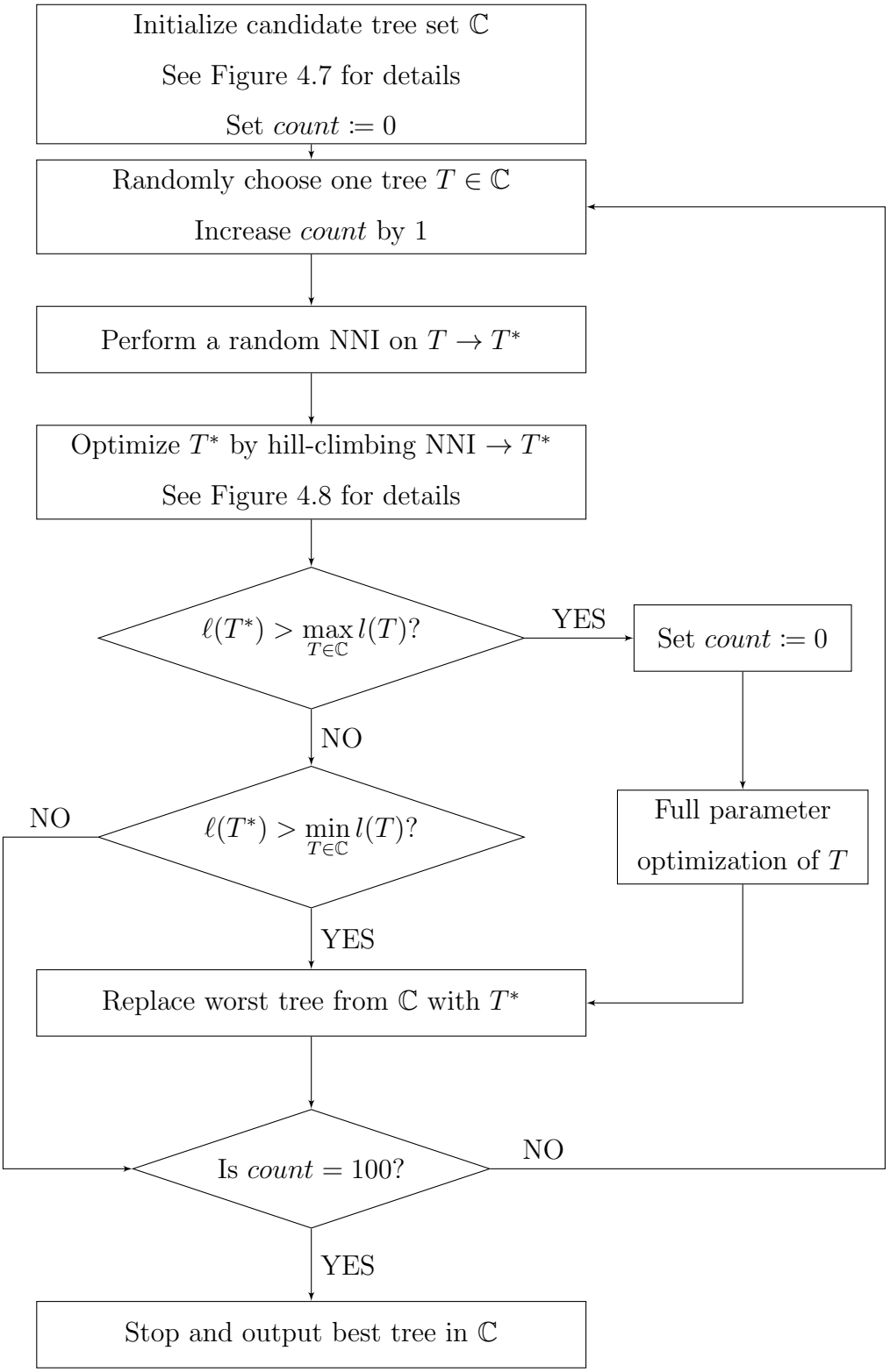

Figure 4.5: Flow chart for IQ-TREE's core optimization algorithm, largely reproduced from Figure 3 of Nguyen *et al.* (2015).

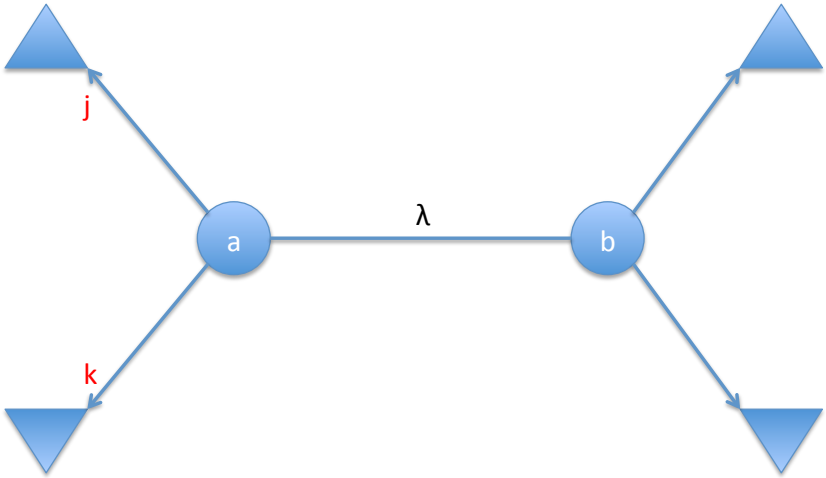

Figure 4.6: Schematic of a phylogenetic tree. The circles represent two nodes on the tree, a and b, connected by a branch of length λ. The triangles represent subtrees.

Then,

$$L_a^h(x) = \left(\sum_y p_{xy}(\lambda_j) L_j^h(y) \right) \left(\sum_y p_{xy}(\lambda_k) L_k^h(y) \right), \tag{4.2}$$

where λ_j and λ_k are the branch lengths of (a, j) and (a, k) respectively. If a is a leaf node, then $L_a^h(x)$ is assigned from the input alignment. $L_b^h(x)$ is computed analogously and we are then able to compute $f_h(\lambda)$:

$$f_h(\lambda) = \sum_x \pi_x \sum_y p_{xy}(\lambda) L_a^h(x) L_b^h(y). \tag{4.3}$$

Holding tree topology, model parameters and all other branch lengths fixed, the optimal branch length λ is the one that maximises the log-likelihood function:

$$\ell(\lambda) = \sum_h \log \big(f_h(\lambda) \big). \tag{4.4}$$

IQ-TREE utilises standard mathematical optimization procedures for optimizing the branch lengths and substitution model parameters. Branch lengths are optimized using the Newton-Raphson method whereas substitution model parameters are optimized using Brent's method (Brent, 2013).

4.4.2 Searching tree space

IQ-TREE has two distinct algorithms for searching tree space. The Candidate Tree Set Algorithm (CTSA) is responsible for generating a set of distinct, data-informed topologies at the start of the optimization procedure. These trees are constructed using stepwise addition, whereby taxa are added one by one by placing them on the most parsimonious branches. The process is repeated to generate up to 100 unique candidate trees. To avoid arriving at the same tree each time, IQ-TREE changes the order that taxa are added throughout the tree building process as well as undertaking subtree pruning and regrafting (SPR) (Stamatakis, 2006). The CTSA can be seen in Figure 4.7. The Hill-climbing Nearest Neighbour Interchange Algorithm (HNNIA) is implemented iteratively throughout the optimization procedure and aims to prevent IQ-TREE from finding solutions that are only locally optimal. In essence it searches for a nearest neighbour interchange (NNI) (Guindon and Gascuel, 2003), or combination of several NNIs, that improves the current best tree. The HNNIA can be seen in Figure 4.8.

4.5 Implementation of the GHOST model in IQ-TREE

4.5.1 Optimizing branch lengths and substitution model parameters of the GHOST model

The framework outlined above computes the likelihood of the data assuming a single tree and substitution model. Implementing the GHOST model requires this framework to be extended to incorporate multiple classes. The GHOST model consists of m classes, $S = \{s_1, s_2, \ldots, s_m\}$. Class z consists of a set of branch lengths, λ_z; an instantaneous rate matrix, Q_z; and a set of equilibrium base frequencies F_z. Each site in the alignment belongs to one of the m classes, but it is not known which one.

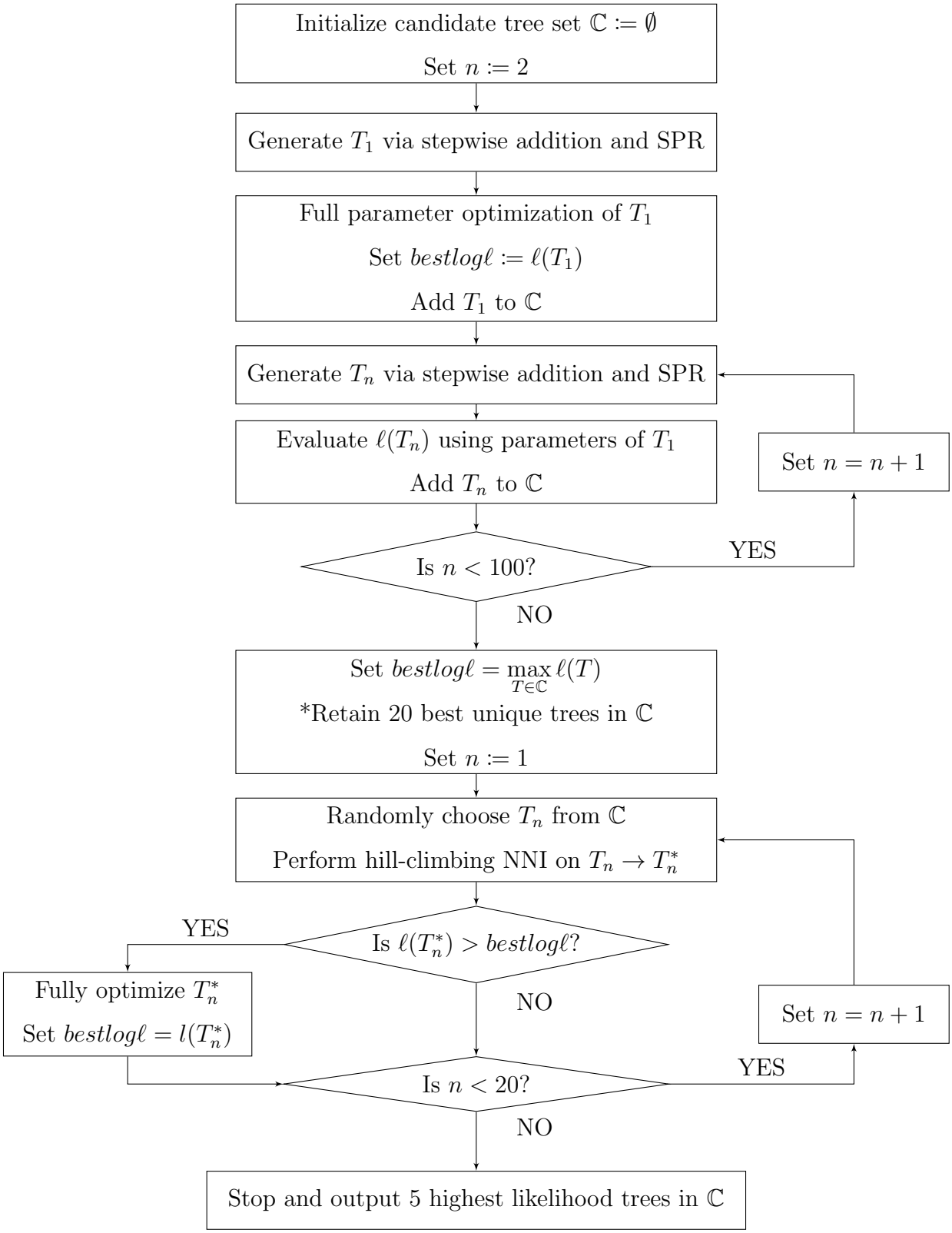

Figure 4.7: Flow chart detailing the Candidate Tree Set Algorithm (CTSA).

*If $\mathbb{C}$ contains less than 100 unique topologies at this point in the algorithm then $\mathbb{C}$ is populated with random unique topologies until it contains the lesser of 100 or all possible topologies.

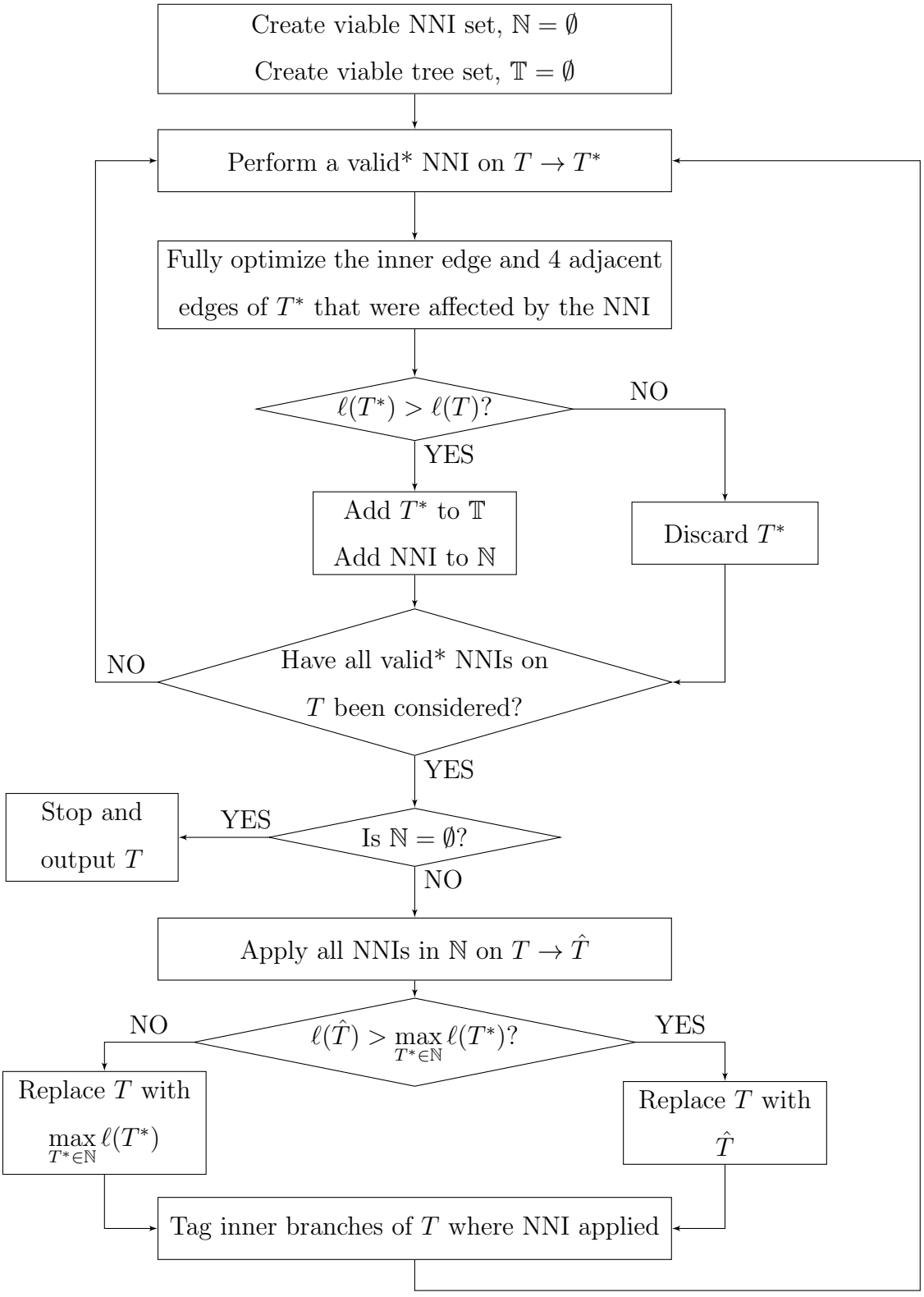

Figure 4.8: Flow chart for the Hill-climbing Nearest Neighbour Interchange Algorithm (HNNIA). *A valid nearest neighbour interchange (NNI) is any NNI upon the initial iteration, or any NNI on an inner edge within 2 branches of a tagged edge upon subsequent iterations.

To progress we must extend the definitions outlined in Section 4.4.1 to include a class index. We define:

1. $p_{xyz}(\lambda)$ to be the probability of change from character x to y during time λ for class z. $p_{xyz}(\lambda)$ are the entries of the transition probability matrix $P(\lambda) = e^{Q_z \lambda}$ where Q_z is the instantaneous rate matrix for class z.

2. $L^h_{az}(x)$ to be the partial likelihood of the sub-tree rooted at a having the character x at alignment site h for class z.

3. $f_{hz}(\lambda)$ to be the site likelihood of T at alignment site h, given branch length λ between a and b for class z.

4. w_z to be the weight of class z. We further define the vector $\mathbf{w} = w_1, w_2, \ldots, w_m$ to be the set of weights for a given mixture model, noting the obvious constraint that $\sum_i w_i = 1$.

Analagous to the construction of the likelihood function in Section 4.4.1, we have:

$$L^h_{az}(x) = \Big(\sum_y p_{xyz}(\lambda_{jz}) L^h_{jz}(y) \Big) \Big(\sum_y p_{xyz}(\lambda_{kz}) L^h_{kz}(y) \Big), \tag{4.5}$$

where λ_{jz} and λ_{kz} are the lengths of the branches (a, j) and (a, k) in class z. Further,

$$f_{hz}(\lambda_z) = \sum_x \sum_y \pi_{xz} p_{xyz}(\lambda_z) L^h_{az}(x) L^h_{bz}(y), \tag{4.6}$$

where λ_z is the length of branch (a, b) in class z; and π_{xz} is the equilibrium frequency of character x for class z. We then compute f_h by simply taking a weighted sum over the m classes:

$$f_h = \sum_z w_z f_{hz}(\lambda_z). \tag{4.7}$$

Finally, the log likelihood is then computed as per Equation (4.4):

$$\begin{aligned} \ell &= \sum_h \log \big(f_h \big) \\ &= \sum_h \log \big(\sum_z w_z f_{hz}(\lambda_z) \big). \end{aligned} \tag{4.8}$$

The existing techniques used for optimizing branch lengths and substitution model parameters can be exended to the GHOST model without alteration. Each parameter is optimized individually, holding all others fixed. Therefore the only difference between optimizing a single class model and the GHOST model is that there are more parameters and consequently more computation time is required.

4.5.2 Optimization of weights for the GHOST model

The weights of the GHOST model cannot be optimized within IQ-TREE's existing architecture, since these parameters do not exist in any simpler, single class models. To optimize the weights in conjunction with the other parameters we employ an expectation-maximization (EM) algorithm (Dempster *et al.*, 1977).

Let $\Theta = \{w_1, \ldots, w_m, \lambda_1, \ldots, \lambda_m, Q_1, \ldots, Q_m, F_1, \ldots, F_m\}$ denote the GHOST model parameters, *i.e.* class weights, branch lengths, rate matrices, and nucleotide or amino-acid frequencies for each of the m classes. We initialize Θ with $\hat{w}_1 = \ldots = \hat{w}_m = 1/m$, all rates $= 1$ in each rate matrix $\hat{Q}_j$, uniform nucleotide or amino-acid frequencies $\hat{F}_j$ (i.e., the Jukes-Cantor model (Jukes and Cantor, 1969)), and $\hat{\lambda}_j$ obtained from the branch lengths of the parisomny tree, rescaled by a discrete Γ distribution with m categories. This becomes the current estimate $\hat{\Theta}$. The EM algorithm iteratively performs an expectation (E) step and a maximization (M) step to update the current estimate until a (local) maximum likelihood is reached.

E-step

For each site h and class z, compute the posterior probability $\hat{p}_{hz}$ of site h belonging to class z based on the current estimate $\hat{\Theta}$:

$$\hat{p}_{hz} = \frac{\hat{w}_z L_{hz}(T, \hat{\Theta})}{\sum_{k=1}^{m} \hat{w}_k L_{hk}(T, \hat{\Theta})}.$$

M-step

For each class z, the log-likelihood function:

$$\ell_z = \sum_{h=1}^{N} \hat{p}_{hz} \log \Big(L_{hz}(T, \lambda_z, Q_z, F_z) \Big).$$

is maximized to obtain the new $\hat{\lambda}_z, \hat{Q}_z, \hat{F}_z$. This can be performed with standard phylogenetic optimization routines for each class, as discussed in Section 4.4.1.

Finally, the weights are updated by:

$$\hat{w}_z = \frac{1}{N} \sum_{i=1}^{N} \hat{p}_{iz}.$$

That is, the new weight for class z is the mean posterior probability of each site belonging to class z. We now have updated all parameters in the model to obtain a new estimate of Θ, which we denote $\hat{\Theta}^{NEW}$. If $\ell(\hat{\Theta}^{NEW}) > \ell(\hat{\Theta}) + \epsilon$ (where $\epsilon = 0.01$), then $\hat{\Theta}$ is replaced by $\hat{\Theta}^{NEW}$ and the E and M steps are repeated. Otherwise, the EM algorithm finishes.

The lack of theoretical obstacles to the implementation of the GHOST model in IQ-TREE is no guarantee that it will be successful in practice. The IQ-TREE implementation of the GHOST model was validated by conducting a set of rigorous simulation studies, which provides the focus of Chapter 5.

Chapter 5

Validation of the GHOST model in IQ-TREE

The implementation of the GHOST model in IQ-TREE was validated by completing a rigorous series of simulations studies. The simulation studies can be broadly categorised as being focused on either recovery of tree topology or recovery of branch lengths and substitution model parameters.

5.1 Tree topology recovery

We validated the ability of IQ-TREE using the ML-GHOST model to recover tree topology from heterotachously-evolved sequences by replicating the simulation study carried out by Kolaczkowski and Thornton (2004) (K&T). We used *Seq-Gen* v1.3.3 (Rambaut and Grassly, 1997) to simulate DNA sequences on two symmetric, 4-taxa trees of identical topology (see Figure 5.1) using the JC model of evolution (Jukes and Cantor, 1969). The branch lengths, parameterised as expected number of substitutions per site, were constructed such that each tree comprised of one pair of non-sister long branches (length p) and one pair of non-sister short branches (length q) separated by an internal branch (length r). Thus when considered individually each tree was susceptible to long branch attraction (Felsenstein, 1978) (LBA). Im-

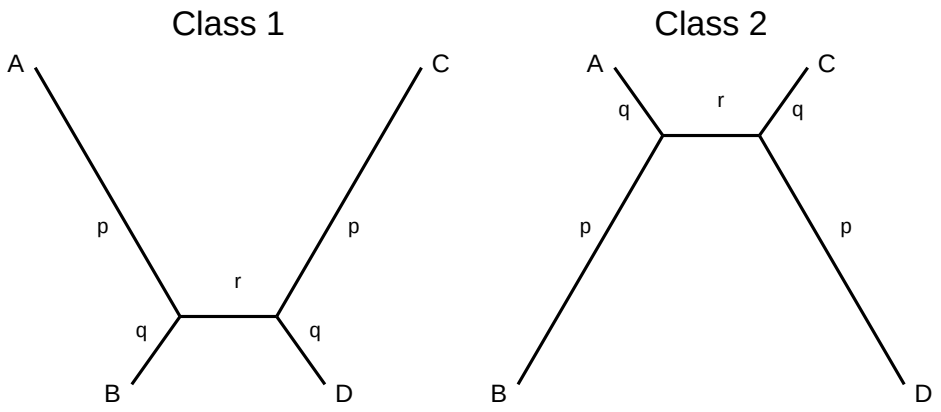

Figure 5.1: The two symmetric, 4-taxa trees of identical topology used in the simulation studies of K&T. The branch lengths were constructed such that each tree comprised of one pair of non-sister long branches and one pair of non-sister short branches.

portantly, the LBA artefact in both trees is complementary - each is biased in the direction of the AC|BD tree. We replicated three separate experiments that were all initially carried out by K&T, hereafter referred to as Experiment 1, Experiment 2 and Experiment 3.

5.1.1 Experiment 1

We fixed $p = 0.75$ and $q = 0.05$ and varied r on the interval $[0.01, 0.4]$ in increments of 0.01. For each value of r, 200 simulated multiple sequence alignments (MSAs) were constructed by concatenating two sub-alignments of equal length, one simulated on each of the trees in Figure 5.1. We carried out phylogenetic inference on each MSA using MP; ML under a JC model of substitution (ML-JC); and ML under the GHOST model with two classes: both assuming a JC model of evolution (ML-JC+H2). The experiment was repeated for sequence lengths of 1,000, 10,000 and 100,000 base pairs. The results can be found in Figure 5.2.

K&T found that both ML and MP had difficulty recovering the correct topology

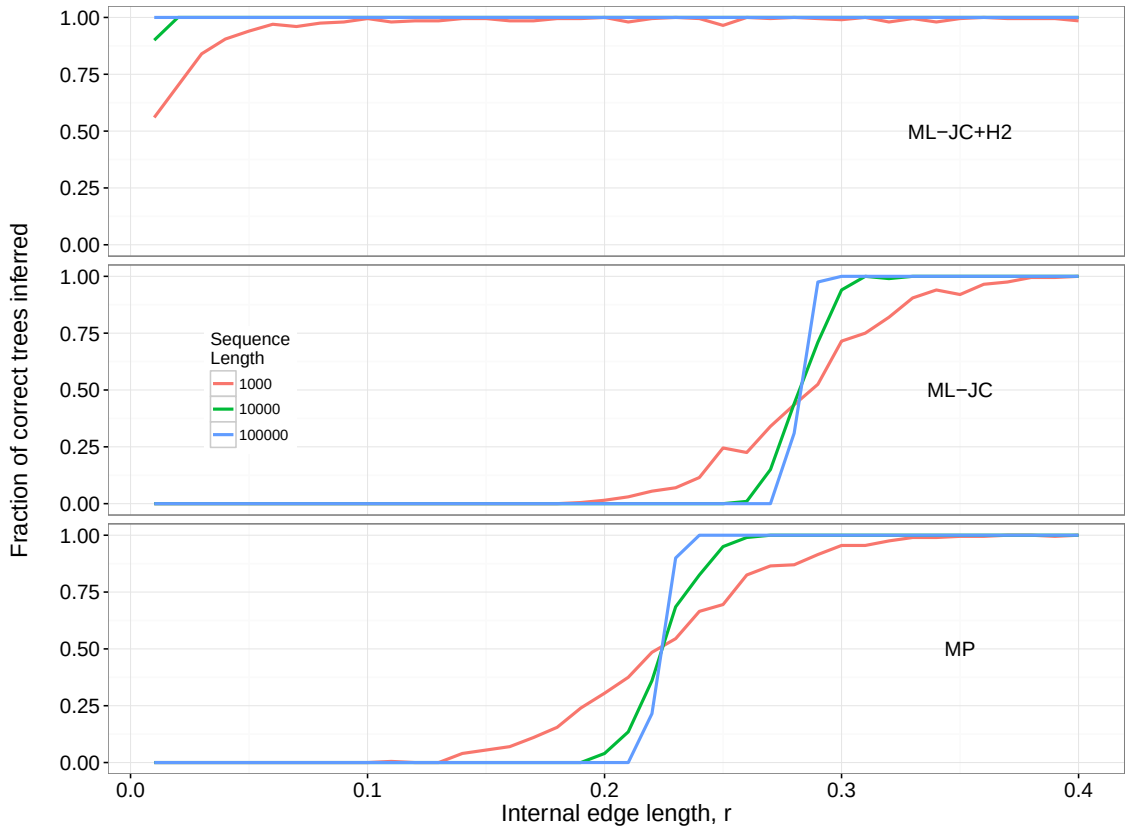

Figure 5.2: Performance of ML-JC+H2, ML-JC and MP for data generated under strong heterotachy, p=0.75 and q=0.05. The length of the internal branch, r, is displayed on the x-axis and was varied between 0.01 and 0.4 with 200 replicates at each value of r. The y-axis displays the fraction of the 200 replicates that recovered the correct topology. The results for MP and ML-JC were identical to the results of K&T, neither performed adequately but MP is able to recover the correct topology for shorter r than ML-JC. However ML-JC+H2 was able to reliably recover the tree topology for this data even when the internal branch is very short.

when the internal branch length, r, was short, but as r increased MP recovered before ML. They also found ML to be inconsistent: for small r, increasing sequence length caused ML to infer the incorrect topology more often. We followed K&T's method precisely and compared the performance of MP, ML-JC (ML under a JC model, identical to the ML used by K&T) and the ML-JC+H2 (ML under JC with 2 GHOST classes). Our results for MP and ML-JC mirrored those of K&T precisely. For a sequence length of 100kb, MP inferred an incorrect topology to some extent for $r < 0.24$ and ML-JC did likewise for $r < 0.3$. The ML-JC+H2 model however always inferred the correct topology. Figure 5.2 shows that given sufficient sequence length, the ML-JC+H2 model inferred the correct topology from the heterogeneous sequences 100% of the time with r as low as 0.01. Our results clearly demonstrate that the ML-JC+H2 model can correctly infer the tree topology when ML-JC and MP both are misled by heterotachy.

5.1.2 Experiment 2

We tested nine different scenarios corresponding to combinations of $p \in \{0.3, 0.5, 0.7\}$ and $q \in \{0, 0.2, 0.4\}$. At each combination of p and q we simulated an MSA of 10,000bp. For each of MP, ML-JC and ML-JC+H2; and at each combination of p and q we determined the smallest value of r (subject to the minimum $r = 0.001$) such that the correct topology was returned at least 50% of the time, over 200 replications. This measure was defined by K&T as BL_{50}. The results can be found in Figure 5.3.

Again the results we observed for MP and ML-JC closely emulated the findings of K&T. ML-JC+H2 comprehensively outperformed the two alternatives, with the difference most apparent when the influence of heterotachy was strongest. This occurs when $q = 0$, for all three values of p ML-JC+H2 achieved a BL50 score of 0.001, which was the smallest value of r tested. When p and q were of similar size the performance of ML-JC+H2 was similar to MP and ML-JC. This is unsurprising

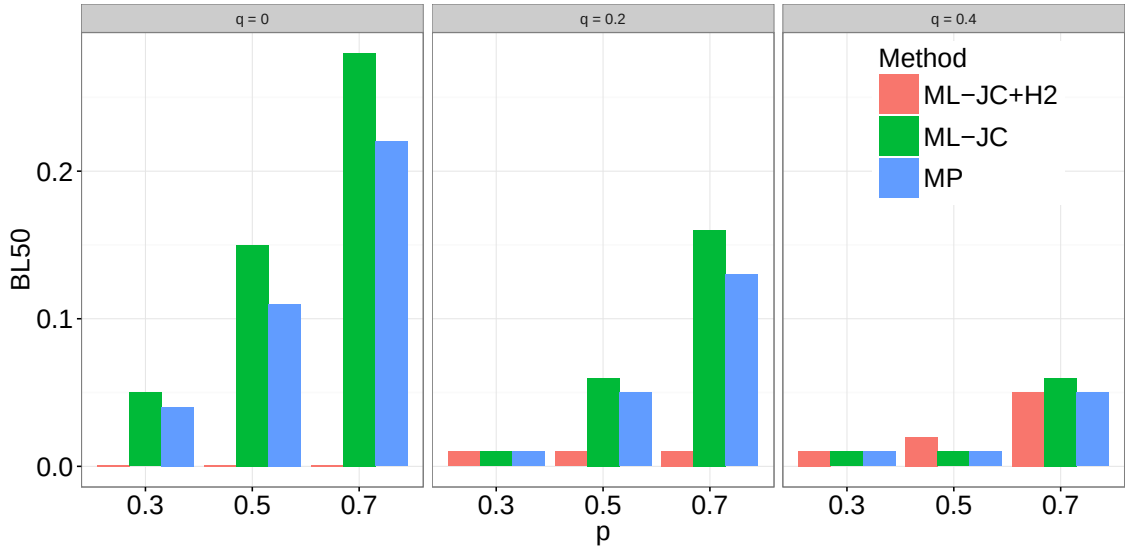

Figure 5.3: Results of K&T's Experiment 2, assessing the performance of MP, ML-JC and ML-JC+H2 for different combinations of p and q. On the x-axis are three different values of p and three different values of q are displayed in the separate facets. On the y-axis is BL_{50},defined by K&T as the minimum internal branch length required for the method to recover the correct tree topology at least 50% of the time, for a sequence length of 10,000bp. Small values of BL_{50} indicate that the model is less likely to infer the incorrect topology given the heterotachously-evolved data. The ML-JC+H2 model clearly outperforms MP and ML-JC over the range of heterotachous conditions tested by K&T. The only cases in which MP and ML perform comparably to ML-JC+H2 is when p and q are similar, that is when the data is not particularly heterotachous, (e.g. $p = 0.3$ & $q = 0.4$, or $p = 0.5$ & $q = 0.4$).

as similar p and q results in the two classes being similar to each other. Thus the influence of heterotachy is not strong and a single class model is most likely sufficient to describe the data.

5.1.3 Experiment 3

We tested the impact of varying the weight, w, of each class in the simulated MSAs for a variety of branch lengths combinations. Initially p and q were fixed at 0.75 and 0.05 respectively, with $r \in \{0.05, 0.15, 0.25\}$ and $w \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$. The process was then repeated, this time with p and r fixed

at 0.75 and 0.15 respectively, with $q \in \{0.05, 0.15, 0.25\}$ and w as before. Sequence length was held fixed throughout at 100,000 bp and 200 replicates were simulated at each combination of branch lengths and weight. The results of K&T indicate that ML-JC could not reliably recover the correct topology for all weights for any of the branch length combinations. Conversely we found that for all branch length combinations ML-JC+H2 was able to recover the correct topology for all weights and 100% of replicates.

5.2 Parameter recovery

The replication of the K&T simulations focused on recovering the tree topology only. The GHOST model is parameter rich and naturally the validation process must address its ability to accurately recover tree and model parameters. We validated the ability of IQ-TREE using ML-GHOST model to recover these parameters from heterotachously-evolved sequences by using two separate simulation procedures. The first procedure, hereafter referred to as the Specific Case, focused on how accurately a specific set of parameters could be recovered under the GHOST model. The second procedure, hereafter referred to as the General Case, focused on whether the results indicated from the Specific Case could be generalised to the broader parameter space. Both simulation procedures used the same tree topology, generated randomly on 12 taxa. *SeqGen* (Rambaut and Grassly, 1997) was again used to simulate all MSAs. Each MSA was constructed as a concatenation of two classes with the j^{th} class having its own set of branch lengths, λ_j, its own GTR rate matrix, Q_j, and its own set of base frequencies F_j. The method for generating the parameters of each class was common to both procedures:

1. A 12 taxa tree topology was generated randomly under the Yule-Harding model (Yule, 1925; Harding, 1971).

2. The branch lengths were drawn randomly from an exponential distribution

with a mean of 0.1.

3. When specifying a GTR rate matrix in *SeqGen*, the G↔T transition rate is fixed at 1 and all other transition rates are expressed relative to the G↔T rate. For each class the 5 transition rates were drawn randomly from a uniform distribution between 0.5 and 5 (U(0.5,5)).

4. Preliminary simulation results indicated that datasets that contained very low base frequency parameters for 1 or more nucleotides returned unreliable parameter estimates. This phenomenon is not directly related to the GHOST model, it was found to be equally present using IQ-TREE when fitting a simple GTR model to an unmixed dataset. To remove any effect of this artefact from our results the base frequencies were each assigned a minimum of 0.1. The remainder was allocated proportionally amongst the four nucleotides by normalising a set of four observations from a $U(0,1)$ distribution and scaling by a factor of 0.6.

5.2.1 Specific Case

Following the steps outlined in Section 5.2 a tree topology and independent sets of parameters for two classes were generated. From these two classes MSAs were constructed by varying the weight of each class. The weight of Class 1, w_1, was varied from 0.2 to 0.8 in increments of 0.05. At each value of w_1, 20 separate datasets were constructed with sequence length 10,000 bp. Each dataset was constructed by concatenating two independently simulated sets of sequences, the first of length $10,000 \times w_1$ simulated using Class 1 parameters, and the second of length $10,000 \times (1-w_1)$ using Class 2 parameters. We used IQ-TREE to infer parameters from each MSA according to ML-GTR+H2.

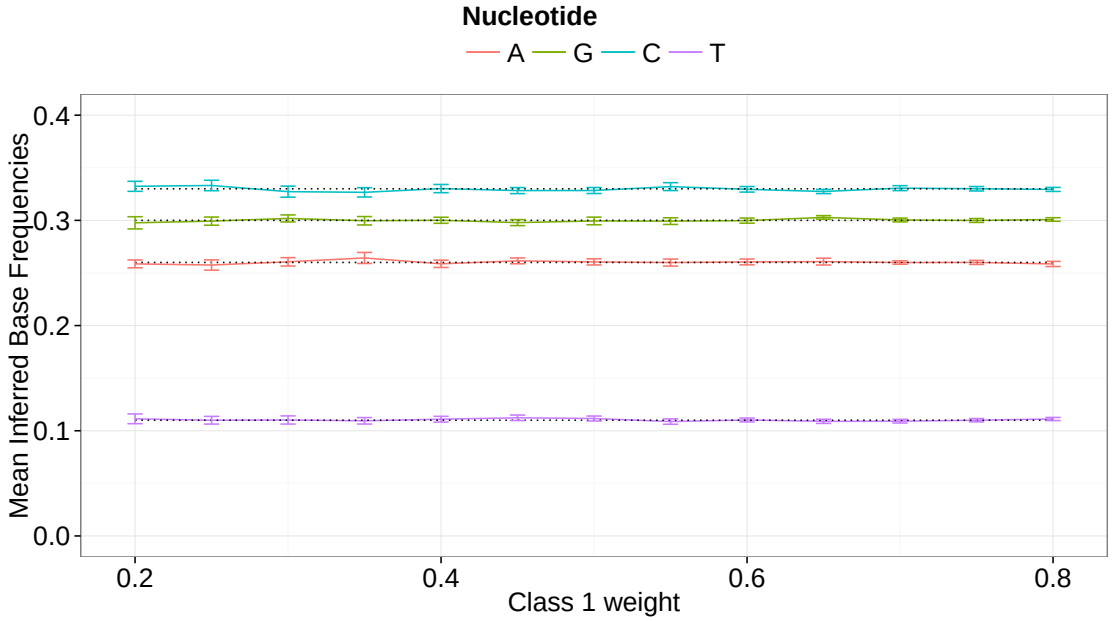

Figure 5.4: The mean inferred base frequency parameters for Class 1 of the Specific Case. The weight of Class 1 is shown on the x-axis, the base frequency is shown on the y-axis. The data points indicate the mean base frequency inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the base frequencies used to simulate the Class 1 component of the MSAs. The results indicate that IQ-TREE was able to accurately recover the base frequencies for Class 1 of the Specific Case simulations.

Results

IQ-TREE was able to recover the correct topology for all replicates at all values of w_1. Figures 5.4 and 5.5 show the inferred base frequencies for each class, averaged over the 20 replications at each value of w_1. The dotted lines give the true value of each base frequency. Clearly IQ-TREE is able to successfully recover the base frequencies of each class accurately. The inferred data points lie very close to the true values with the minimal deviance attributable to stochastic variation in the simulation of the MSAs.

Figures 5.6 and 5.7 show the inferred substitution rates for each class, averaged over the 20 replications at each value of w_1. The dotted lines give the true value of each substitution rate. Again we see that the inferred substitution rates closely

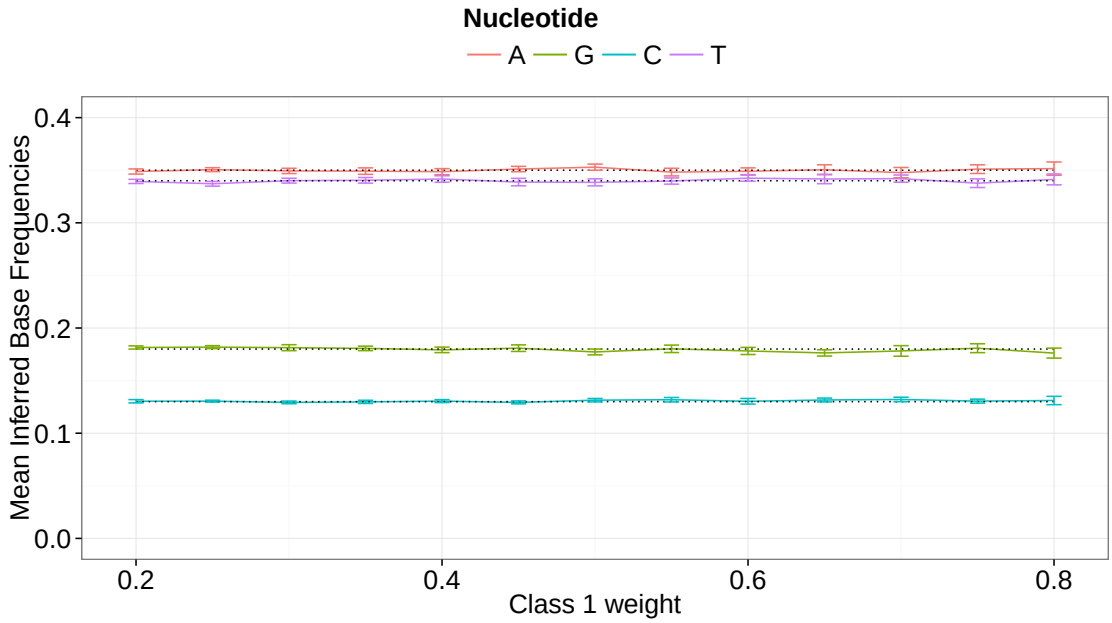

Figure 5.5: The mean inferred base frequency parameters for Class 2 of the Specific Case. The weight of Class 1 is shown on the x-axis, the base frequency is shown on the y-axis. The data points indicate the mean base frequency inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the base frequencies used to simulate the Class 2 component of the MSAs. The results indicate that IQ-TREE was able to accurately recover the base frequencies for Class 2 of the Specific Case simulations.

match the true values, although the deviance is noticeably greater than that seen for the base frequencies in Figures 5.4 and 5.5. This suggests that the stochastic variation in the simulation process has a greater effect on substitution rates than it does on base frequencies. Another feature noticeable in both figures is that the accuracy of the inference appears to be negatively correlated with the magnitude of the true substitution rate. This feature also appears to be related to the stochastic variation of the simulation process as it is more noticeable when the sequence length of the class in question is short. Referring to Figure 5.6 and focussing on the A↔T substitution rate, when w_1 is low the substitution rate is consistently underestimated. As w_1 increases, the length of sequence within the MSA generated under Class 1 increases and the accuracy of inference improves accordingly. The same feature is observed for the C↔T substitution rate in Class 2.

Assessing the performance of the GHOST model with respect to accurately recovering branch lengths is not as straightforward as it is for base frequencies or substitution rates. Our 12-taxon tree has 21 branch lengths, directly comparing true and inferred branch lengths individually would be cumbersome and not practical at all for larger trees. Moreover it is not possible if an incorrect topology is inferred. Many metrics for quantifying the distance between two trees have been proposed in the literature. We make use of the branch score, BS, as introduced by Kuhner and Felsenstein (1994). Another difficulty in assessing accuracy of branch length recovery is that since we use the BS, an absolute distance metric, we must establish some frame of reference so that we can assess whether the results we obtain are suitably close to the truth or not. To do this we make use of the partition model. The fundamental difference between the partition model and the GHOST model is that the partition model has knowledge of which sites in the alignment belong to which class. So in effect (and excluding the possibility of inferring the incorrect topology) the results of the partition model are identical to those that would be obtained by fitting GTR models to the Class 1 and Class 2 sequences independently. Thus we can consider the trees inferred by the partition model as a benchmark.

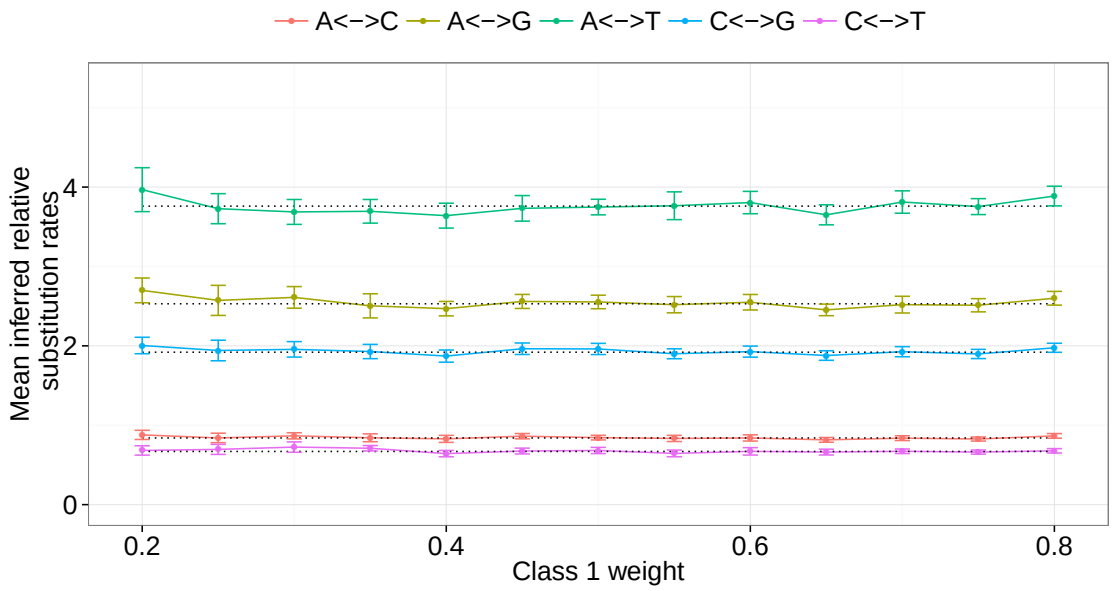

Figure 5.6: The mean inferred substitution rate parameters for Class 1 of the Specific Case. The weight of Class 1 is shown on the x-axis, the substitution rate is shown on the y-axis. The data points indicate the mean substitution rate inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the substitution rates used to simulate the Class 1 component of the MSAs. All rates are recovered by IQ-TREE with a reasonable level of accuracy. The error appears to be greater for substitution rates with higher true values, most notably the A↔T rate. The fact that the error decreases as w_1 increases suggests that it is primarily an artefact of stochastic variation in the simulation process, the effect is diminished as the length of the Class 1 component of the MSA increases.

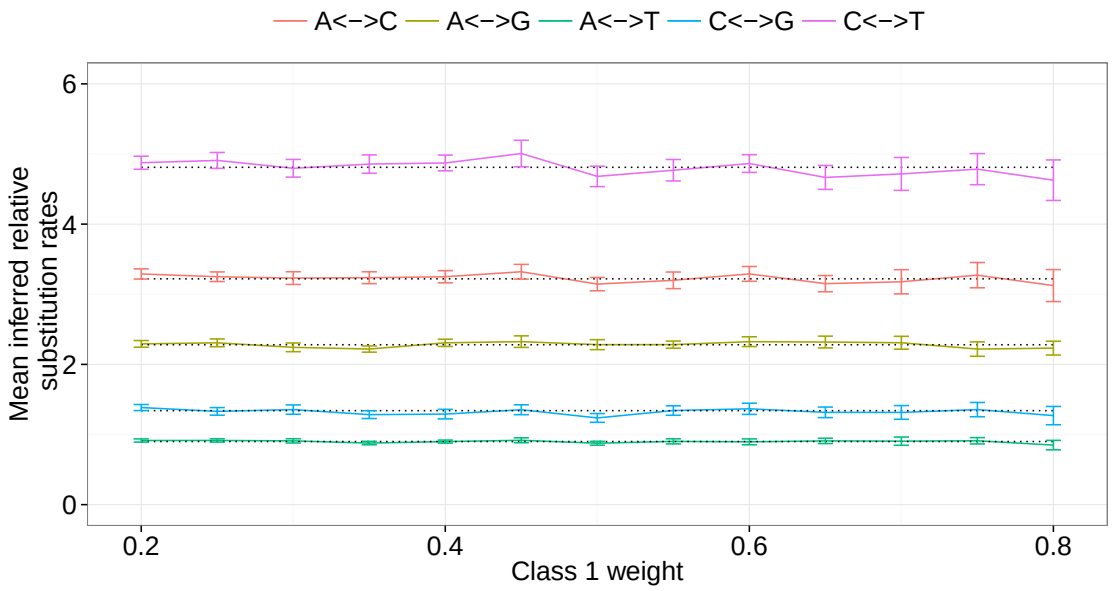

Figure 5.7: The mean inferred substitution rate parameters for Class 2 of the Specific Case. The weight of Class 1 is shown on the x-axis, the substitution rate is shown on the y-axis. The data points indicate the mean substitution rate inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the substitution rates used to simulate the Class 2 component of the MSAs. All rates are recovered by IQ-TREE with a reasonable level of accuracy. The error appears to be greater for substitution rates with higher true values, most notably the C↔T rate. The fact that the error decreases as w_1 increases suggests that it is primarily an artefact of stochastic variation in the simulation process, the effect is diminished as the length of the Class 1 component of the MSA increases.

Furthermore, we can consider the magnitude of the BS for the partition model as a measure of the noise introduced by the simulation process. This is due to the fact that the partition model is in no way misspecified, and a BS of zero would imply that the inferred tree is identical to the true tree. So when assessing the performance of the GHOST model we should hope to achieve a BS that approaches those achieved by the partition model. Figures 5.8 and 5.9 show the distance between the inferred and true trees for Class 1 and Class 2 respectively, as measured by the BS, averaged over the 20 replicates at each value of w_1. Each Figure displays the BS for both the GHOST and partition models. Both figures show that the difference in BS between the GHOST and partition models is small relative to the magnitude of the BS for the partition model.

5.2.2 General case

As in the Specific Case, each dataset was constructed as a partition of two classes with total sequence length being 10000 base pairs. Each of the 1000 datasets had its own set of parameters and its own tree topology. For each dataset the branch lengths, transition rates and base frequencies were selected randomly as described above. The topology for each dataset was generated randomly on 12 taxa. Finally, w_1 for each dataset was drawn from a $U(0.2, 0.8)$ distribution.

For each simulated dataset in both the Specific and General Cases *iqtree* was used to infer parameters under both the GHOST2 model and the partition model. In order to compare inferred parameters to the true parameters it is necessary to allocate each of the inferred classes to one of the true classes. This was done based on the BS, whichever inferred class had the smallest BS when compared to the true Class 1 tree was deemed to be the inferred Class 1, with the other class therefore becoming the inferred Class 2.

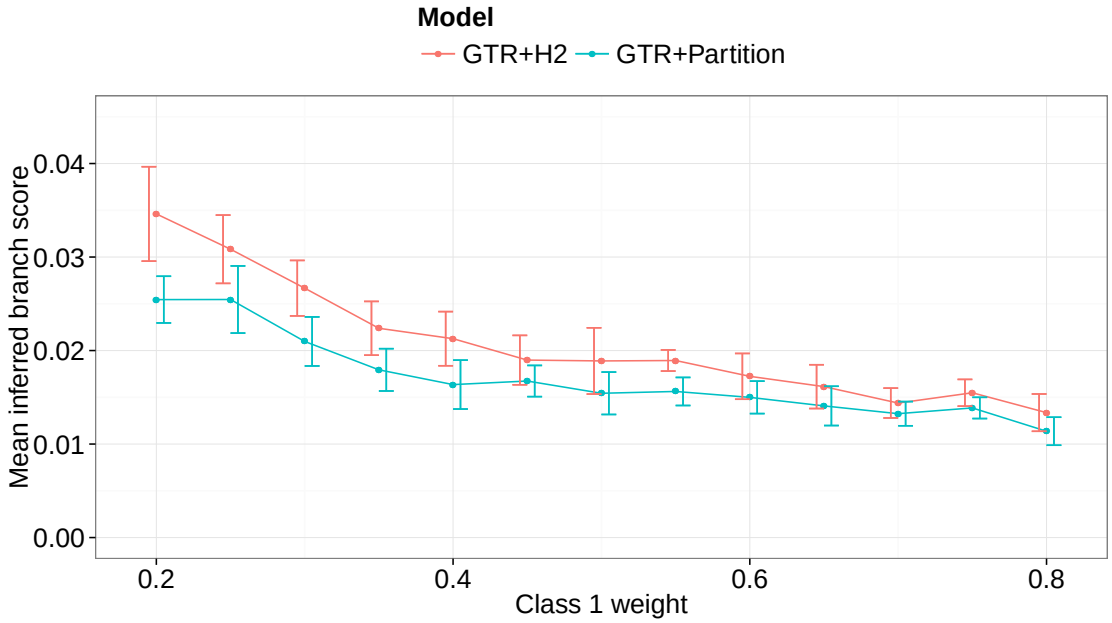

Figure 5.8: The mean inferred Branch Score (BS) for Class 1 of the Specific Case, for both the GHOST and partition models. The weight of Class 1 is shown on the x-axis, the BS is shown on the y-axis. The data points indicate the mean BS inferred by IQ-TREE using ML-GTR+H2 or ML-GTR+PART over the 20 replicate MSAs at that Class 1 weight. The difference in BS between the partition and the GHOST models is small in comparison to the magnitude of the partition model BS, suggesting that with respect to branch length recovery IQ-TREE using the GHOST model performs as well as we could expect. Furthermore this distance decreases as w_1 increases (as the sequence generated under Class 1 becomes longer), implying consistency.

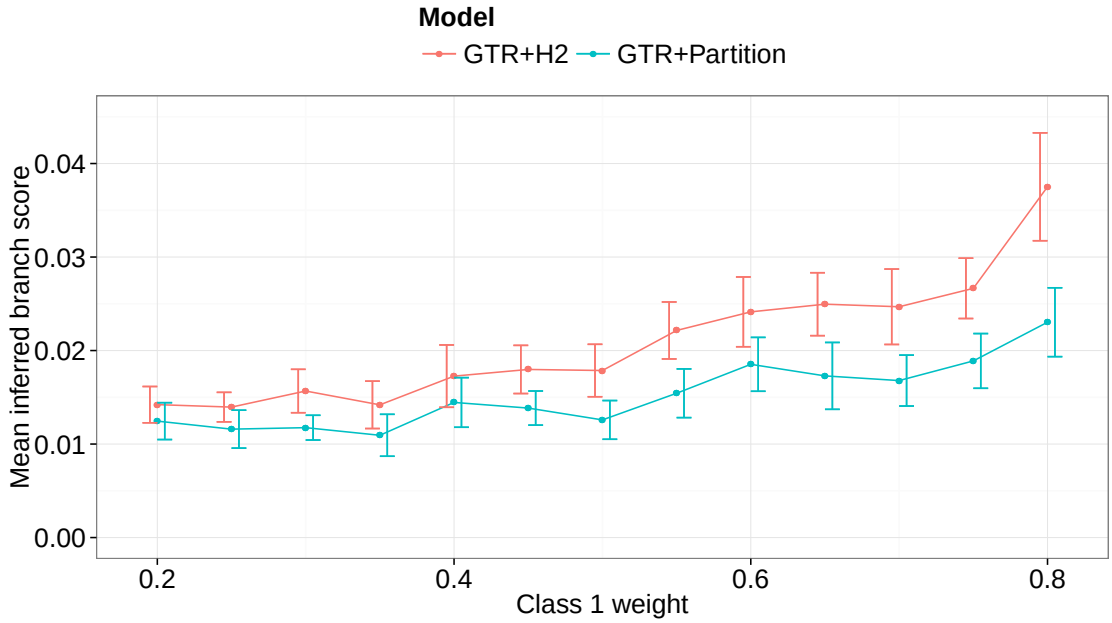

Figure 5.9: The mean inferred Branch Score (BS) for Class 2 of the Specific Case, for both the GHOST and partition models. The weight of Class 1 is shown on the x-axis, the BS is shown on the y-axis. The data points indicate the mean BS inferred by IQ-TREE using ML-GTR+H2 or ML-GTR+PART over the 20 replicate MSAs at that Class 1 weight. The difference in BS between the partition and the GHOST models is small in comparison to the magnitude of the partition model BS, suggesting that with respect to branch length recovery IQ-TREE using the GHOST model performs as well as we could expect. Furthermore this distance increases as w_1 increases (as the sequence generated under Class 2 becomes shorter), implying consistency.

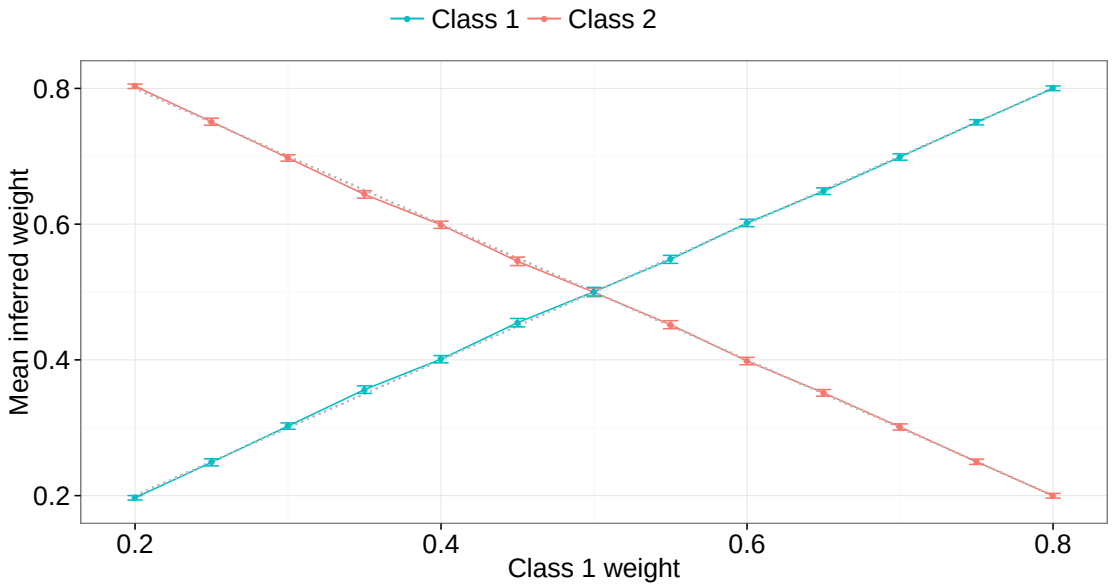

Figure 5.10: The mean inferred weights for Classes 1 and 2 of the Specific Case. The weight of Class 1 is shown on the x-axis, the inferred weight is shown on the y-axis. The data points indicate the mean weight inferred by IQ-TREE using ML-GTR+H2 over the 20 replicate MSAs at that Class 1 weight. The dotted lines indicate the true values of the weights for Class 1 and Class 2. The results indicate that IQ-TREE was able to accurately recover the weights of the two classes for Specific Case simulations.

Metrics

For the General Case the performance of the GHOST model for each dataset had to be quantified in a standardised manner, so that an overall indication of performance could be established over the 1000 independent datasets. The rate score, RS, is defined as the mean absolute relative error between inferred and true rate parameters, over each of the 10 rate parameters (5 in each of the 2 classes). More formally, if $\mathbf{R}_j^T = (R_{jAC}^T, R_{jAG}^T, R_{jAT}^T, R_{jCG}^T, R_{jCT}^T)$ is the vector of true rate parameters for the j^{th} class and $\mathbf{R}_j^I$ is the corresponding vector of inferred rate parameters for the j^{th} class then:

$$RS = \frac{1}{5m} \sum_{j=1}^{m} \sum_{k=AC}^{CT} \frac{|R_{jk}^I - R_{jk}^T|}{R_{jk}^T}.$$

The frequency score, FS, is defined analagously for the 8 base frequency parameters (4 in each of the 2 classes). The inferred trees are measured against the true trees using 2 distinct metrics. The Robinson-Foulds distance (Robinson and Foulds, 1981), RF, was used to check for any topological discrepancy. As for the Specific Case, BS, was used to quantify the distance between trees taking branch lengths into account. WS is defined simply as the absolute difference between the inferred and true value of w_1.

Results

The RF distances indicated that the GHOST model was able to recover the correct topology for 996 of the 1000 simulated datasets. The four anomalies returned an RF score of 2, meaning the inferred trees differed from the true trees on only one internal edge. Further investigation of the true parameters for these datasets indicated that in each case the internal edge in question was exceedingly short in at least one of the classes. The partition model also failed to recover the correct topology for these four datasets.

As with the assessment of the BS for the Specific Case, we must again compare our results to the partition model in order to assess the performance of the GHOST model. The mean and standard deviations of RS, FS, BS and WS for both the GHOST and partition models are displayed in Table 5.1.

For the GHOST model the mean RS of 0.083 implies that on average the inferred substitution rates differed from their true values by 8.3%. This result is acceptable given that the RS score for the partition model is 0.058, an average difference of 5.8%. So the error attributable to the use of the GHOST model is just less than half that attributable to the noise in the simulation process. Similarly the FS for the GHOST model is 0.024, meaning inferred base frequencies were accurate to within 2.4% of the true values. The partition model reported an average error of 1.6%, so once again the error attributable to the use of the GHOST model is around half that attributable to the noise in the simulation process. We cannot interpret the BS in terms of percentage but the same pattern emerges. The GHOST model has a mean BS of 0.026, compared to the partition model mean BS of 0.019, again showing that the majority of the error in the GHOST model is attributable to the noise in the simulation process. The mean WS of 0.012 simply means the average difference between the inferred weight and the true weight was just 0.012, a self-evidently positive result.

5.3 Soft classification of sites to classes

An auxiliary benefit of the ML implementation of the GHOST model in IQ-TREE is that we can easily soft classify the sites according to their probability of belonging to a particular class. We define P_{ij} as the probability of the i^{th} site belonging to the j^{th} class. A simple application of Bayes Law utilising components of the likelihood expression yields:

Metric	GHOST	Partition
Mean RS (SD)	0.083 (0.042)	0.058 (0.026)
Mean FS (SD)	0.024 (0.010)	0.016 (0.005)
Mean BS (SD)	0.026 (0.009)	0.019 (0.006)
Mean WS (SD)	0.012 (0.012)	- (-)*

* The true weight is a known input of the partition model therefore these scores are 0 by definition.

Table 5.1: General Case simulation results - Summary statistics for the rate score (RS), frequency score (FS), branch score (BS) and weight score (WS) for comparisons of the GHOST model and the partition model to the true parameters for the 1000 simulations conducted under the General Case.

$$P_{ij} = \frac{w_j L_{ij}}{L_i},$$

where w_j is the weight of the j^{th} class in the mixture model; L_{ij} is the likelihood of the data at site i evolving under the j^{th} class in the mixture model; and L_i is the weighted likelihood across all classes of the data at site i evolving under the mixture model.

This classification can be used to identify sites in the alignment that belong with high probability to a particular class of interest. To demonstrate we can perform a soft classification on one of the MSAs generated for the Specific Case simulations, for simplicity we have chosen an MSA where Class 1 and Class 2 are of equal weight. Figure 5.11 displays the results of the soft classification. As one would expect the probability of a site belonging to Class 1 is significantly higher for those sites that were simulated from the Class 1 parameters than those simulated from Class 2 parameters. While there are individual sites that belong to Class 2 that are classified as having a high probability of belonging to Class 1 and vice versa, the distinction between the distributions of Class 1 and Class 2 sites is clear.

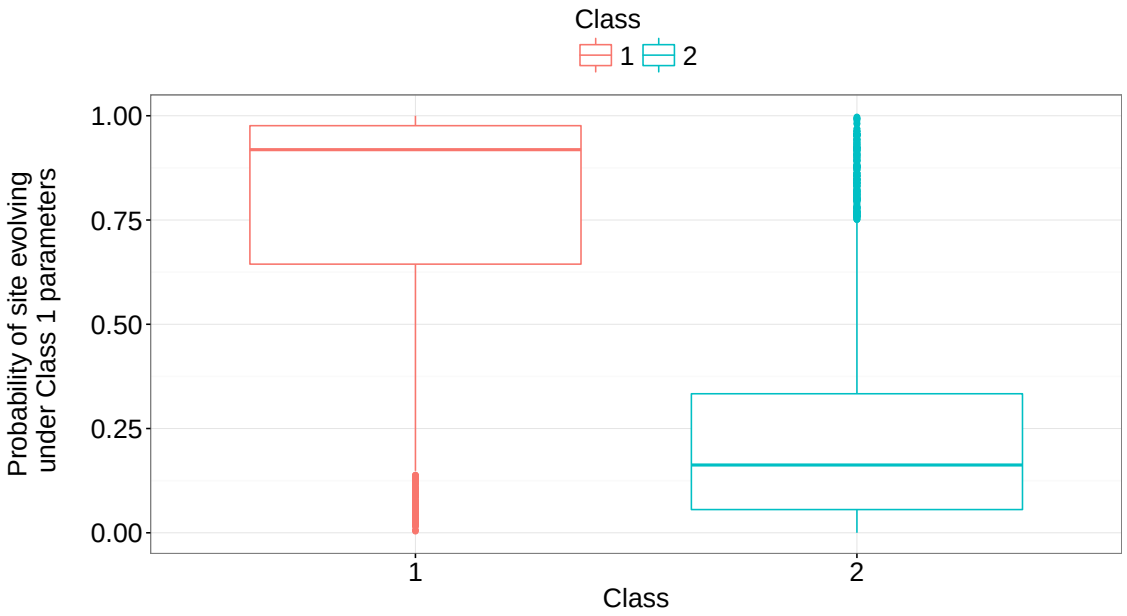

Figure 5.11: Soft classification of sites to classes - the probability of a site belonging to Class 1 is shown on the y-axis, the two Classes are shown on the x-axis. The boxplots clearly show that sites generated under Class 1 parameters are classified as having a higher probability of belonging to Class 1 than sites generated under Class 2.

Chapter 6

The Convergent Evolution of Electric Fishes

6.1 Background

To investigate the performance of the GHOST model using real data we applied it to part of the coding region of a sodium channel gene, $Na_v1.4a$. Zakon *et al.* (2006) demonstrated the role that this gene has played in the convergent evolution of the electric organ amongst electric fish species from South America and Africa. Their dataset consists of 11 fish species: four gymnotiforms, an order of electric fish species native to South America, one mormyrid, an order of electric fish species from Africa, and six non-electric fishes from a variety of locations. Specifically the gymnotiforms were the Black Knifefish (*Sternopygus macrurus*), the Electric Eel (*Electrophorus electricus*), the Pintailed Knifefish (*Brachyhypopomus pinnicaudatus*) and the Brown Ghost Knifefish (*Apteronotus leptorhynchus*); the mormyrid was the Elephantnose Fish (*Gnathonemus petersii*); and the non-electric fishes were the Silver Arowana (*Osteoglossum bicirrhosum*), the Channel Catfish (*Ictalurus punctatus*), the Zebrafish (*Danio rerio*), the Green Spotted Pufferfish (*Tetraodon nigroviridus*), the Panther Pufferfish (*Takifugu pardalis*) and the Clown Knifefish (*Notopterus chi-*

tala).

Zakon *et al.* (2006) used reverse transcription polymerase chain reaction to examine the expression pattern of $Na_v1.4a$ in both electric and non-electric fishes. They found the gene was expressed in muscle tissue in non-electric fishes but has lost this expression in electric fishes, instead being expressed in the electric organ. Critically there was one exception to this finding - $Na_v1.4a$ is still expressed in muscle in the Brown Ghost Knifefish, and not in its electric organ. This is due to the fact that the electric organ of the Brown Ghost Knifefish is derived from neural tissue, whereas the electric organs of the other electric fishes are derived from muscle tissue (Zakon *et al.*, 2006).

They also analysed synonymous and non-synonymous substitutions using the PAML software package (Yang, 2007), finding evidence that this gene was evolving under positive selection in both the gymnotiforms and the mormyrids. They performed the analysis both by amino acid site and by branch, finding that approximately 57% of the sites are highly constrained. They found evidence of sites evolving under positive selection on all branches leading to electric fish species except the Brown Ghost Knifefish, suggesting that positive selection played a role in the transition of $Na_v1.4a$ expression in the electric fishes (excluding the Brown Ghost Knifefish), from muscle tissue in their ancestral lineages to the electric organ in their current form.

They identify specific amino-acid replacements that occur convergently in domains of the gene that have been shown experimentally to influence sodium channel inactivation. This inactivation is critical in producing the unique electric waveform that each species uses for communication and identification purposes.

6.2 Data

$Na_v1.4a$ consists of four domains (DI - DIV), each of which has 6 membrane spanning regions (S1-S6). The dataset consists of 2178 base pairs from the coding region of

Common Name	Scientific Name	GenBank
Black Knifefish	*S. Macrurus*	AF378144
Pintailed Knifefish	*B. Pinnicaudatus*	DQ351534
Electric Eel	*E. Electricus*	M22252
Brown Ghost Knifefish	*A. Leptorhynchus*	DQ351533
Elephantnose Fish	*G. Petersii*	DQ275142
Silver Arowana	*O. Bicirrhosum*	DQ336343
Channel Catfish	*I. Punctatus*	AY204537
Zebrafish	*D. Rerio*	DQ149506
Green Spotted Pufferish	*T. Nigroviridus*	CAAE01014976
Panther Pufferfish	*T. Pardalis*	AB030482
Clown Knifefish	*N. Chitala*	DQ336344

Table 6.1: The 11 fish species in the dataset and the GenBank accession numbers for the $Na_v1.4a$ gene of each species.

$Na_v1.4a$. The 2178 base pairs are a concatenation of six separate exons from S6 of DII to S2 of DIV. Using the sequence for the Electric Eel (GenBank accession number M22252) as a reference the regions correspond to sites 2149 - 2880, 2893 - 2916, 2998 - 3027, 3037 - 3081, 3148 - 3192 and 3229 - 4530. Table 6.1 lists the 11 species and their GenBank accession numbers.

6.3 Identifying the optimal GHOST model

We used Akaike's Information Criterion (Akaike, 1974) (AIC) to determine the number of classes that provided the best fit between tree, model and data. We fitted a ML-GTR+Hm model to the electric fish data for $m \in \{2, 3, 4, 5, 6, 7, 8\}$. We then compared the AIC score for each model. The results can be seen in Figure 6.1 and indicate that 4 classes provides the best fit for the electric fish dataset.

In order to test whether this result was reliable we repeated the fitting process

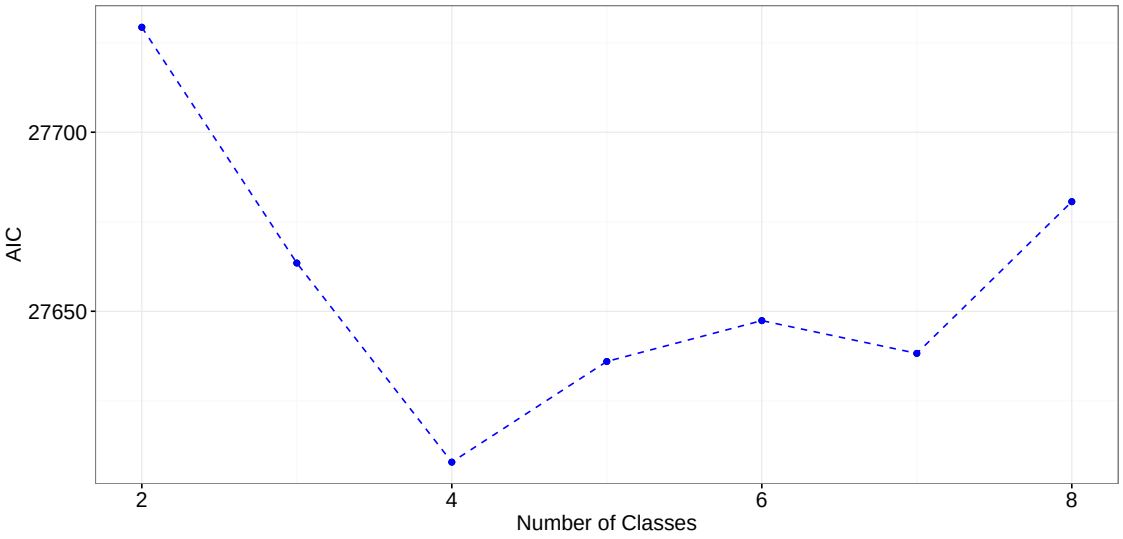

Figure 6.1: The AIC scores achieved by varying the number of classes while fitting an ML-GTR+H model. The results indicate that 4 is the optimal number of classes for this dataset.

100 times each for $m \in \{3, 4, 5, 6, 7\}$, (2 and 8 classes were both ruled out upon observation of Figure 6.1). Figure 6.2 shows that 4 classes consistently provides the best fit to the data. It is interesting to note the apparent increase in variance of AIC scores as the number of classes is increased. For the electric fish dataset each additional class carries with it an extra 28 parameters to be inferred by the model. It is not surprising then that as the parameter space increases so too does the difficulty in navigating the likelihood surface. The EM-algorithm becomes more susceptible to finding locally optimal solutions when the parameter space is high dimensional. Given this finding we recommend repeated fitting of the GHOST model, paying attention to the consistency of the inferred parameters.

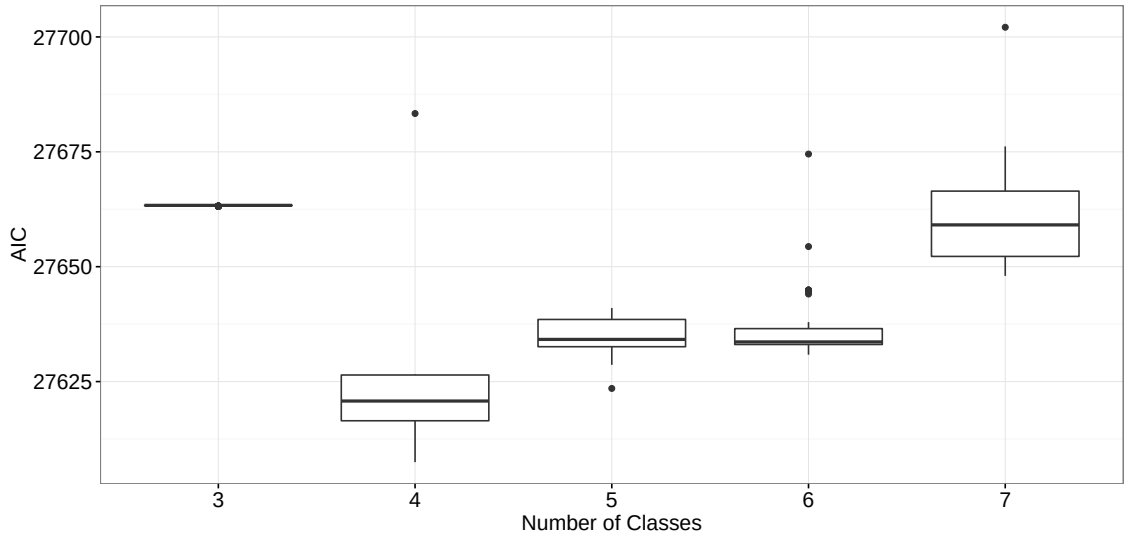

Figure 6.2: The AIC scores achieved by varying the number of classes while fitting an ML-GTR+H model. For each class, m, 100 ML-GTR+Hm models were fitted to the data independently.

6.4 Analysis of classes inferred by ML-GTR+H4 model

Tables 6.2 and 6.3 give the substitution rates and base frequencies inferred by ML-GTR+H4 for each class while Figure 6.3 shows the four trees inferred by ML-GTR+H4 along with the class weights. Careful attention must be paid to the individual branch length scales for each tree in Figure 6.3. The trees have been scaled appropriately so that the detail of each can be seen, the apparent similarity in their size is therefore an illusion. Notably the tree of the largest class by weight, Class 1, is significantly smaller than those of the other three classes. The sum of its branches, or total tree length (TTL), is 0.19. The next smallest tree belongs to Class 4 which is an order of magnitude larger with a TTL of 1.9. Recall that the PAML analysis of Zakon *et al.* (2006) found that 57% of the sites in the alignment were highly conserved. The ML-GTR+H4 appears to be in agreement, approximately 54.47% of sites belong to Class 1 and the size of the branch lengths indicates these sites are also highly conserved.

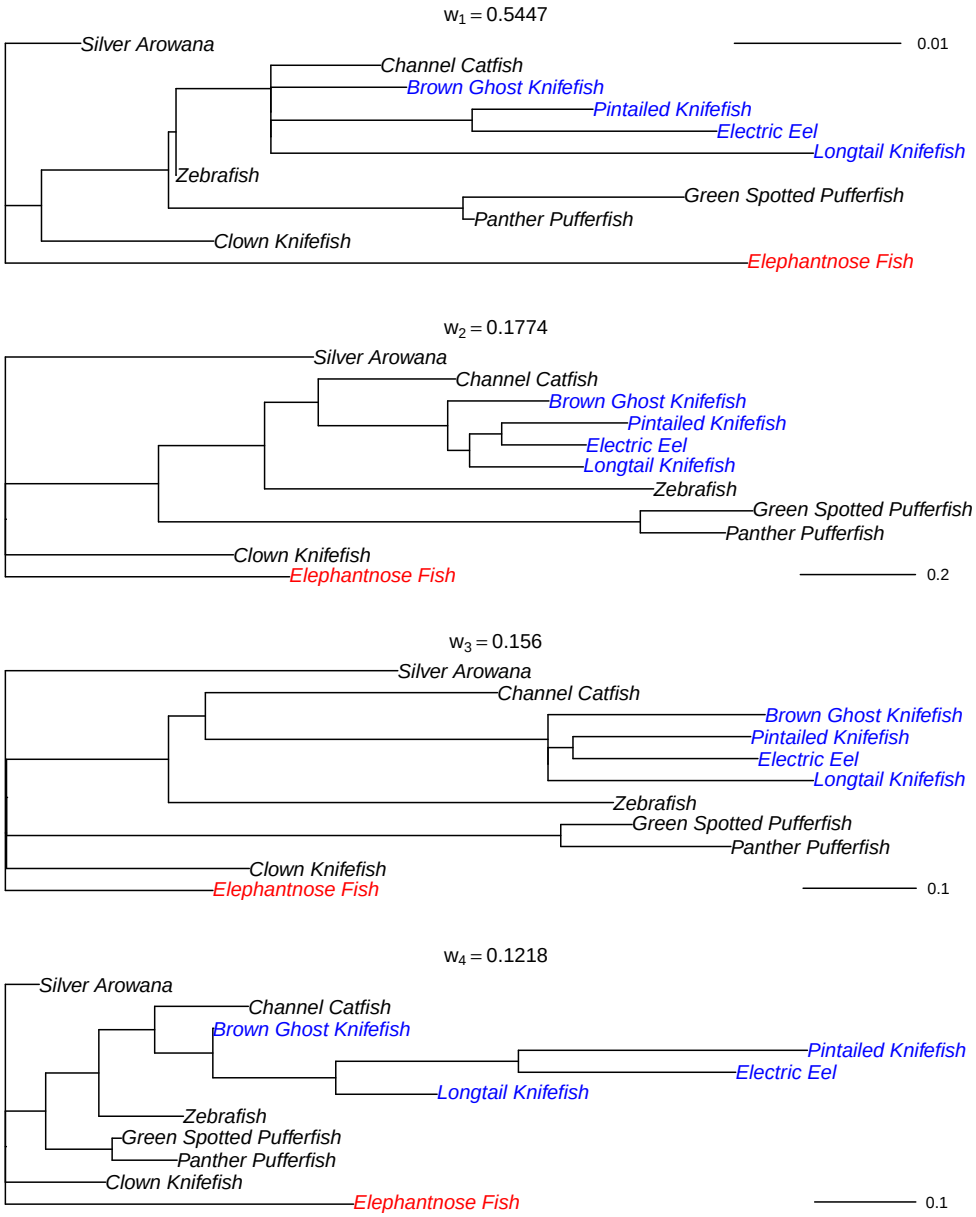

Figure 6.3: The four trees obtained from fitting the ML-GTR+H4 model to the electric fish data. The inferred weight of each class is indicated above each tree. Note the different scales for each tree, the dominant class (by weight) is much slower evolving than the three smaller classes. An indication of this is the total tree length (TTL) for the four classes: $\mathrm{TTL}_1 = 0.19$, $\mathrm{TTL}_2 = 5.12$, $\mathrm{TTL}_3 = 3.35$ and $\mathrm{TTL}_4 = 1.90$.

Class	A↔ C	A↔ G	A↔ T	C↔ G	C↔ T
1	56.78	39.08	16.54	22.15	52.61
2	1.21	3.12	3.07	1.39	3.58
3	0.66	25.47	0.86	0.0001	8.70
4	0.81	0.56	0.79	0.85	0.71

Table 6.2: The relative substitution rates inferred by ML-GTR+H4 for the electric fish dataset. Rates are shown relative to the G↔T substitution rate which is fixed at 1.

The smallest of the four classes (by weight) inferred by the model corresponds well to Zakon's hypothesis of convergent evolution of $Na_v1.4a$ between the South American and African electric fishes. We will denote this class the 'convergent' class, its tree is shown in more detail in Figure 6.4. Sites in the convergent class were much faster evolving in electric than non-electric fishes. This is indicative of either a relaxation of purifying selection pressure; an introduction of positive selection pressure; or a combination of both. The notable exception is the Brown Ghost Knifefish, which appears relatively conserved. This concurs with the conclusions of Zakon *et al.* who found that $Na_v1.4a$ was expressed in muscles in the non-electric fish and the Brown Ghost Knifefish, but in the electric organ of the other electric fishes. Amongst the electric fish in this dataset the Brown Ghost Knifefish is unique in that its electric organ has evolved from neural tissue, whereas the other electric species have electric organs which were evolved from muscle tissue. The ability of the GHOST model to isolate such a small phylogenetic signal (the inferred weight of the convergent class being 12.18%, the smallest of the 4 classes) is most encouraging. The clear distinction between the Brown Ghost Knifefish and the other electric fish adds significant support to the hypothesis that the sites that belong to the convergent class are likely to be influential in the transition of gene expression from muscle to electric organ.

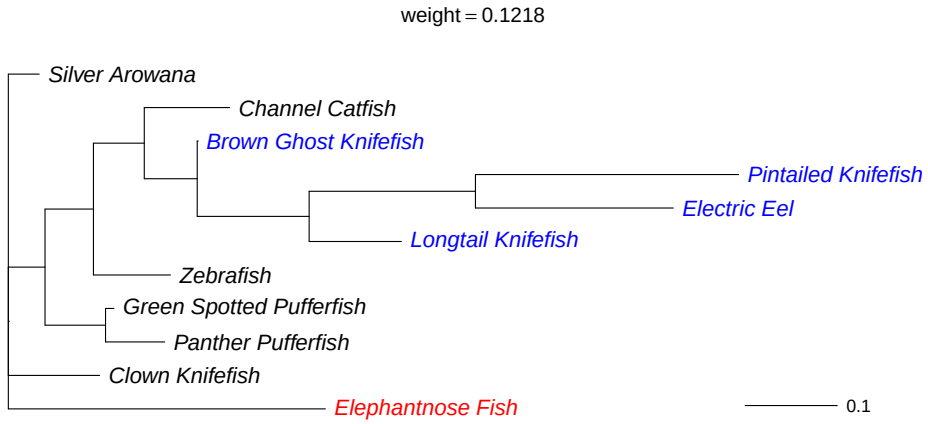

Figure 6.4: The convergent class inferred by ML-GTR+H4. The 11 fish species comprised four South American electric fish (blue), one African electric fish (red), and six non-electric fish from various locations. The smallest class from the GHOST4 model shows that in comparison to the electric fish the non-electric species are relatively conserved.

Class	A	C	G	T
1	0.271	0.161	0.234	0.334
2	0.227	0.325	0.239	0.209
3	0.122	0.437	0.221	0.220
4	0.311	0.134	0.371	0.184

Table 6.3: The base frequencies inferred by ML-GTR+H4 for the electric fish dataset.

6.5 Soft classification of sites to classes

As shown in Chapter 5 the GHOST model enables sites to be soft classified according to the probability that they belong to individual classes. This feature facilitates the prospective identification of functionally important sites in an alignment. Zakon *et al.* (2006) report several sites from the dataset that are influential in the inactivation of the sodium channel, a process critical to electric organ pulse duration. Figure 6.5 shows that these sites have a higher than average probability of belonging to the convergent class. This finding adds further evidence that the GHOST model is able to identify a subtle phylogenetic signal in a sequence alignment. Given that there are many other sites in the alignment with a high probability of belonging to the convergent class the GHOST model may become an important tool to identify sites of potential functional importance in an alignment, helping to focus the experimental work of biologists.

In addition to providing insight on an individual site basis, the soft classification can also help to inform us about the nature of the classes themselves. Recall that the nucleotides in coding sequences can be grouped into three, referred to as codons, which ultimately determine the amino acid that is formed. With four bases there are 64 unique codons but there are only 20 amino acids. Thus there is significant redundancy, the result of which is that the amino acid can often be determined by the first and second codon positions, with the third bearing no influence. Table 6.4 shows the relative frequency of codon position for each class. If class membership and codon position were independent attributes of each site then we should expect the relative frequency of each codon position to be approximately one third for each class. This is not what we observe. Sites in codon positions 1 and 2 are heavily under-represented in Classes 2 and 3, totalling just 27.5% and 32.0% respectively. We therefore suspect that a comparatively large proportion of the substitutions that occur on sites in these classes are synonymous: not resulting in an amino acid replacement and therefore having no functional impact on the organism. Conversely

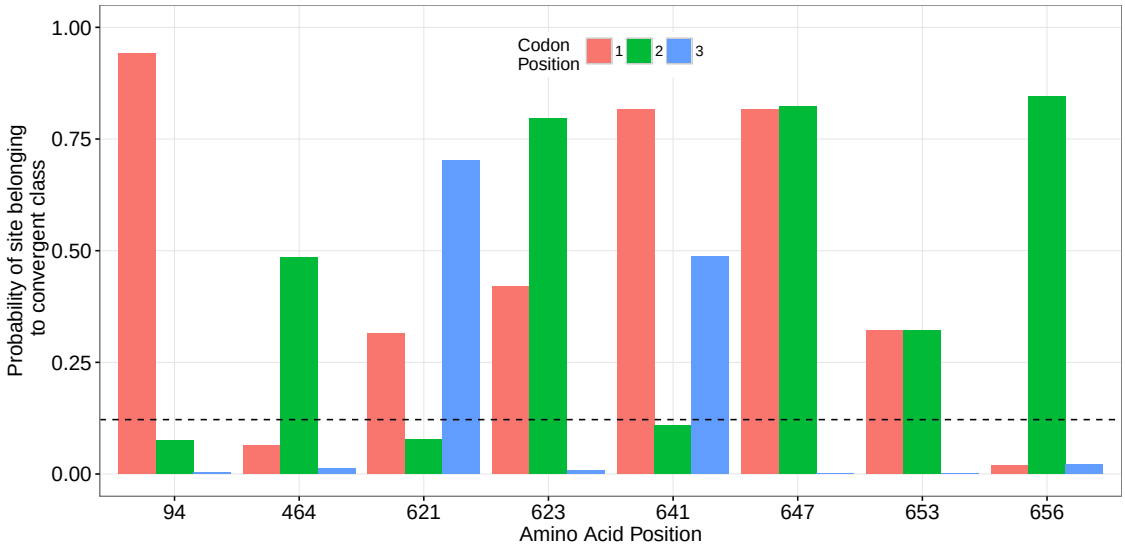

Figure 6.5: Probability of sites belonging to the convergent class by codon position. The amino acid positions selected correspond with those identified by Zakon *et al.* as being critical to the inactivation of the Na^+ gene. The line at 0.1218 represents the average probability of belonging to the convergent class over all sites in the alignment. Sites at which nucleotide substitutions lead to functionally important amino acid replacements have a high probability of belonging to the convergent class. For example, at amino acid site 647 an otherwise conserved proline (codon CCN) is replaced by a valine (GTN) in the Pintailed Knifefish and a cysteine (TGY) in the Electric Eel. Substitutions at codon position 1 and 2 are necessary for both of these amino acid replacements and we find these sites have a high probability of belonging to the convergent class.

Codon Position	Class 1	Class 2	Class 3	Class 4
1	0.404	0.183	0.203	0.404
2	0.482	0.092	0.126	0.284
3	0.114	0.725	0.671	0.312

Table 6.4: The relative frequency of codon position for each of the four inferred classes.

Class 1 has an overrepresentation of sites in codon position 1 and 2. This would suggest that a comparatively large proportion of the substitutions that occur in Class 1 sites are non-synonymous: resulting in amino acid replacements that influence the fitness of the organism. Bear in mind however that the branch lengths on the Class 1 tree (see Figure 6.3) are very short compared to the other classes. Even taking into account the fact that it is the largest class by weight ($w_1 = 0.5447$), the total number of substitutions attributable to Class 1 would be negligible compared to the other classes. This leaves only the convergent class, Class 4. Approximately 69% of the sites in the convergent class lie in codon positions 1 or 2. Thus we can expect a significant number of the substitutions attributable to the convergent class to be non-synonymous, leading to amino acid replacements and functional changes in the organism. We can therefore conclude that even though the convergent class is smallest by weight, it appears to be the primary catalyst of evolution via natural selection within $Na_v1.4a$ amongst these species.

6.6 ML-GHOST vs comparable models and methods

The performance of the IQ-TREE implementation of the GHOST model was measured against that of the models of Pagel & Meade (Pagel and Meade, 2004; Meade and Pagel, 2008), implemented in a Bayesian framework in their software package BayesPhylogenies (Pagel and Meade, 2006). Comparisons were made between the

Site Index	Codon Position	Probability
1999	1	0.9947
1297	1	0.9941
178	1	0.9870
762	3	0.9841
1660	1	0.9785
469	1	0.9764
676	1	0.9741
1266	3	0.9723
680	2	0.9692
1839	3	0.9640

Table 6.5: The ten sites in the alignment with the highest probability of belonging to the convergent class. Note the over-representation of codon position 1, suggesting these sites are likely to have non-synonymous substitutions present.

GHOST model with 4 classes (GHOST4), the Pagel & Meade MRM with 4 classes (PMR4), the Pagel & Meade MBL with 4 classes (PMB4) and the two Pagel & Meade models implemented simultaneously with 4 classes (PMRB4). These models were selected from the wider literature as they are nested within each other (PMR4 and PMB4 are special cases of GHOST4; and GHOST4 is a special case of PMRB4). A fair and reasonable method to compare the results of a ML analysis to a Bayesian analysis was not immediately apparent. Instead of constructing a distribution of results from the converged Monte-Carlo Markov Chain as would normally be the case, it was decided to treat the MCMC as a rudimentary algorithm for searching the parameter space. Thus we decided to run the MCMC for 2 million iterations and use the highest likelihood obtained for purposes of calculating the AIC.

As seen in Table 6.7 the GHOST model performs significantly better than the models of Pagel & Meade in terms of both AIC and computation time. Furthermore, it also offers a much more straightforward biological interpretation. It might

Taxa	Nucleotide Sequence
Elephantnose Fish	G T G G T A G G T A
Clown Knifefish	T T G G G A A G G G
Silver Arowana	T T G G G A A G G G
Panther Pufferfish	T T G G G A A G G G
Zebrafish	T T G G G A A G G G
Green Spotted Pufferfish	T T G G G A A G G G
Channel Catfish	T T G G G A A T G G
Brown Ghost Knifefish	T T G G G A A T G G
Longtailed Knifefish	T G T T G A A T T G
Electric Eel	T G T T T G T A A T
Pintailed Knifefish	G T G T G T G A A T

Table 6.6: The sequence alignment corrsponding to the ten sites identified in Table 6.5, ordered from highest probability of belonging to the convergent class to lowest. Clearly the overwhelming majority of substitutions (highlighted in magenta) occur in the electric fish.

Model	AIC	Time (s)
GHOST4	27164	64
PMR4	27512	2026
PMB4	27594	2018
PMRB4	27281	8938

Table 6.7: The results show that when applied to the electric fish dataset the GHOST model in IQ-TREE clearly outperformed all of the Pagel & Meade models in Bayes Phylogenies. Their best fitting model, the PMRB4, was inferior to GHOST in terms of AIC by 117 units and it took approximately 140 times longer to run.

appear at first that the GHOST4 model is identical in theory to the PMRB4 model, however this is not the case. The PMRB4 model is a simultaneous and independent implementation of both the PMR4 and the PMB4. This means that 2 separate sets of weights are reported, one set applying to the 4 rate classes and one set applying to the 4 branch length classes. So in effect the results of the PMRB4 model actually imply 16 classes, not 4 as desired. A model that has 4 different substitution rate matrices and 4 different branch length sets independent of each other lacks biological plausibility. In fairness to Pagel & Meade, they have not advocated the simultaneous implementation of their PMR and PMB models in the literature, but it is an available option in their software.

Chapter 7

Conclusion

We have shown that the majority of the popular methods and models used for phylogenetic inference are not well suited when the multiple sequence alignment (MSA) evolved under heterotachous conditions. Using both simulation and theoretical analysis we establish that the heart of the problem lies in the misspecification between the evolutionary processes that gave rise to the data and the model of sequence evolution adopted in the phylogenetic inference process. Current methods that address this misspecification have significant drawbacks: they may be prohibitively slow, as in Bayesian MCMC methods; impose additional restrictive and potentially unrealistic assumptions, as in partition models; or lack critical features, such as tree search heuristics. The absence of computationally efficient inference software that can effectively model heterotachously-evolved MSAs is therefore a thorn in the side of modern phylogenetic analysis.

We have addressed this issue by proposing GHOST, a mixture model that offers significant flexibility in its ability to model heterogeneity of evolutionary rates, across both sites and lineages. We have implemented GHOST in a ML framework within the phylogenetic inference software IQ-TREE. We conducted several rigorous simulation studies which established that the IQ-TREE was able to use ML and the GHOST model to correctly infer tree topology, branch lengths and substitution model parameters from heterotachously-evolved MSAs. Furthermore, it was able to

consistently do so for a large number of MSAs, evolved from randomly generated tree topologies, branch lengths and substitution model parameters.

We applied the GHOST model to a MSA from the coding region of a sodium channel gene from 11 species of fish (Zakon *et al.*, 2006). This MSA had been shown in the literature to offer evidence of the convergent evolution of the electric organ amongst two distinct lineages of electric fishes, one from South America and the other from Africa. Applying the GHOST model to this MSA, IQ-TREE was able to extract a subtle phylogenetic signal which identified sites that were fast evolving in the electric fishes while remaining largely conserved in the non-electric fishes. Some sites identified as fitting this evolutionary pattern had been previously shown by biologists to play a critical role in the function of the electric organ.

The GHOST model shows great promise for improving the accuracy of phylogenetic inference. However, there is still much work to be done on three fronts: simulation; extension; and application. The simulation studies we carried out were successful, but they were limited to 12 taxa, two model classes and nucleotide sequence data. These simulation studies need to be extended. The size and availablility of MSAs is continually growing. The GHOST model must be shown to perform satisfactorily with increased taxon sampling before it can be used with confidence on larger MSAs. The GHOST model must also be shown to be able to reliably infer parameters when more than two classes are present. We have already seen that Akaike's Information Criterion (AIC) Akaike (1974) indicated four was the optimal number of classes for the electric fish MSA. A great deal of phylogenetic inference is carried out on data at the amino acid or codon level. Our simulation studies have not assessed the performance of the GHOST model with these types of data. Simualtion studies should therefore be carried out to verify its suitability before the model is used to analyse these types of data.

The implementation of the GHOST model in IQ-TREE is still in its infancy. There are several opportunities to refine and extend the implementation. The model currently requires the user to define the desired number of classes. This results in

the user having to run a set of analyses with a range of different class numbers and then compare results to determine the best fit. It would be desirable for the number of classes to be an extra parameter, inferred automatically by IQ-TREE to provide the best fit to the data. The GHOST model is currently not compatible with IQ-TREE's model selection tool. Given the reason for its development (that most current models cannot adequately model heterotachously-evolved data), it is essential that the GHOST model is an available option for a tool specifically designed to inform phylogeneticists of the best fitting model for their data. The GHOST model currently cannot be implemented in conjunction with other models such as an invariable sites model or the discrete gamma model. The ability to do so could provide some economy. If the GHOST model, as currently implemented, infers a primarily invariable class (*i.e.* one class has a tree with very short branch lengths) then a significant parameter saving could be made by capturing this class via an invariable sites model. Similarly if some classes were found to have branch lengths that were approximately scalar multiples of each other, then the incorporation of the discrete gamma model could have similar benefits.

We believe that the GHOST model in combination with its efficient implementation in the established IQ-TREE software will find a broad range of applications in the field. There are many examples in the literature whereby phylogenies obtained from traditional ML inference conflict with the well-established consensus. In these cases the GHOST model has the potential to resolve disputed phylogenies. The GHOST model should also lead to better estimation of branch lengths for heterotachously-evolved MSAs which may in turn lead to the revision of accepted divergence dates among species. We show in the manuscript that the GHOST model can detect subtle phylogenetic signals related to particular features of the taxa. Furthermore, it is able to identify candidate sites in the alignment that may belong with high probability to a particular class. These features may help to focus the work of biologists endeavouring to identify sites within a MSA linked to particular morphological traits. By the same token, the GHOST model could prove beneficial insight

to medical researchers as they attempt to identify sites associated with particular pathologies.

Bibliography

Jun Adachi and Masami Hasegawa. Improved dating of the human/chimpanzee separation in the mitochondrial dna tree: heterogeneity among amino acid sites. *Journal of Molecular Evolution*, 40(6):622–628, 1995.

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

Elizabeth S Allman, Sonia Petrovic, John A Rhodes, and Seth Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(3):710–722, 2011.

Guy Baele, Jeroen Raes, Yves Van de Peer, and Stijn Vansteelandt. An improved statistical method for detecting heterotachy in nucleotide sequences. *Molecular Biology and Evolution*, 23(7):1397–1405, 2006.

Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

Peter Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the Archaeological and Historical Sciences*, 1971.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Belinda SW Chang and Dana L Campbell. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Molecular Biology and Evolution*, 17(8):1220–1231, 2000.

Gary A Churchill, A Von Haeseler, and William C Navidi. Sample size for a phylogenetic inference. *Molecular Biology and Evolution*, 9(4):753–769, 1992.

Charles Darwin. *The origin of species*. John Murray, 1859.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

Anthony WF Edwards and Cavalli LL Sforza. The reconstruction of evolution. *Heredity*, 18, 1963.

Joseph Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3): 240–249, 1973.

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

Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

Joseph Felsenstein. PHYLIP version 3.6 a3. 2002.

Joseph Felsenstein. *Inferring phylogenies*. Sinauer Associates Sunderland, 2004.

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

Walter M Fitch and Emanuel Margoliash. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. *Biochemical Genetics*, 1(1):65–71, 1967.

Walter M Fitch and Etan Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4(5):579–593, 1970.

Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

Sudhindra R Gadagkar and Sudhir Kumar. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Molecular Biology and Evolution*, 22(11):2139–2141, 2005.

Nicolas Galtier. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, 18(5):866–873, 2001.

Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17 (6):333–351, 2016.

Xun Gu, Yun-Xin Fu, and Wen-Hsiung Li. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution*, 12(4):546–557, 1995.

Stéphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5): 696–704, 2003.

EF Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, pages 44–77, 1971.

Masami Hasegawa and Tetsuo Hashimoto. Ribosomal RNA trees misleading? *Nature*, 361(6407):23, 1993.

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(2):160–174, 1985.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Michael D Hendy and David Penny. A framework for the quantitative study of evolutionary trees. *Systematic Biology*, 38(4):297–309, 1989.

Simon Y Ho and Lars Jermiin. Tracing the decay of the historical signal in biological sequence data. *Systematic Biology*, 53(4):623–37, Aug 2004. doi: 10.1080/10635150490503035.

John P Huelsenbeck. Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution*, 19(5):698–707, 2002.

Vivek Jayaswal, Thomas KF Wong, John Robinson, Leon Poladian, and Lars S Jermiin. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Systematic Biology*, 63(5):726–742, 2014.

Lars Jermiin, Simon Y Ho, Faisal Ababneh, John Robinson, and Anthony W Larkum. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology*, 53(4):638–43, Aug 2004. doi: 10.1080/10635150490468648.

Thomas Jukes and Charles Cantor. Evolution of protein molecules. In Munro H.N. *Mammalian Protein Metabolism*, pages 21–123, New York: Academic Press, 1969.

Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

Motoo Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences*, 78(1):454–458, 1981.

Bryan Kolaczkowski and Joseph W Thornton. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431(7011): 980–984, 2004.

Mary K Kuhner and Joseph Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468, 1994.

Cecilia Lanave, Giuliano Preparata, Cecilia Sacone, and Gabriella Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1):86–93, 1984.

Nicolas Lartillot and Hervé Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109, 2004.

Pietro Lio and Nick Goldman. Models of molecular evolution and phylogeny. *Genome Research*, 8(12):1233–1244, 1998.

Peter Lockhart, Phil Novis, Brook G Milligan, Jamie Riden, Andrew Rambaut, and Tony Larkum. Heterotachy and tree building: a case study with plastids and eubacteria. *Molecular Biology and Evolution*, 23(1):40–45, 2006.

Peter J Lockhart, Michael A Steel, Michael D Hendy, and David Penny. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, 11(4):605–612, 1994.

Philippe Lopez, Didier Casane, and Hervé Philippe. Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, 19(1):1–7, 2002.

Thomas Madden. *The BLAST sequence analysis tool*. National Center for Biotechnology Information (US), 2013.

Frederick A Matsen and Mike Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 56(5):767–775, 2007.

Allan M Maxam and Walter Gilbert. A new method for sequencing dna. *Proceedings of the National Academy of Sciences*, 74(2):560–564, 1977.

Andrew Meade and Mark Pagel. A phylogenetic mixture model for heterotachy. In *Evolutionary Biology from Concept to Application*, pages 29–41. Springer, 2008.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Kary B Mullis and Fred A Faloona. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, 155:335–350, 1987.

SB Needleman. Needleman-Wunsch algorithm for sequence similarity searches. *Journal of Molecular Biology*, 48:443–453, 1970.

Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2015.

Mark Pagel and Andrew Meade. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4):571–581, 2004.

Mark Pagel and Andrew Meade. BayesPhylogenies version 1.0. 2006.

Hervé Philippe, Yan Zhou, Henner Brinkmann, Nicolas Rodrigue, and Frédéric Delsuc. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, 5(1):50, 2005.

Tal Pupko, Dorothée Huchon, Ying Cao, Norihiro Okada, and Masami Hasegawa. Combining multiple data sets in a likelihood analysis: which models are the best? *Molecular Biology and Evolution*, 19(12):2294–2307, 2002.

Andrew Rambaut and Nicholas C Grassly. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences: CABIOS*, 13(3):235–238, 1997.

Jaxk H Reeves. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution*, 35(1): 17–31, 1992.

David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, 1981.

Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12): 5463–5467, 1977.

Heiko A Schmidt, Korbinian Strimmer, Martin Vingron, and Arndt von Haeseler. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3):502–504, 2002.

C. Semple and A. Steel. *Phylogenetics*. Oxford Lecture Series in Mathematics and Its Applications, 24. Oxford University Press on Demand, 2003. ISBN 9780198509424.

Liat Shavit Grievink, David Penny, Mike D Hendy, and Barbara R Holland. LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. *BMC Evolutionary Biology*, 8:317, 2008.

Liat Shavit Grievink, David Penny, Michael D Hendy, and Barbara R Holland. Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree. *Systematic Biology*, 59(3):288–97, May 2010.

Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

Matthew Spencer, Edward Susko, and Andrew J Roger. Likelihood, parsimony, and heterogeneous evolution. *Molecular Biology and Evolution*, 22(5):1161–1164, 2005.

Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

Alexandros Stamatakis, Thomas Ludwig, and Harald Meier. RAxML-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463, 2005.

Mike Steel. Should phylogenetic models be trying to fit an elephant? *Trends in Genetics*, 21(6):307–309, 2005.

David L Swofford, Gary J Olsen, Peter J Waddell, and David M Hillis. *Molecular Systematics, 2nd Edition*. Sinauer Associates, Inc., 1996.

Koichiro Tamura and Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526, 1993.

Rosa Tarrío, Francisco Rodríguez-Trelles, and Francisco J Ayala. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the drosophilidae. *Molecular Biology and Evolution*, 18(8):1464–1473, 2001.

Chris Tuffley and Mike Steel. Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, 147(1):63–91, 1998.

Le Sy Vinh and Arndt von Haeseler. IQPNNI: moving fast through tree space and stopping in time. *Molecular Biology and Evolution*, 21(8):1565–1571, 2004.

Peter J Waddell and David Penny. Evolutionary trees of apes and humans from dna sequences. *Handbook of Symbolic Evolution*, pages 53–73, 1996.

Huai-Chun Wang, Matthew Spencer, Edward Susko, and Andrew J Roger. Testing for covarion-like evolution in protein sequences. *Molecular Biology and Evolution*, 24(1):294–305, 2007.

James D Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

Jihua Wu and Edward Susko. General heterotachy and distance method adjustments. *Molecular Biology and Evolution*, 26(12):2689–2697, 2009.

Jihua Wu and Edward Susko. A test for heterotachy using multiple pairs of sequences. *Molecular Biology and Evolution*, 28(5):1661–1673, 2011.

Ziheng Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6): 1396–1401, 1993.

Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, 1994.

Ziheng Yang. Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution*, 42(5):587–596, 1996.

Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

Udny Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, containing papers of a biological character*, 213:21–87, 1925.

Harold H Zakon, Ying Lu, Derrick J Zwickl, and David M Hillis. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3675–3680, 2006.

KA Zaretskii. Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Matematicheskikh Nauk*, 20(6):90–92, 1965.