Novel Mathematical Aspects of Phylogenetic Estimation

A thesis submitted by

Mareike Fischer

in partial fulfillment of the requirements for the degree

Doctor of Philosophy

at the University of Canterbury

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Examining Committee

Prof. Mike Steel: Internal Examiner

Prof. David Penny: External (NZ) Examiner

Prof. Arndt von Haeseler: External (overseas) Examiner

Abstract

In evolutionary biology, genetic sequences carry with them a trace of the underlying tree that describes their evolution from a common ancestral sequence. Inferring this underlying tree is challenging. We investigate some curious cases in which different methods like Maximum Parsimony, Maximum Likelihood and distance-based methods lead to different trees. Moreover, we state that in some cases, ancestral sequences can be more reliably reconstructed when some of the leaves of the tree are ignored – even if these leaves are close to the root. While all these findings show problems inherent to either the assumed model or the applied method, sometimes an inaccurate tree reconstruction is simply due to insufficient data. This is particularly problematic when a rapid divergence event occurred in the distant past. We analyze an idealized form of this problem and determine a tight lower bound on the growth rate for the sequence length required to resolve the tree (independent of any particular branch length). Finally, we investigate the problem of intermediates in the fossil record. The extent of 'gaps' (missing transitional stages) has been used to argue against gradual evolution from a common ancestor. We take an analytical approach and demonstrate why, under certain sampling conditions, we may not expect intermediates to be found.

Acknowledgements

I take this opportunity to thank my supervisor Mike Steel for his support over the last three years. By explaining relevant concepts and pointing out various open problems he made it easy for me to access the area of phylogenetics, which was new to me when I first came to New Zealand. Moreover, he always took the time to discuss the progress of my work, such that any problems at which I got stuck could be sorted out quickly. Most importantly, he encouraged me right from the beginning to present my results at various conferences, where I also got to meet many researchers from a variety of different areas. These experiences have definitely contributed to the value the last three years had both for my personal as well as my professional life.

Additionally, I want to thank my other two co-authors Hans-Jürgen Bandelt and Bhalchandra Thatte. With both I had many fruitful discussions, which gave rise to various mathematical ideas. Moreover, Bhalchandra often took the time to look at my other projects and to give me helpful comments on them. Also, I thank Wolfgang Fischl for the great collaboration concerning the relevance of 'misleading sequences' in practice.

I wish to thank my co-supervisors Barbara Holland and Charles Semple, as well as my external PhD examiners Arndt von Haeseler and David Penny. I will remember all my visits to Palmerston North, during which I had many opportunities to present my own work as well as to be introduced to open questions on which Barbara and David were working, as great times – in spite of the weather in Palmy, which tried very hard to convince me otherwise. The same applies to my visit to Arndt's working group in Vienna, where despite the unbearable heat many fruitful mathematical discussions took place. I also owe many thanks to Marta Casanellas, who invited me to her working

group in Barcelona and introduced me to Elizabeth Allman and John Rhodes. All three of them, Marta, Elizabeth and John, showed a lot of interest in my work and made many useful suggestions. I also thank the Alfréd Rényi Institute of Mathematics in Budapest for enabling me to stay there and collaborate with Bhalchandra, but also for giving me the opportunity to discuss my work with Péter Erdös and Lászlo Skékeley. This was very helpful.

During the past three years, I attended about 15 conferences and meetings and I additionally visited four different working groups. At all these occasions, I had useful mathematical discussions with many different people – so, naturally, it is impossible to thank all of them. But I do wish to explicitly thank Elliott Sober, who brought the problem of intermediate fossils to my attention, as well as Matt Philips, who told me about the importance of clades in biological research, and Oliver Will, who gave me some additional information on the fossil record.

All of the above mentioned people made my time in New Zealand worthwhile and unforgettable. But there are several others who made it even more invaluable: Beáta Faller, Ramona Schmid and Simone Linz – I thank you for all the coffees, walks, trips and thoughts we shared. I remember that I once told you that our friendship developed in time-lapse speed, and I still think that this is the reason why I felt at home in New Zealand so quickly. Thanks, Mädels, for everything!

I also wish to thank all other postgraduate students and postdocs I met in New Zealand and who brightened up many of my days. To name just a few of them: Alethea, Tanja, Klaas, Erick, Scott, Daniel, Hannes, Michael, Iris, Maarten and Meghan. Moreover, thanks to all my fellow Stats tutors, most importantly Irene David, Hilary Seddon, Anna MacDonald, Miriam Hodge and, of course, the best colleague and officemate ever: Kyoko Fukuda. Thanks for tolerating all my boxes for the last couple of months! I also absolutely have to thank Rachael, Peter, Nicole, Joshua, Michael, Shannon and Liene for the very best night during the last three years – as well as their support on the worst morning ever, which directly followed that night.

Despite all these great people, my time in New Zealand would not have been the same without the backup and support of my family and friends back home in Europe. First of all, thanks to everyone who visited me here: Eva and Andreas, I knew you would like the cave! Christian and Taiga, I will never forget our evening at that very peculiar bar my sister suggested. And Julia, I am glad you could come here after all – it has been great to see you growing up, and it was even better to be able to show you one of the most beautiful countries in the world. Thanks also to you, Britta, Kata and Oli – I know you wanted to visit me, but it did not work out in the end. I am thankful for all your support over the last decade. Particular thanks also to Thomas and Hannah – I lack the words to describe how glad I am to have friends like you.

Most importantly, I want to thank my family for their support and, particularly my parents, for coping with the long journey just to visit me here – I know that for you these long flights are even worse than for me. My sister Janina has always been invaluable – she is inspiring and funny, and we are the best team ever (even though some of the things we take up often just cause incredulous head shaking of our parents and others). Thanks for being the only chemistry student in the world voluntarily proofreading a maths PhD thesis, and even more thanks for coming to New Zealand twice and joining every weird activity this country has to offer – the photo taken during our paraflight is still my favorite.

I also wish to thank the New Zealand Mathematical Society for awarding me the Aitken Prize. And, most importantly, I am much obliged to the Allan Wilson Centre for Molecular Ecology and Evolution and the Marsden Fund for financing my research. Without their support, all this would not have been possible.

Thanks to all of you and to this country, Aotearoa, for the best 3 years of my life.

Mareike Fischer

Contents

Chapter 1

Introduction

All sciences, arts and religions are branches of the same tree.

Albert Einstein

Ever since Charles Darwin published his work ‘The Origin of Species’, originally published as ‘On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life’ [7], in which the idea of a phylogenetic tree as underlying concept of inter-species relationships was first introduced, the search for the ‘true tree’ and the ‘most recent common ancestor’ has never lost its appeal. Even so, the means and methods used for this purpose have changed over the decades: while Darwin strongly depended on morphological (phenotype) data, most conclusions drawn nowadays are based on genetic (genotype) data. Thus, the enthusiasm for evolutionary research in general and phylogenetics in particular was strongly enlivened by the sequencing of the human genome in 2003.

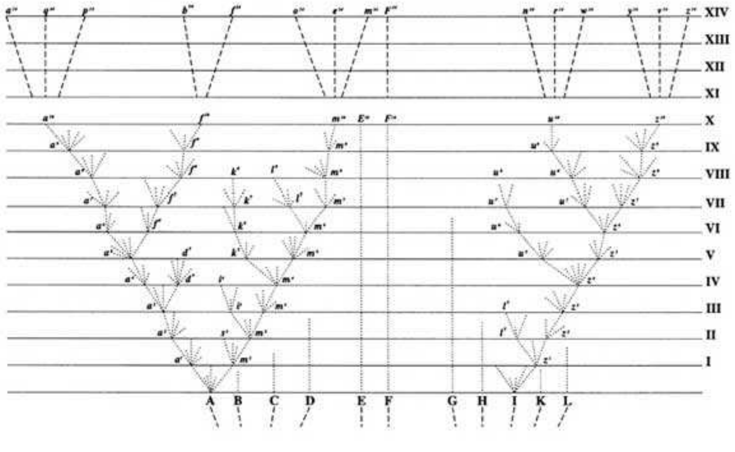


Figure 1.1: The only illustration in Darwin’s book ‘The Origin of Species’ was a simple sketch of a set of rooted phylogenetic trees.

Due to improved methods and more efficient DNA sequencing machines, the amount of available DNA sequence data is ever-growing. Unsurprisingly, this makes phylogenetic models and methods, which help interpret the data, more and more essential. In this thesis, we present the most important of these models and methods along with some innovative results. We first present some general preliminaries in Chapter 2, in which the methods and models used in this thesis are introduced along with some required terminology. Then, we analyze these methods more in-depth. First, we show in Chapter 3 that there are cases in which the Fitch algorithm for Maximum Parsimony (MP), which is one of the most frequently used phylogenetic methods, gives more accurate estimates of the ancestral root state when some taxa, i.e., leaves of the underlying tree, are ignored – even if these taxa are close to the root. In Chapters 4, 5 and 6, we analyze the performance and accuracy of different tree reconstruction methods: first, in Chapter 4, we show that for all numbers of taxa ternary sequences can be constructed for which the perfect MP tree does not coincide with the distance-wise perfect tree. We show how to construct such 'misleading' sequences and we present a full characterization of them for the 4-taxa case. In Chapter 5, we show that MP and another tree inference method, Maximum Likelihood (ML), may choose conflicting trees for certain biologically relevant modifications of the so-called 'no common mechanism' N_r-model, even though they are known to be equivalent without these modifications. In particular, we show that the equivalence fails when a molecular clock is assumed or when nucleotide substitution probabilities are bounded by an upper bound less than that assumed by the N_r-model (under 'no common mechanism'). We will also explain to what extent the latter inequivalence is related to the misleading sequences as introduced in Chapter 4. Then, in Chapter 6, we analyze the general accuracy of tree reconstruction methods for a worst-case scenario: when speciation events occurred in rapid succession (leading to short branch lengths) in the distant past (leading to the large branch lengths for the incident edges), this makes the underlying tree particularly difficult to reconstruct. We analyze a somewhat idealized 4-taxa case and provide tight lower bounds (up to a constant factor) on the sequence length needed to reconstruct the true tree with high probability.

Finally, in Chapter 7 we introduce two simple stochastic models (one based on the amount of evolutionary history and another one based on clades of the underlying tree) for the estimation of the degree of relatedness of fossils from different times. We show that there are cases in which fossils from an intermediate time cannot be expected to be morphological intermediates – and thus will cause alleged 'gaps' in the fossil record. This is a novel approach to explain anomalies in the fossil record, as traditionally only reasons like the rare conditions needed for fossilization and the discontinuous fossil discovery (as opposed to continuous evolutionary processes) are specified.

Chapter	**Authors**	**Journal**	**Reference**	**Status**
3	Fischer, Thatte	J. Theo. Biol.	[FT09a]	submitted
4.1	Bandelt, Fischer	Syst. Biol.	[BF08]	published
5	Fischer, Thatte	Bull. Math. Biol.	[FT09b]	accepted s.t. rev.
6	Fischer, Steel	J. Theo. Biol.	[FS09]	published
7.2	Fischer, Steel	Evol. Bioinf. Onl.	[FS08]	published

Table 1: List of publications resulting from this thesis. Mareike Fischer is lead author or contributed equally to all the publications listed here.

Three articles resulting from this thesis have already been published and two more are currently under review. Table 1 gives an overview of these publications and the corresponding chapters in this thesis. Due to the interdisciplinary nature of this work, much collaboration was involved – in total there are four authors on the publications resulting from this thesis. It should be noted that I was lead author or equal contributor on all these papers. Various chapters draw heavily on these publications, however the work presented here is either completely my own work or work to which I contributed substantially. The only exception to this are some of the findings of Fischl quoted in Chapter 4.3 where I only analyzed the particular alignment used as an example, but where other results presented by Fischl in an unpublished paper [13], such as Table 5, were cited to complement our common results.

Chapter 2

Preliminaries

2.1 Basic Definitions and Notations

We start with some basic definitions and general notations which will be used throughout this thesis. Notations or definitions which are used only in one chapter will be introduced where needed.

An *unrooted phylogenetic X-tree* is a tree $\mathcal{T} = (V(\mathcal{T}), E(\mathcal{T}))$ with vertices $V(\mathcal{T})$ and edges $E(\mathcal{T})$ on a leaf set $X = \{1, \dots, n\} \subseteq V(\mathcal{T})$ with no vertices of degree 2. If $\mathcal{T}$ is binary, all internal nodes have degree 3. Note that a binary phylogenetic tree on X has n leaves (end nodes) that are labeled by the n taxa in X and $n-2$ interior nodes that are unlabeled. The definition of *rooted phylogenetic X-trees* is analogous except that there one node of degree 2, namely the root, is allowed. A *character* f is a function $f : X \to \mathcal{C}$ for some set $\mathcal{C} := \{c_1, c_2, c_3, \dots, c_r\}$ of r *character states* ($r \in \mathbb{N}$) and is often denoted by $f = \alpha_1 \dots \alpha_n$, where $\alpha_i \in \mathcal{C}$ for all $i = 1, \dots, n$, and where $f(i) = \alpha_i$. A character is said to be *informative (with respect to parsimony)* if at least two distinct character states occur more than once in X, otherwise it is called *uninformative* or *non-informative.* Moreover, we say that each character induces a *partition* of X: if, for instance, $f = \alpha\beta\gamma\dots\gamma$, f induces the partition $1|2|3\dots n$, or, more formally, the partition of X into the sets $\{1\}$, $\{2\}$ and $\{3, \dots, n\}$. If f is binary, the induced partition is called an *X-split* or *split* for short. Partitions and splits of X are said to be *trivial* exactly when they are induced by non-informative characters on X. Throughout this thesis, we will use the following notation (see [38]): Let $\Sigma(X)$ be the set of all X-splits $\sigma = A|B = B|A$ of a set X of n taxa, i.e. $\Sigma(X) = \{\sigma = A|B = B|A : A \cup B = X, A \cap B = \emptyset\}$. Furthermore, let $\Sigma^*(X)$ be the set of non-trivial X-splits only (i.e.

the set of X-splits for which $|A|, |B| > 1$). Analogously, let $\Sigma(\mathcal{T})$ denote the X-splits induced by the edges of $\mathcal{T}$ (where $\mathcal{T}$ is an binary phylogenetic X-tree), and $\Sigma^*(\mathcal{T})$ the non-trivial ones.

An *extension* of a character f to $V(\mathcal{T})$ is a map $g : V(\mathcal{T}) \rightarrow \mathcal{C}$ such that the restriction of g to $X \subseteq V(\mathcal{T})$ is f. For such an extension g of f, we denote by $l_{\mathcal{T}}(g)$ the number of edges $e = (u, v)$ in $\mathcal{T}$ on which a *substitution* occurs, i.e., where $g(u) \neq g(v)$.

Often, we analyze a sequence of characters rather than a single character. If a sequence S consists of characters $f_1, \ldots, f_k$ for some integer $k \geq 1$, we denote this by $S = f_1 \ldots f_k$. If two sequences $S = f_1 \ldots f_k$ and $\tilde{S} = \tilde{f}_1 \ldots \tilde{f}_l$ are concatenated, the concatenation is denoted by sequence $\hat{S} = S\tilde{S} = f_1 \ldots f_k\tilde{f}_1 \ldots \tilde{f}_l$.

2.2 The N_r-model

We now introduce the nucleotide substitution model on which most of our results are based. Unless stated otherwise, our results will correspond to the so-called *N_r-model*, also known as r-state symmetric model, Neyman r-state model or Cavender-Farris-Neyman model. For $r = 4$, the N_r-model is also known as Jukes-Cantor model [20].

Let $\mathcal{T}$ be a phylogenetic X-tree, and let $c_1, \ldots, c_r$ be r distinct character states. If $\mathcal{T}$ is not rooted, we may arbitrarily choose one of the leaves to be the root (note that because of the so-called reversibility of the N_r-model, the actual root position does not matter, see [11]). Then, the N_r-model assumes a uniform distribution of states at the root, and assumes equal rates of substitutions between any two distinct character states [31]. For any edge $e \in E(\mathcal{T})$, let p_e denote the probability that a substitution of a character state c_i by another character state c_j occurs on edge e for $c_i \neq c_j$. Furthermore, let q_e denote the probability that no substitution occurs on edge e. Then, in the N_r-model we have $0 \leq p_e \leq \frac{1}{r}$ for all $e \in E(\mathcal{T})$ and $(r-1)p_e + q_e = 1$. Furthermore, the N_r-model assumes that substitutions occur independently on different edges.

Although it is the simplest non-trivial Markov process on a tree, the N_r-model allows for an exact analysis of many scenarios. Moreover, stochastic results for this model typically extend to more general finite-state models where an exact analysis is usually more complex [28].

Note that for the interpretation of character sequences $S = f_1 \dots f_m$ as opposed to single characters, we often make the additional model assumption of *'no common mechanism'*. This means that substitution probabilities on edges of the underlying tree $\mathcal{T}$ may differ for each character in the sequence without any correlation between the sites. That is, we suppose that for each character f_i in the sequence and for each edge e of the tree, there is a parameter $p_{e,i}$ that gives the substitution probability for f_i on edge e, and that the parameters $p_{e,i}$ are all independent. For all $i = 1, \dots, m$, we will denote by $\bar{p}_i$ the vector of substitution probabilities $p_{e,i}$ assigned to the edges e of $\mathcal{T}$. So the difference between the N_r-model and the N_r-model with the additional assumption of no common mechanism is that for a sequence $S = f_1 \dots f_m$ of characters, both assume that all characters evolved independently, but when there is a common mechanism, the distributions of the characters additionally have to be identical (i.e., $p_{e,i} = p_e$ for all i and some fixed value p_e). So whenever characters evolve 'i.i.d.', there is a common mechanism.

We will introduce other models, such as the Random Cluster Model, in Chapters 6.4 and 6.5.

2.3 Phylogenetic Methods

2.3.1 Maximum Parsimony

Maximum Parsimony, or *MP* for short, is one of the most frequently used tree inference methods, and it is not based on a specific nucleotide substitution model. It selects the tree that requires the smallest number of substitutions to extend the sequences at the tips of the tree to (ancestral) sequences at all the interior vertices of the tree. Its simplest form is the so-called *Fitch parsimony* [14], which will be considered in this thesis unless stated otherwise. Given a character f and a tree $\mathcal{T}$, MP assigns sets of character states to all internal nodes such that they provide possible choices to achieve the minimum number of changes to realize f. Along the way, the so-called *parsimony score* or *MP-score* $l_{\mathcal{T}}(f)$, which denotes the minimum number of changes required to realize f on $\mathcal{T}$, is calculated. It is obtained by minimizing $l_{\mathcal{T}}(g)$ over all possible extensions g of f. The parsimony score of a sequence of characters $S := f_1 f_2 \dots f_m$ is given by $l_{\mathcal{T}}(S) = \sum\limits_{i=1}^{m} l_{\mathcal{T}}(f_i)$. In order to use MP for tree inference, all possible trees have to be analyzed. A (not necessarily unique) tree with minimal parsimony score is called *Maximum Parsimony tree*, or *MP-tree* for short. Note that a parsimoniously non-informative character has the same MP-score on all trees, whereas informative characters have different scores on at least two trees. If an r-state character f has parsimony score $r-1$ on a tree $\mathcal{T}$, it is said to be *convex* or *homoplasy-free* on $\mathcal{T}$. A sequence of characters is called convex or homoplasy-free on $\mathcal{T}$ when all its characters have this property. If a sequence is convex on a tree $\mathcal{T}$, $\mathcal{T}$ is called its *perfect phylogeny.*

Note that when MP is used for tree inference in Chapters 4, 5 and 6, unless stated otherwise, the term *most parsimonious extension* g of a character f on a tree $\mathcal{T}$ refers to *any* extension g for which $l_{\mathcal{T}}(g)$ is minimal. The minimal MP-score can be calculated with the Fitch algorithm given in [14]. However, in Chapter 3, where the *reconstruction accuracy* of MP concerning the ancestral root state is examined, we consider only those interior node labels and thus extensions which are suggested by the Fitch algorithm. It is important to state this as it is known that while the Fitch algorithm calculates the parsimony score and gives some most parsimonious extensions, it does not necessarily find *all* most parsimonious extensions (see [12], pp. 12–13 for details). For further background on MP, the reader can consult, for example, [12] or [38].

2.3.2 Maximum Likelihood

Now we define *Maximum Likelihood trees*, or for short *ML-trees.* As opposed to MP, ML is based on nucleotide substitution models and thus may give different results for different models. In this thesis, we assume the N_r-model when inferring ML-trees. Note that in the N_r-model, the parameter space is compact, which is why we can here define ML-trees using only maxima rather than suprema.

For any phylogenetic X-tree $\mathcal{T}$, any character f on the leaf set X and any vector $\bar{p}$ of substitution probabilities assigned to all edges of $\mathcal{T}$, we denote by $P(f|\mathcal{T},\bar{p})$ the likelihood of observing the character f on $\mathcal{T}$ for the given parameter values $\bar{p}$. The maximum likelihood of a character f on $\mathcal{T}$, denoted by $\max P(f|\mathcal{T})$, is then the maximum of $P(f|\mathcal{T},\bar{p})$ over the space of all $\bar{p}$, i.e., $\max_{\bar{p}} P(f|\mathcal{T},\bar{p})$.

When the likelihood of a character sequence S is calculated, we assume 'no common mechanism', i.e., if $p_{e,i}$ gives the substitution probability for f_i on edge e, we assume that the parameters $p_{e,i}$ are all independent. For a sequence $S := f_1 \dots f_m$ the likelihood is the product of the likelihoods of the individual characters, i.e., $P(S|\mathcal{T},\bar{p}_1,\dots,\bar{p}_m) = \prod\limits_{i=1}^{m} P(f_i|\mathcal{T},\bar{p}_i)$. Then, under no common mechanism, the maximum likelihood of S can be calculated as the product of the maximum likelihoods of the individual characters, i.e., $\max_{(\bar{p}_1,\dots,\bar{p}_m)} P(S|\mathcal{T},\bar{p}_1,\dots,\bar{p}_m) = \prod\limits_{i=1}^{m} \max_{\bar{p}_i} P(f_i|\mathcal{T},\bar{p}_i)$. Finally, the (not necessarily unique) *Maximum Likelihood tree*, or *ML-tree* for short, of S is a tree for which this value is maximal, i.e., $\operatorname{argmax}_{\mathcal{T}} \max_{(\bar{p}_1,\dots,\bar{p}_m)} P(S|\mathcal{T},\bar{p}_1,\dots,\bar{p}_m)$.

An algorithm for the explicit calculation of ML-trees is given by Felsenstein [11].

2.3.3 Distance-based Methods

Apart from various variations of MP and ML, the third group of most frequently used tree inference methods are *distance-based methods* such as, for instance, *Neighbor-Joining* [3]. In this thesis, we only consider the simplest kind of distances, namely the so-called non-normalized *Hamming distance*, which is also known as *mismatch distance*. In practice, these distances are normally used in a 'corrected' form to account for so-called hidden changes in sequence evolution (for instance, if both a sequence and one of its ancestral sequences have an A at a particular site, this does not mean that no change has occurred during the sequence evolution – e.g., the A could have changed to a G and back to an A. Since such changes are not observable when the two sequences are compared, such changes are called 'hidden'). However, all distance results of this thesis, in particular the misleading sequences as introduced in Chapter 4, are not affected by this correction – i.e., they remain valid even when the derived Hamming distances are corrected. Therefore, it is here sufficient to formally introduce the Hamming distance in its basic form.

From any finite sequence $S = f_1 f_2 \dots f_m$ of characters on X (as e.g. obtained from aligned DNA sequences) one derives the Hamming (mismatch) distance function $d = d_S$ as follows:

$$d_S(x, y) := |\{i : f_i(x) \neq f_i(y)\}| \text{ for taxa } x, y \in X.$$

Note that for $S = f$, i.e., a single character f, d_f only depends on the partitions associated with f, which is why we may index d by the corresponding partitions instead. So in the case of a k-ary character f, which partitions the taxon set into k parts $A_1, \dots, A_k$, we also denote the mismatch distance as $d_{A_1|\dots|A_k} := d_f$.

Regarding phylogenetic research, an important property of any distance function is *treelikeness*: an arbitrary function $d : X \times X \to \mathbb{R}$, where X is a leaf set, is called a *tree metric* when it satisfies the so-called *4-point condition.* As in [38], Definition 7.1.5, we define this as follows: A function $d : X \times X \to \mathbb{R}$, where X is a leaf set, for which $d(x, x) = 0$ and $d(x, y) = d(y, x)$ for all $x, y \in X$, satisfies the 4-point condition if, for every four (not necessarily distinct) elements $w, x, y, z \in X$, two of the sums $d(w, x) + d(y, z)$, $d(w, y) + d(x, z)$ and $d(w, z) + d(x, y)$ are equal and not less than the remaining one. In the following, we will also often refer to tree metrics as *treelike* or *additive* distances. A tree metric $d_{A|B}$, i.e., a metric induced by an X-split, will be referred to as *split metric.*Note that if the Hamming distances are treelike on a binary (and thus fully resolved) phylogenetic tree $\mathcal{T}$, they are treelike *only* on $\mathcal{T}$ (see [4]).

A tree metric $d : X \times X \to \mathbb{R}$ for an X-tree $\mathcal{T}$ is called *ultrametric* if, for every three distinct elements $x, y, z \in X$, two of the distances $d(x, y)$, $d(x, z)$ and $d(y, z)$ are equal and not less than the third. By Theorem 7.2.5 in [38], this is the case if and only if $d : X \times X \to \mathbb{R}$ can be extended to $d : V(\mathcal{T}) \cup \{\rho\} \times V(\mathcal{T}) \cup \{\rho\} \to \mathbb{R}$ such that for some distinguished point ρ (which either is in $V(\mathcal{T})$ or is a newly introduced vertex of degree 2), we have $d(x, \rho) = d(y, \rho)$ for all $x, y \in X$. In this case, we say the distances are *clocklike* or conform to a *molecular clock* with *root* ρ. This is why in this thesis, we often use the term 'clocklike' as a synonym for 'ultrametric'. Note that this distance-based definition of molecular clocks differs from the purely probability-based definition which will be introduced and used in Chapters 3.3 and 5.3. There, rather than the distances from all leaves to a distinguished root being equal, the probabilities of a change along the paths from the root to all leaves are supposed to be equal.

Chapter 3

Maximum Parsimony on Subsets of Taxa

> Learning to ignore things is one of the great paths to inner peace.
>
> Robert J. Sawyer

In a recent study [22], a likelihood analysis of Fitch's maximum parsimony method [14] for the reconstruction of the ancestral state at the root was conducted. It was shown that in a rooted phylogenetic tree if one leaf is closer to the root than all the other leaves, then the character state at this leaf may sometimes be a more accurate guess of the ancestral state than the ancestral state constructed by MP applied to all taxa. The authors also provided an example of a phylogenetic tree for which MP for the reconstruction of the root state works more reliably on a subset of taxa closer to the root than on all taxa.

Generally the root state is more likely to be conserved on taxa that are nearer to the root than on taxa that are further away. Therefore, it is not surprising that on some trees the root state can be more reliably estimated by looking at only taxa nearer to the root. But can the reconstruction accuracy of MP improve when a taxon or a subset of taxa close to the root is ignored? In Chapter 3.2, we present a surprising example of a tree on which MP on a subset of taxa is more likely to reconstruct the correct ancestral state. In our example, the reconstruction accuracy improves when we ignore a taxon close to the root from our analysis. Moreover, the ignored taxon may be arbitrarily close to the root compared to the taxa that are not ignored. On the other hand, in Chapter 3.3, we show that under a molecular clock, considering a single taxon is never better than considering all taxa for the purpose of ancestral state reconstruction. Our analysis resolves a conjecture of Li, Steel and Zhang [22] for the

2-state case. They conjectured that under a molecular clock, MP on all taxa is expected to generally perform at least as good (in the sense of the reconstruction accuracy) as reconstructing the ancestral state based on the character state at a single taxon. In Section 3.3, we make the conjecture precise and answer it affirmatively for the case of the 2-state symmetric model.

3.1 The MP Reconstruction Accuracy of the Root State

We first state some general properties of the (Fitch) MP reconstruction accuracy of the root state, before we provide an example for a misleading taxon. In the following, let $\mathcal{T}$ be a rooted binary phylogenetic X-tree with root ρ. We assume that each vertex in $\mathcal{T}$ takes one of the two states α and β. The states evolve from the root state under the N_2-model as described in Chapter 2.2.

In this chapter we analyze the probability that MP, in the form of the Fitch algorithm, applied to a subset of the set of taxa correctly estimates the true state at the root. Suppose that Y is a subset of X, i.e., $Y \subseteq X$. Y induces a subtree $\mathcal{T}_Y$, rooted at a vertex y. Here, y is the most recent common ancestor of vertices in Y. It is possible that $y = \rho$. Let f_Y denote the restriction of a binary character f to Y. MP assigns states α or β to all internal nodes (including the root ρ) so that the total number of substitutions is minimized. Such an assignment is not necessarily unique: MP computes a set S_z of possible states at each internal vertex z, so that each most parsimonious assignment must assign one of the states in S_z to the vertex z. When MP is applied to a binary character f, we have either $S_\rho = \{\alpha\}$ or $S_\rho = \{\beta\}$ or $S_\rho = \{\alpha, \beta\}$ at the root ρ. If S_ρ is either $\{\alpha\}$ or $\{\beta\}$, then we say that MP *unambiguously reconstructs* the root state; otherwise (when $S_\rho = \{\alpha, \beta\}$) we say that MP *ambiguously reconstructs* the root state.

The MP algorithm may also be applied to f_Y on the subtree $\mathcal{T}_Y$. It returns a state set $S_y = \{\alpha\}$ or $S_y = \{\beta\}$ or $S_y = \{\alpha, \beta\}$ for the root y of $\mathcal{T}_Y$. We will denote by $\mathtt{MP}(f_Y, \mathcal{T}_Y)$ the set of states which MP assigns to the root y of the subtree $\mathcal{T}_Y$ for character f_Y.

In the following, we will denote by $\mathtt{MP}(f, T)$ the set of character states chosen by Fitch's maximum parsimony algorithm as possible root states when applied to a character f on a tree T.

Li, Steel and Zhang defined the *unambiguous reconstruction accuracy* $UA(Y)$ and the *ambiguous reconstruction accuracy* $AA(Y)$ as follows:

$$\begin{aligned} UA(Y) &:= P\left(\mathtt{MP}(f_Y, \mathcal{T}_Y) = \{\alpha\} | \rho = \alpha\right), \\ AA(Y) &:= P\left(\mathtt{MP}(f_Y, \mathcal{T}_Y) = \{\alpha, \beta\} | \rho = \alpha\right). \end{aligned}$$

In other words, $UA(Y)$ is the probability that the root state α evolves to a character f for which MP on Y assigns the state set $\{\alpha\}$ to the root y of $\mathcal{T}_Y$. Similarly, $AA(Y)$ is the probability that the root state α evolves to a character f for which MP on Y assigns the state set $\{\alpha, \beta\}$ to the root y of $\mathcal{T}_Y$.

Furthermore, they defined the *reconstruction accuracy* as

$$RA(Y) = UA(Y) + \frac{1}{2}AA(Y), \tag{1}$$

where the second term indicates that when MP reconstructs the state at the root ambiguously, we select one of the states with equal probability.

Note that MP, when applied to Y, estimates a state at the root vertex y of the subtree $\mathcal{T}_Y$ induced by Y. Since it is possible that the root y of $\mathcal{T}_Y$ is different from the root ρ of $\mathcal{T}$, we define the reconstructed state at y to be the estimate of the state at the root based on the subset Y of taxa.

Li, Steel and Zhang gave an example of a tree for which the reconstruction accuracy of MP on a proper subset of taxa is higher than the reconstruction accuracy of MP on all taxa, i.e., $RA(Y) > RA(X)$ for some proper subset Y of X. But their example requires that the taxa in Y are closer to the root than the other taxa, i.e., that the probability of a substitution from the root to any taxon in Y is smaller than the probability of a substitution from the root to the other taxa. The example that we present in the following does not require any taxa to be closer to the root. On the contrary, our example shows that a *misleading taxon or taxa* (a taxon or taxa that have an adverse effect on the reconstruction accuracy) may be arbitrarily close to the root.

3.2 An Example of a Misleading Taxon

The main result of this chapter is the following theorem which shows that there are trees on which the reconstruction accuracy improves when a taxon close to the root is ignored in an MP based ancestral state reconstruction. Moreover, such a misleading taxon may be arbitrarily close to the root. Note that we consider a taxon $x_1 \in X$ to be closer to the root than a taxon $x_2 \in X$ whenever the substitution probability from ρ to x_1 is smaller than that from ρ to x_2.

Theorem 3.1. *Let p_z be any real number such that $0 < p_z < \frac{1}{2}$, and assume the N_2-model. Then there exists a binary phylogenetic tree $\mathcal{T}$ on a leaf set X and rooted at ρ such that the following conditions are satisfied:*

1. *for some leaf z, the substitution probability from ρ to z is p_z;*
2. *$RA(X - \{z\}) > RA(X)$; and*
3. *for each leaf $v \neq z$, the substitution probability p_v from ρ to v is more than p_z, i.e., z is closer to the root than any other taxon.*

To prove the above theorem, we first need some notation and a lemma. Let y be a vertex in a binary phylogenetic tree $\mathcal{T}$, and let Y be the set of leaves below y. We associate three probabilities with Y as follows.

$$\begin{aligned} P_\alpha(Y) &:= P\left(\mathtt{MP}(f_Y, \mathcal{T}_Y) = \{\alpha\} | y = \alpha\right), \\ P_\beta(Y) &:= P\left(\mathtt{MP}(f_Y, \mathcal{T}_Y) = \{\beta\} | y = \alpha\right), \\ P_{\alpha\beta}(Y) &:= P\left(\mathtt{MP}(f_Y, \mathcal{T}_Y) = \{\alpha, \beta\} | y = \alpha\right). \end{aligned}$$

Let $\mathcal{T}_h$ be a balanced binary tree of depth h, i.e., with leaf set X such that $|X| = 2^h$. Suppose that the substitution probability on each edge of $\mathcal{T}_h$ is q. For this particular symmetric tree, we denote $P_\alpha(X)$, $P_\beta(X)$ and $P_{\alpha\beta}(X)$ by $P_\alpha(h,q)$, $P_\beta(h,q)$ and $P_{\alpha\beta}(h,q)$, respectively. The convergence properties of these probabilities (for $h \to \infty$ and for various values of q) have been studied in detail, [for example, 40; 48]. We state the following result on the convergence of $P_\alpha(h,q)$ that additionally provides a lower bound on $P_\alpha(h,q)$ which is independent of h.

Lemma 3.2 (Charleston and Steel, 1995, Yang, 2008)**.** *Let $\mathcal{T}_h$ be a binary balanced phylogenetic tree of depth $h \geq 2$. Let $q < \frac{1}{8}$ be the probability of a substitution on each edge of the tree under the N_2-model. Then $P_\alpha(h,q)$ approaches*

$$\frac{1}{2}\left(1 - \frac{2q}{1-2q} + \frac{\sqrt{(1-8q)(1-4q)}}{(1-2q)^2}\right)$$

from above as $h \to \infty$. Moreover, as q goes to 0, the above limiting value approaches 1.

We are now in a position to prove the above theorem.

Proof of Theorem 3.1. Let $\mathcal{T}$ be a phylogenetic tree rooted at ρ constructed as follows. The left subtree of $\mathcal{T}$ contains a single leaf z. The right subtree of $\mathcal{T}$ is $\mathcal{T}_Y$ with leaf set Y and root y. Therefore, the leaf set of $\mathcal{T}$ is $X = Y \cup \{z\}$. We choose $\mathcal{T}_Y$ to be a balanced binary tree of depth h and substitution probability q on each edge. Let the substitution probabilities on (ρ, z) and (ρ, y) be p_z and p_y, respectively, where p_z is any given real number such that $0 < p_z < \frac{1}{2}$. An illustration of these parameters is provided by Figure 3.1.

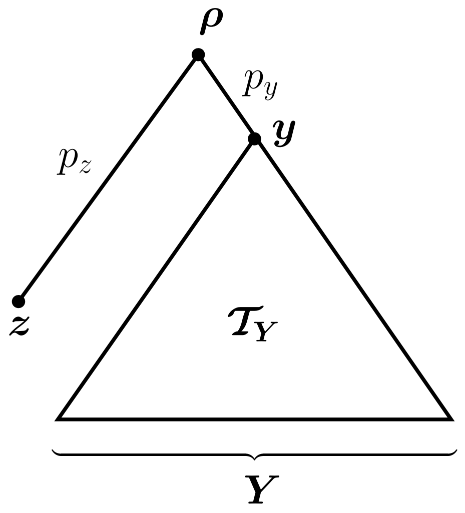


Figure 3.1: A tree on which MP is more accurate when applied to $Y \subset X$.

For the above tree, the reconstruction accuracy on X is given by

$$\begin{aligned} RA(X) &= (1-p_z)\left((1-p_y)P_\alpha(h,q) + p_y P_\beta(h,q) + P_{\alpha\beta}(h,q)\right) \\ &\quad +\frac{1}{2}\,p_z\left((1-p_y)P_\alpha(h,q) + p_y P_\beta(h,q)\right) \\ &\quad +\frac{1}{2}\,(1-p_z)\left(p_y P_\alpha(h,q) + (1-p_y)P_\beta(h,q)\right). \end{aligned}$$

The reconstruction accuracy on Y is given by

$$RA(Y) = (1-p_y)P_\alpha(h,q) + p_y P_\beta(h,q) + \frac{1}{2}\,P_{\alpha\beta}(h,q).$$

In order to satisfy $RA(Y) > RA(X)$, we therefore must have

$$(p_z - p_y)P_\alpha(h,q) > (1-2p_z)P_{\alpha\beta}(h,q) + (1-p_z-p_y)P_\beta(h,q). \qquad (2)$$

We now show that for any value of p_z, however small, the remaining substitution probabilities q and p_y and the depth h of $\mathcal{T}_Y$ can be chosen such that $RA(Y) > RA(X)$ (Condition 2 in Theorem 3.1), and for every vertex v in Y, the probability of a change of state from the root to v is more than p_z (Condition 3 in Theorem 3.1).

We express the third condition in Theorem 3.1 in a different form. Let $Q := 1-2q$, $P_z := 1-2p_z$ and $P_y := 1-2p_y$. Since the tree $\mathcal{T}_h$ is symmetric, the probability of a change of state from the root to any leaf v in Y is the same, and is given by $p_v = \frac{1-P_yQ^h}{2}$. Therefore, the third condition may now be written as $P_yQ^h < P_z$, or equivalently as

$$(1-2q)^h < \frac{1-2p_z}{1-2p_y}. \tag{3}$$

It follows from Lemma 3.2 that, for all $h \geq 2$, as q approaches 0, the left hand side of Equation (2) approaches $p_z - p_y$ and the right hand side approaches 0. Therefore, there is a real number ϵ such that $0 < \epsilon < \frac{1}{8}$, and whenever $q < \epsilon$, Equation (2) is satisfied. Now given a value of p_z, we first arbitrarily fix p_y such that $0 < p_y < p_z$, and then fix a value of $H := (1-2q)^h$ satisfying the constraint in Equation (3). We then choose h sufficiently large so that $q = \frac{1}{2}(1 - H^{\frac{1}{h}}) < \epsilon$ and the constraint given in Equation (2) is satisfied as well. This completes the proof. □

Note that when $q \geq \frac{1}{8}$, the sequence $P_\alpha(h,q)$ has quite different convergence properties than when $q < \frac{1}{8}$, and the bound provided by Lemma 3.2 does not apply, [see 40; 48, for details]. Therefore, our construction of a misleading taxon given in the proof of Theorem 3.1 strongly depends on q being sufficiently small.

3.3 A Single Taxon under a Molecular Clock

In this section, we consider binary characters on a binary phylogenetic tree $\mathcal{T}$ with leaf set X under a molecular clock and the N_2-model introduced earlier. Let p be the probability that a leaf is in a different state than the root. Therefore, if we were to guess the root state by looking at only one taxon, the probability of success would be the probability that the root state was conserved at this taxon, which is $1 - p$. That is, if $Y = \{x_1\}$ is a single taxon subset of X, then $RA(Y) = 1 - p$. In the following, we show that $1 - p$ is in fact a lower bound on $RA(X)$, implying that MP applied to all taxa reconstructs the root state at least as successfully as reconstructing the root state from a single taxon.

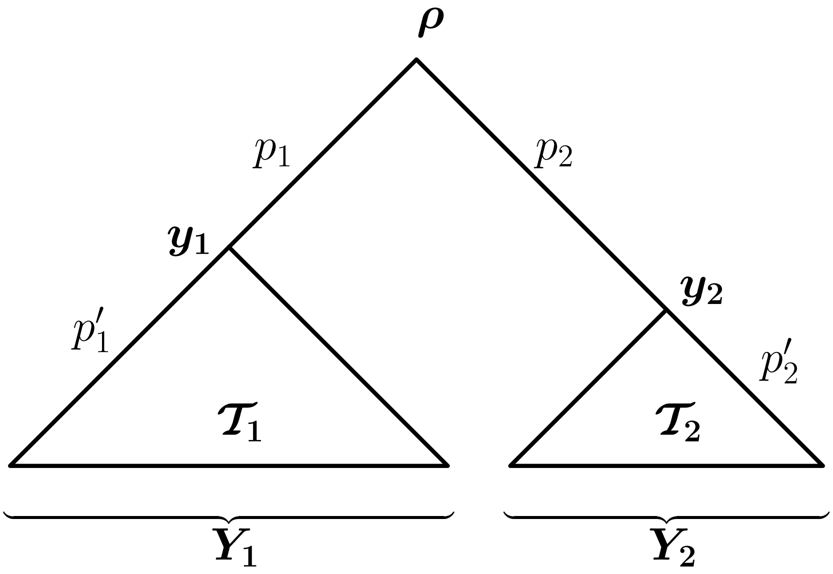


Figure 3.2: Illustration for Theorem 3.3: For any clocklike binary phylogenetic tree $\mathcal{T}$ the reconstruction accuracy of MP based on all leaves is at least as good as the one based on a single leaf.

As shown in Figure 3.2, we denote the children of ρ by y_1 and y_2, and define $\mathcal{T}_i$ to be the subtrees rooted at y_i for $i \in \{1, 2\}$. Let the probabilities of a change of state from ρ to y_i be p_i. The probabilities of a change of state from y_i to any leaf under y_i are p'_i. For i in $\{1, 2\}$, we define $P_i := 1 - 2p_i$. Similarly, we define $P := 1 - 2p$.

In the above notation, we prove the following lower bound on $RA(X)$.

Theorem 3.3. *For any rooted binary phylogenetic ultrametric (clocklike) tree $\mathcal{T}$ with leaf set X, the reconstruction accuracy of MP is at least equal to the conservation probability from the root to any leaf, that is,*

$$RA(X) \geq 1 - p.$$

Proof. We first state three recursions, which we use later to give an inductive proof of the theorem.

$$\begin{aligned}
P_\alpha(X) &= \left(\frac{1+P_1}{2}P_\alpha(Y_1)+\frac{1-P_1}{2}P_\beta(Y_1)\right)\left(\frac{1+P_2}{2}P_\alpha(Y_2)+\frac{1-P_2}{2}P_\beta(Y_2)\right)\\
&\quad +P_{\alpha\beta}(Y_1)\left(\frac{1+P_2}{2}P_\alpha(Y_2)+\frac{1-P_2}{2}P_\beta(Y_2)\right)\\
&\quad +\left(\frac{1+P_1}{2}P_\alpha(Y_1)+\frac{1-P_1}{2}P_\beta(Y_1)\right)P_{\alpha\beta}(Y_2)
\end{aligned}$$

$$\begin{aligned}
P_\beta(X) &= \left(\frac{1-P_1}{2}P_\alpha(Y_1)+\frac{1+P_1}{2}P_\beta(Y_1)\right)\left(\frac{1-P_2}{2}P_\alpha(Y_2)+\frac{1+P_2}{2}P_\beta(Y_2)\right)\\
&\quad +P_{\alpha\beta}(Y_1)\left(\frac{1-P_2}{2}P_\alpha(Y_2)+\frac{1+P_2}{2}P_\beta(Y_2)\right)\\
&\quad +\left(\frac{1-P_1}{2}P_\alpha(Y_1)+\frac{1+P_1}{2}P_\beta(Y_1)\right)P_{\alpha\beta}(Y_2)
\end{aligned}$$

$$\begin{aligned}
P_{\alpha\beta}(X) &= \left(\frac{1+P_1}{2}P_\alpha(Y_1)+\frac{1-P_1}{2}P_\beta(Y_1)\right)\left(\frac{1-P_2}{2}P_\alpha(Y_2)+\frac{1+P_2}{2}P_\beta(Y_2)\right)\\
&\quad +\left(\frac{1-P_1}{2}P_\alpha(Y_1)+\frac{1+P_1}{2}P_\beta(Y_1)\right)\left(\frac{1+P_2}{2}P_\alpha(Y_2)+\frac{1-P_2}{2}P_\beta(Y_2)\right)\\
&\quad +P_{\alpha\beta}(Y_1)P_{\alpha\beta}(Y_2)
\end{aligned}$$

We define $D(X) := P_\alpha(X)+\frac{1}{2}P_{\alpha\beta}(X)-\frac{1}{2}(1+P)$, and similarly we define $D_1 := D(Y_1)$ and $D_2 := D(Y_2)$. The above recursions can be manipulated with a computer algebra system to verify that

$$4D(X) = 2P_{\alpha\beta}(Y_1)D_2P_2 + 2P_{\alpha\beta}(Y_2)D_1P_1 + 2D_2P_2 + 2D_1P_1 + P_{\alpha\beta}(Y_1)P + P_{\alpha\beta}(Y_2)P.$$

Now, by induction on the number of leaves, we show that $D(X)$ is non-negative. The base case of the inductive proof is when Y_1 and Y_2 are singleton sets, in which case $D(Y_1)$, $D(Y_2)$ and $D(X)$ are all equal to 0, that is $RA(X)$ is $1-p$. Suppose that the tree $\mathcal{T}$ has n taxa, and suppose that $D(X)$ is non-negative for all trees having fewer than n taxa. Since both Y_1 and Y_2 contain fewer than n taxa, $D(Y_1)$ and $D(Y_2)$ are both non-negative. Since $P_{\alpha\beta}(Y_1)$, $P_{\alpha\beta}(Y_2)$, P_1 and P_2 are all non-negative, the right hand side of the above equation is non-negative, implying the theorem. □

3.4 Interpretation

In this chapter, we analyzed the question of how the Fitch MP algorithm performs when used to reconstruct the ancestral root state. In particular, we considered the problem for phylogenetic trees on which the probability of a change of state from the root vertex to any leaf is constant. Earlier simulation studies [e.g., 36; 52] suggested that the reconstruction accuracy is generally increased when more taxa are considered. But simulations conducted by Li, Steel and Zhang showed that even under a molecular clock, MP may perform better on certain subsets of taxa. In Chapter 3.1 we presented an example of a tree in which one of the subtrees at the root consists of a single leaf and a pending edge, and the other subtree is a balanced binary tree of depth h, for some large h, and small ($< \frac{1}{8}$) substitution probabilities on all edges. On this tree, we observed that the ancestral state reconstruction is more accurate if only the set of taxa on the balanced subtree is considered. This is in contrast to the example given by Li, Steel and Zhang in which an outgroup taxon closer to the root or a single fossil record may give a better estimate of the root state than considering the whole tree. As our example shows, even a very short edge connecting the root with a leaf cannot guarantee an accurate root state estimation if the remaining taxa induce a balanced tree with a large number of taxa. For such trees, it may be better to ignore the fossil or a taxon closer to the root. Thus, there seems to be no general theoretical guideline to decide what subsets of taxa are to be used for a more reliable reconstruction of the root state. In general, we believe that very long leaf edges would have an adverse effect on the ancestral state reconstruction using MP.

While using the data on a subset of taxa may give a more accurate estimate of the root state, in general a single taxon subset does not give a better reconstruction accuracy. We showed this in Section 3.3 by resolving a conjecture of Li, Steel and Zhang for the 2-state case. They conjectured that for r-state characters on an ultrametric (clocklike) tree and a symmetric model of substitution, ancestral state reconstruction using all taxa is at least as accurate as that using a single taxon. We expect such a result to be true even when there are more than two states.

Chapter 4

Misleading Sequences

> An educated person is one who has learned that information almost always turns out to be at best incomplete and very often false or misleading.
>
> Russell Baker

In the previous chapter, the reconstruction accuracy of Fitch's MP algorithm for the estimation of the ancestral root state was analyzed. However, the ever-growing amount of available genetic sequence data requires stochastic models for nucleotide substitution and tree reconstruction methods not only to allow for ancestral state reconstructions but also for the inference of phylogenetic trees. Unsurprisingly, such models and methods have therefore been widely discussed in the last decades (see, e.g., [10], [12], [38], [50]). One common method to infer phylogenetic trees is to transform large DNA data sets into distance matrices, to which distance methods can then be applied. This transformation, however, inevitably leads to some loss of information, but more seriously, distances can be positively misleading in extreme cases. Huson and Steel [19] have constructed sequences yielding a unique most parsimonious tree that is totally different from the tree univocally supported by the corresponding distances. But their construction required $n-1$ character states for n taxa and therefore cannot be realized with DNA sequences whenever $n > 5$.

We will show here that no more than three states are actually needed, that is, binary and ternary characters suffice for generating the extreme contrast between parsimony- and distance-based trees. We do this in a constructive way, i.e., for any choice of binary phylogenetic trees $\mathcal{T}_1$, $\mathcal{T}_2$, we show how to generate a sequence of binary and ternary characters for which $\mathcal{T}_1$ is the unique MP-tree and $\mathcal{T}_2$ is the best possible tree concerning the induced distances. For this purpose, parsimoniously non-informative characters are very useful because they can yield some signal for distance-based methods but do not discriminate between trees under the parsimony criterion. Specifically, not only can non-informative ternary characters jointly generate any non-trivial split metric (along with some trivial split metrics), they can totally cancel any non-trivial split metric (i.e., generate a star tree), too. We can thus freely navigate between perfectly additive distances by using only non-informative ternary characters. Last but not least, we compare the sequence length of our construction to that of the Huson-Steel construction and show that our sequences can be significantly shorter. We then introduce an approach on how to characterize all misleading sequences for the four taxa case. Finally, we analyze the relevance of misleading sequences for practical purposes and show a misleading example from a DNA alignment including a human gene.

4.1 Construction of Misleading Sequences from Ternary Characters

We begin with some notation. In the following, we will assume that X is a set of $n \geq 4$ taxa.

Let $d_f : X \times X \to \mathbb{N}$ denote the Hamming distance induced by a character f and $d_S := \sum\limits_{i=1}^{k} d_{f_i}$ for sequences $S = f_1...f_k$ of k characters. Note that since d_f only depends on the partitions associated with the character f, we may index d by the corresponding partitions instead. In the case of a k-ary character f, which partitions the taxon set into k parts $A_1, \ldots, A_k$, we also denote the Hamming distance as $d_{A_1|\ldots|A_k} := d_f$.

Recall that a non-informative ternary character f has three character states α, β, γ for which α and β are attained in X exactly once: $f(a) = \alpha$, $f(b) = \beta$, and $f(x) = \gamma$ for all taxa a, b, and x different from a and b. Such a character on X featuring the two taxa a, b thus induces a 'trivial' partition of X into two singletons and a remainder, for which we use the shorthand $a|b|X{-}a, b$. Similarly, the 'trivial' partition (alias trivial split) of X associated with a non-informative binary character distinguishing taxon a

is denoted by $a|X{-}a$.

We now give a formal definition of misleading sequences and distances, before we continue with some more preliminaries required to prove Theorem 4.4, which is the main result of this chapter.

Definition 4.1. *A sequence S of characters on a leaf set X, which is convex only on a binary phylogenetic X-tree $\mathcal{T}_1$ and whose derived Hamming distances are additive on a binary phylogenetic X-tree $\mathcal{T}_2$ such that $\mathcal{T}_1 \neq \mathcal{T}_2$, is called a* misleading sequence*, and its derived Hamming distances d_S are called* misleading distances*.*

We now explore some useful properties of the Hamming distance which are needed as prerequisites to construct misleading sequences from binary and ternary characters.

First, note that the mismatch distances derived from a partition with k parts $A_1, \ldots, A_k$ cannot be distinguished from the sum of the mismatch distances derived from the half-weighted splits $A_i|X{-}A_i$ $(i = 1, \ldots, k)$:

$$d_{A_1|\ldots|A_k} = \frac{1}{2}\left(d_{A_1|X-A_1} + \ldots + d_{A_k|X-A_k}\right). \tag{4}$$

In particular, if f is a non-informative ternary character featuring the two taxa a, b, then the derived mismatch distance $d_f = d_{a|b|X-a,b}$ equals the sum of the three mismatch distances derived from the half-weighted splits $a|X{-}a$, $b|X{-}b$, and $a{,}b|X{-}a{,}b$:

$$d_{a|b|X-a,b} = \frac{1}{2}\left(d_{a|X-a} + d_{b|X-b} + d_{a,b|X-a,b}\right). \tag{5}$$

This obvious relationship is illustrated in Figure 4.1. This representation thus permits substituting $d_{a|b|X-a,b}$ by $d_{a,b|X-a,b}$, or vice versa, modulo trivial split metrics.

Linear dependence of the split metrics on X leads to several equations; for instance, the most fundamental one equates the sum of all split metrics where one part has

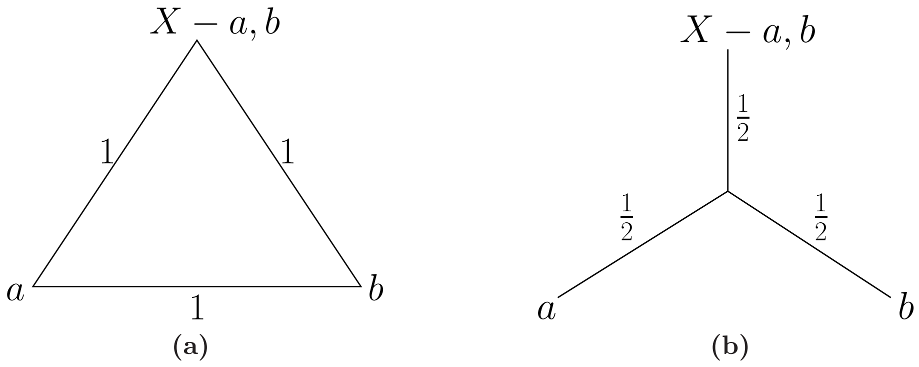


Figure 4.1: Illustration of Equation 5: Two ways of representing the distances derived from a non-informative ternary character by a network: a) quasi-median network (triangle); b) splits-tree (3-star).

cardinality exactly 2 with a multiple of the sum of all trivial split metrics:

$$\sum_{\substack{\{c,\, c'\} \,\subset\, X \\ c \,\neq\, c'}} d_{c,c'|X-c,c'} = (n-2) \cdot \sum_{z \in X} d_{z|X-z}. \tag{6}$$

This formula is straightforward to verify by evaluating either side on a pair x, y of different taxa, yielding $2(n-2)$ on either side.

For any X-split $A|B$, from (6) and the following trivial equation (7) one immediately obtains the subsequent equation (8):

$$\sum_{\substack{a \,\in\, A \\ b \,\in\, B}} d_{a,b|X-a,b} = \sum_{\substack{\{c,\, c'\} \,\subset\, X \\ c \,\neq\, c'}} d_{c,c'|X-c,c'} - \sum_{\substack{\{a,\, a'\} \,\subset\, A \\ a \,\neq\, a'}} d_{a,a'|X-a,a'} - \sum_{\substack{\{b,\, b'\} \,\subset\, B \\ b \,\neq\, b'}} d_{X-b,b'|b,b'}, \tag{7}$$

$$\sum_{\substack{a \,\in\, A \\ b \,\in\, B}} d_{a,b|X-a,b} = (n-2) \cdot \sum_{z \in X} d_{z|X-z} - \sum_{\substack{\{a,\, a'\} \,\subset\, A \\ a \,\neq\, a'}} d_{a,a'|X-a,a'} - \sum_{\substack{\{b,\, b'\} \,\subset\, B \\ b \,\neq\, b'}} d_{X-b,b'|b,b'}. \tag{8}$$

The latter equation is then used to express the split metric $d_{A|B}$ as a linear combination of split metrics where one split part has cardinality at most 2.

Lemma 4.2. *The split metric $d_{A|B}$ for a non-trivial split $A|B$ of an n-taxa set X satisfies the following two equations:*

$$d_{A|B} = \frac{1}{2}\left[|B|\sum_{a\in A} d_{a|X-a} + \quad |A|\sum_{b\in B} d_{X-b|b} - \sum_{\substack{a\,\in\, A \\ b\,\in\, B}} d_{a,b|X-a,b}\right], \tag{9}$$

$$d_{A|B} = \frac{1}{2}\left[\sum_{\substack{\{a,\,a'\}\,\subset\, A \\ a\,\neq\, a'}} d_{a,a'|X-a,a'} + \sum_{\substack{\{b,\,b'\}\,\subset\, B \\ b\,\neq\, b'}} d_{X-b,b'|b,b'} - \ (|A|-2)\sum_{a\in A} d_{a|X-a} - \ (|B|-2)\sum_{b\in B} d_{X-b|b}\right]. \tag{10}$$

Proof. To prove (9), one needs to evaluate either side on a pair x, y where x is from A, say, and y is either from $A-x$ or from B. The result on the right side is $\frac{1}{2}[2\cdot|B|-2\cdot|B|] = 0$ in the former case and equal to $\frac{1}{2}[|B|+|A|-(n-2)] = 1$ in the latter case, as required. From (9) one readily derives (10) by substituting the sum of all $d_{a,b|X-a,b}$ $(a \in A, b \in B)$ by the right-hand side of (8). □

This lemma and the preceding equations constitute the ingredients for the following proposition, which describes how one can either cancel or create a single split metric by means of metrics derived from non-informative ternary characters.

Proposition 4.3. *The split metric $d_{A|B}$ for a non-trivial split $A|B$ of an n-taxa set X can, up to a sum of trivial split metrics, either be canceled via*

$$d_{A|B} + \sum_{\substack{a\,\in\, A \\ b\,\in\, B}} d_{a|b|X-a,b} = |B|\sum_{a\in A} d_{a|X-a} + |A|\sum_{b\in B} d_{b|X-b}, \tag{11}$$

or be created via

$$\sum_{\substack{\{a,\,a'\}\,\subset\, A \\ a\,\neq\, a'}} d_{a|a'|X-a,a'} + \sum_{\substack{\{b,\,b'\}\,\subset\, B \\ b\,\neq\, b'}} d_{X-b,b'|b|b'} = d_{A|B} + \left(|A| - \frac{3}{2}\right)\sum_{a\in A} d_{a|X-a} + \left(|B| - \frac{3}{2}\right)\sum_{b\in B} d_{b|X-b} \tag{12}$$

through metrics derived from non-informative ternary characters.

Note that in Equation (11), the right-hand side consists of multiples of trivial split metrics only and thus corresponds to a star tree, i.e., unresolved tree, which is why we refer to this as a split 'cancelation'. Similarly, the right-hand side of Equation (12) consists only of a particular (non-trivial) split metric plus some multiples of trivial split metrics, which may change the length of pending edges but carry no information about interior edges, which is why it is adequate to refer to this as a split 'generation'.

Proof. Using the preceding lemmas, we compute

$$\begin{aligned}
d_{A|B} + \sum_{\substack{a \in A \\ b \in B}} d_{a|b|X-a,b} &\overset{(5)}{=} d_{A|B} + \frac{1}{2} \cdot \sum_{\substack{a \in A \\ b \in B}} \big(d_{a|X-a} + d_{b|X-b} + d_{a,b|X-a,b} \big) \\
&= d_{A|B} + \frac{1}{2} \cdot \left[|B| \cdot \sum_{a \in A} d_{a|X-a} + |A| \cdot \sum_{b \in B} d_{b|X-b} + \sum_{\substack{a \in A \\ b \in B}} d_{a,b|X-a,b} \right] \\
&\overset{(9)}{=} |B| \cdot \sum_{a \in A} d_{a|X-a} + |A| \cdot \sum_{b \in B} d_{b|X-b},
\end{aligned}$$

thus establishing (11). To prove (12), observe that

$$\begin{aligned}
&\sum_{\substack{\{a,a'\} \subset A \\ a \neq a'}} d_{a|a'|X-a,a'} + \sum_{\substack{\{b,b'\} \subset B \\ b \neq b'}} d_{X-b,b'|b|b'} \\
&\overset{(5)}{=} \frac{1}{2} \left[(|A|-1) \sum_{a \in A} d_{a|X-a} + (|B|-1) \sum_{b \in B} d_{b|X-b} + \sum_{\substack{\{a,a'\} \subset A \\ a \neq a'}} d_{a,a'|X-a,a'} + \sum_{\substack{\{b,b'\} \subset B \\ b \neq b'}} d_{X-b,b'|b,b'} \right] \\
&\overset{(10)}{=} \frac{1}{2} \left[2 d_{A|B} + (2 \cdot |A| - 3) \sum_{a \in A} d_{a|X-a} + (2 \cdot |B| - 3) \sum_{b \in B} d_{b|X-b} \right],
\end{aligned}$$

from which (12) follows immediately. □

With these prerequisites we can readily prove the following theorem.

Theorem 4.4. *For any two distinct binary phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ on the same taxon set X, there exists a sequence S of binary and ternary characters for which*

(i) S is homoplasy-free for $\mathcal{T}_1$ and for no other phylogenetic tree on X,

(ii) the distance function d_S derived from S is an ultrametric that is perfectly additive on $\mathcal{T}_2$ and on no other phylogenetic tree on X.

Proof. Let S_1 be a sequence of k binary characters that induces the system $\Sigma^*(\mathcal{T}_1)$ of the k non-trivial splits of the first tree, which correspond to the $k = |X| - 3$ interior edges of $\mathcal{T}_1$. Trivially, S_1 is homoplasy-free on $\mathcal{T}_1$ and on no other tree (this can be seen, for example, by using Theorem 3.1.4 of [38]). By adding suitable non-informative ternary characters we will cancel every split in $\Sigma^*(\mathcal{T}_1)$ with respect to the resulting mismatch distances, then create all members of $\Sigma^*(\mathcal{T}_2)$, that is, all non-trivial splits of $\mathcal{T}_2$, metrically and finally take care of ultrametricity by adding the necessary non-informative binary characters.

Specifically, for every split $A|B$ from $\Sigma^*(\mathcal{T}_1)$ and every pair $a \in A, b \in B$ we take a ternary character featuring the pair a, b, that is, inducing the partition $a|b|X{-}a, b$, and add it to the sequence. Then according to Proposition 4.3, Equation (11), the thus resulting expanded sequence S_2 yields pairwise distances conforming to a star metric. Next for each split $A_i|B_i$ from $\Sigma^*(\mathcal{T}_2)$ and all pairs a, a' from A_i and b, b' from B_i we take non-informative ternary characters featuring those pairs a, a' and b, b' and call the sequence consisting of these characters S'_i. Let $S' := S'_1 \ldots S'_{n-3}$ be the concatenation of all S'_i. Then, in view of Proposition 4.3, Equation (12), the mismatch distances with respect to sequence S_3, which is the concatenation of S_2 and S', are perfectly additive on $\mathcal{T}_2$ but no other phylogenetic tree on X.

Finally, in order to obtain an ultrametric, we first select a point r at distance either 0 or $\frac{1}{2}$ to the midpoint on a longest path of $\mathcal{T}_2$ which has integer distances to the taxa (labeling the leaves of $\mathcal{T}_2$) relative to d_{S_3}. Let μ be the maximum distance from r to a taxon in X. Then, for every $x \in X$, add $\mu - d_{S_3}(r, x)$ many non-informative binary characters featuring x to the current sequence S_3. Then the final sequence S delivers an ultrametric d_S and still supports $\mathcal{T}_2$ and only $\mathcal{T}_2$, as required. This finishes the proof of the Theorem. □

We will now illustrate our construction with the help of an example.

4.1.1 Illustration

$$\begin{array}{r cc||cccccc|cccccc||cccc|cccc||c}
1: & \alpha & \alpha & \alpha & \alpha & \alpha & \gamma & \gamma & \gamma & \alpha & \alpha & \gamma & \gamma & \gamma & \gamma & \alpha & \gamma & \gamma & \gamma & \alpha & \alpha & \gamma & \gamma & \beta \\
2: & \alpha & \alpha & \gamma & \gamma & \gamma & \alpha & \alpha & \alpha & \gamma & \gamma & \alpha & \alpha & \gamma & \gamma & \gamma & \alpha & \alpha & \gamma & \gamma & \gamma & \gamma & \alpha & \beta \\
3: & \beta & \alpha & \beta & \gamma & \gamma & \beta & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \alpha & \alpha & \beta & \gamma & \gamma & \gamma & \beta & \gamma & \alpha & \gamma & \alpha \\
4: & \beta & \beta & \gamma & \beta & \gamma & \gamma & \beta & \gamma & \beta & \gamma & \beta & \gamma & \beta & \gamma & \gamma & \beta & \gamma & \alpha & \gamma & \beta & \beta & \gamma & \beta \\
5: & \beta & \beta & \gamma & \gamma & \beta & \gamma & \gamma & \beta & \gamma & \beta & \gamma & \beta & \gamma & \beta & \gamma & \gamma & \beta & \beta & \gamma & \gamma & \gamma & \beta & \beta
\end{array}$$

Figure 4.2: Twenty-three characters discriminating five taxa (1 - 5), which support a perfect phylogeny disconcordant with the 'additive' tree faithfully representing the distance function derived from the character sequence.

In Figure 4.2, we display a sequence on five taxa consisting of two compatible binary, 20 non-informative ternary characters and one non-informative binary character that illustrates the character sequence g constructed in the proof of the Theorem. The first two characters induce the compatible splits 12|345 and 123|45. The next six ternary characters cancel the former split, whereas the subsequent six characters cancel the latter split metrically. This means that the first three blocks indicated in Figure 4.2, i.e., the first 14 characters, correspond to a star tree. Then the next block of four ternary characters establishes the novel distance split 13|245, whereas the subsequent block erects 134|25. The final binary character helps to make the distances ultrametric. These data support the perfect phylogeny shown in Figure 4.3(a), where the parsimony scores are indicated along the links. The ultrametric tree shown in Figure 4.3(b) (where the branch lengths are indicated) represents the corresponding mismatch distances.

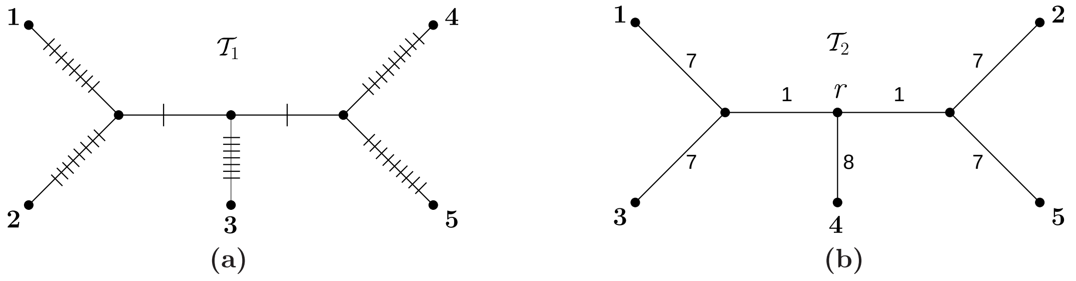


Figure 4.3: a) The perfect phylogeny supported by the data from Figure 4.2; b) the ultrametric representation by a tree rooted at node r for the distances derived from the data of Figure 4.2.

4.1.2 Sequence Length Analysis

We now analyze the sequence length of the misleading sequences as constructed in Theorem 4.4. Note that as there are various ways to derive sequence S, which consists of sequence S_3 and some added uninformative binary characters to make it an ultrametric, the following formula for the sequence length disregards the molecular clock property, i.e., it only refers to S_3 rather than S.

Corollary 4.5. *For X with $|X| = n \geq 4$ and trees $\mathcal{T}_1$ and $\mathcal{T}_2$ on X, let sequence S_3 be as constructed in the proof of Theorem 4.4. Then, the number $k(S_3)$ of characters in S_3 is given by*

$$k(S_3) = (n-3) + \sum_{i=1}^{n-3} |A_i| \cdot |B_i| + \sum_{i=1}^{n-3} \left(\binom{|\tilde{A}_i|}{2} + \binom{|\tilde{B}_i|}{2} \right),$$

where $\sigma_i = A_i|B_i \in \Sigma^*(\mathcal{T}_1)$, $\tilde{\sigma}_i = \tilde{A}_i|\tilde{B}_i \in \Sigma^*(\mathcal{T}_2)$.

Proof. The formula follows directly from the construction introduced in the proof of Theorem 4.4. □

In the following, we will show that this sequence length can be significantly shorter than the one of the Huson-Steel construction as presented in [19]. In order to see this, recall that Huson and Steel start with the following distances: $D_1(x, y)$ denotes the number of interior edges of $\mathcal{T}_1$ separating taxa x, y, and D_2 is an arbitrary ultrametric on $\mathcal{T}_2$ such that all interior edges have edge weight 1 and all pending edges have a non-negative integer edge weight. Moreover, they set $d_{ij} := |D_2(i,j) - D_1(i,j)|$, $s = \sum\limits_{i \in X} \sum\limits_{j \neq i \in X} d_{ij}$. Finally, they define $n_{ij} = r - B \cdot d_{ij} + s$, where $B = \left(\binom{n}{2} - 1\right)$ and r is such that $r - d_{ij} + s \geq 0$ for all d_{ij}.

Note that this construction does not lead to a unique sequence as there are various choices of D_2, which lead to different sequences and different edge lengths. The basic approach used by Huson and Steel is this: they first add B copies of all binary characters induced by the $n-3$ interior edges of $\mathcal{T}_1$, i.e., by $\Sigma^*(\mathcal{T}_1)$, and then they concatenate this sequence with n_{ij} copies of a character in which the (distinct) taxa i and j are in state α, and all other taxa are in states other than α and different from one another. The latter characters are all non-informative. So as in our approach, non-informative characters are used to cancel splits induced by the binary characters defining the perfect phylogeny. But the non-informative characters of the Huson-Steel approach employ $n-1$ character states (where $n = |X|$) as opposed to the three states required by the Bandelt-Fischer approach.

Concerning the sequence length of the Huson-Steel construction, we can now state the following lemma.

Lemma 4.6. *The number of characters $k(\tilde{S})$ of the misleading sequence $\tilde{S}$ constructed in* [19] *is*

$$k(\tilde{S}) = \left(\binom{n}{2} - 1\right)(n-3) + \sum_{i \in X} \sum_{j \neq i \in X} n_{ij}. \tag{13}$$

In order to compare $k(S_3)$ and $k(\tilde{S})$, we have to make sure that we do not make $k(\tilde{S})$ unnecessarily large. Note that as soon as D_2 is chosen, the values for d_{ij} and s are fixed. So one can only minimize $\sum n_{ij}$ by making r minimal. This means that r has to be chosen such that $r = B \cdot \max_{ij} d_{ij} - s$. Let $d^*_{ij} = \max_{ij} d_{ij}$.

Lemma 4.7. *The number of characters $k(\tilde{S})$ of the misleading sequence $\tilde{S}$ constructed in* [19] *is bounded from below by*

$$k(\tilde{S}) \geq \left(\binom{n}{2} - 1\right) \cdot \left((n-3) + \binom{n}{2} \cdot d^*_{ij} - s\right).$$

Proof. From Lemma 4.6 we have

$$\begin{aligned} k(\tilde{S}) &= \left(\binom{n}{2} - 1\right)(n-3) + \sum_{i \in X} \sum_{j \neq i \in X} (r - B \cdot d_{ij} + s) \\ &= \left(\binom{n}{2} - 1\right)(n-3) + \binom{n}{2}(r+s) - B \cdot \sum_{i \in X} \sum_{j \neq i \in X} d_{ij} \\ &= \left(\binom{n}{2} - 1\right)(n-3) + \binom{n}{2}(r+s) - B \cdot s \\ &\geq \left(\binom{n}{2} - 1\right)(n-3) + \left(\binom{n}{2} - 1\right) \cdot \left(\binom{n}{2} \cdot d^*_{ij} - s\right) \\ &= \left(\binom{n}{2} - 1\right) \cdot \left((n-3) + \binom{n}{2} \cdot d^*_{ij} - s\right), \end{aligned}$$

where the lower bound in the second to last step is obtained by using the the minimal value for r, i.e., $r = B \cdot d^*_{ij} - s$. □

Even the lower bound on the sequence length of $\tilde{S}$ is not independent of the original choice of D_2. Therefore, the sequence lengths of the Huson-Steel approach and the Bandelt-Fischer approach cannot directly be compared. However, we will show with the following example that the difference can in fact be huge.

Example 4.8. Consider again trees $\mathcal{T}_1$ and $\mathcal{T}_2$ given by Figure 4.3. We will now follow the Huson-Steel approach for these trees, and we will also use D_2 as given in Figure 4.3.

Table 2 shows the values required for the Huson-Steel approach. Note that here, r is chosen to be minimal. Since $B = \binom{5}{2} - 1 = 9$ by definition, and since $s = 144$

and $\max_{ij} d_{ij} = 16$ according to Table 2, the minimal choice for r turns out to be $9 \cdot 16 - 144 = 0$. In the following, let $N_r := \sum\limits_{i \in X} \sum\limits_{j \neq i \in X} n_{ij}$.

(i,j)	$D_1(i,j)$	$D_2(i,j)$	d_{ij}	n_{ij}
$(1,2)$	0	16	16^*	0
$(1,3)$	1	14	13	27
$(1,4)$	2	16	14	18
$(1,5)$	2	16	14	18
$(2,3)$	1	16	15	9
$(2,4)$	2	16	14	18
$(2,5)$	2	14	12	36
$(3,4)$	1	16	15	9
$(3,5)$	1	16	15	9
$(4,5)$	0	16	16^*	0
Sum:			$s = 144$	$N_0 = 144$

Table 2: Overview of the values needed for the Huson-Steel approach in order to construct a misleading sequence which is homoplasy-free on tree $\mathcal{T}_1$ and additive on tree $\mathcal{T}_2$ given by Figure 4.3. Note that the value of r minimizing the sequence length is used, which here means $r = 0$.

Figure 4.4 depicts the minimal sequence achieved by the Huson-Steel approach, i.e., the minimal sequence constructable by their approach with $\mathcal{T}_1$ being the perfect phylogeny and the derived distances being treelike only on $\mathcal{T}_2$. The sequence length is, according to Lemma 4.6, $\left(\binom{5}{2} - 1\right)(5-3) + N_r = 9 \cdot 2 + 144 = 162$. Recall that the sequence length $k(S_3)$ of the approach presented above was 22 (and the whole sequence S given by Figure 4.2, which contained an additional character to achieve an ultrametric has length 23) and was thus significantly shorter than the Huson-Steel sequence. This shows that sequences consisting of binary and ternary characters only may be more efficient in constructing misleading sequences.

It is important to state, however, that the Bandelt-Fischer way of generating misleading sequences is at least not generally 'best possible' with regard to sequence length, even if we do not insist on the sequence to fit a molecular clock (but recall that the construction is 'best possible' concerning the number of employed character states). First of all, the construction given in the proof of Theorem 4.4 disregards the fact that $\mathcal{T}_1$ and $\mathcal{T}_2$ might have some non-trivial splits in common. In this case, these splits do

$$\begin{array}{lccccccccc|ccccccccc||ccccccccc}
1: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\
2: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta \\
3: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\
4: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma \\
5: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta
\end{array}$$

$$\begin{array}{lcccccccccccccccccc|ccccccccc}
1: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\
2: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta \\
3: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma \\
4: & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\
5: & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta
\end{array}$$

$$\begin{array}{lccccccccc|cccccccccccccccccc|}
1: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\
2: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta \\
3: & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma \\
4: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta \\
5: & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha
\end{array}$$

$$\begin{array}{lccccccccc|cccccccccccccccccc|}
1: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\
2: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta \\
3: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma \\
4: & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta \\
5: & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta
\end{array}$$

$$\begin{array}{lccccccccccccccccccccccccccc}
1: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\
2: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta \\
3: & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma \\
4: & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta \\
5: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta
\end{array}$$

$$\begin{array}{lccccccccc|ccccccccc|ccccccccc}
1: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\
2: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta \\
3: & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma \\
4: & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta \\
5: & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \beta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \delta & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma & \gamma
\end{array}$$

Figure 4.4: The misleading Huson-Steel sequence of minimal length for trees $\mathcal{T}_1$ and $\mathcal{T}_2$ as given in Figure 4.3. The sequence length is 162, which is roughly seven times longer than the sequence of length 23 given by Figure 4.2.

not have to be metrically cancelled and then created again via ternary non-informative characters in order to obtain a sequence which is homoplasy-free only on $\mathcal{T}_1$ and whose derived Hamming distances are additive only on $\mathcal{T}_2$. Moreover, in Chapter 4.2 we will show that for $n = 4$ we can also find a shorter misleading sequence than that provided by the proof of Theorem 4.4 (in particular, we will see that a sequence length of 6 is achieveable, which is not possible with the instructions given above). Naturally, since 4-taxa trees only have one internal edge, different 4-taxa trees have no non-trivial split in common. So even in such cases, there may exist misleading sequences which are shorter than suggested by the above construction. Moreover, Huber, Moulton and Steel showed that in general, i.e., for any binary phylogenetic tree $\mathcal{T}$ on any number of taxa, there are always four characters which are convex on $\mathcal{T}$ and no other tree [18]. This suggests that the first part of the misleading sequence construction, namely the one consisting of informative characters, can be made significantly shorter than suggested in the proof

	pattern	character type	corresponding variable
1 :	$\alpha\alpha\beta\beta$	binary, informative	x_1
2 :	$\alpha\beta\beta\beta$	binary, non-informative	x_2
3 :	$\alpha\beta\alpha\alpha$	binary, non-informative	x_3
4 :	$\alpha\alpha\beta\alpha$	binary, non-informative	x_4
5 :	$\alpha\alpha\alpha\beta$	binary, non-informative	x_5
6 :	$\alpha\alpha\beta\gamma$	ternary, non-informative	x_6
7 :	$\alpha\beta\alpha\gamma$	ternary, non-informative	x_7
8 :	$\alpha\beta\gamma\alpha$	ternary, non-informative	x_8
9 :	$\alpha\beta\beta\gamma$	ternary, non-informative	x_9
10 :	$\alpha\beta\gamma\beta$	ternary, non-informative	x_{10}
11 :	$\alpha\gamma\beta\beta$	ternary, non-informative	x_{11}
12 :	$\alpha\beta\alpha\beta$	binary, informative	x_{12}
13 :	$\alpha\beta\beta\alpha$	binary, informative	x_{13}
14 :	$\alpha\beta\gamma\delta$	quaternary, non-informative	x_{14}
15 :	$\alpha\alpha\alpha\alpha$	unitary, non-informative	x_{15}

Table 3: This table lists the 15 different character patterns or partitions which are possible on four taxa and their corresponding variable.

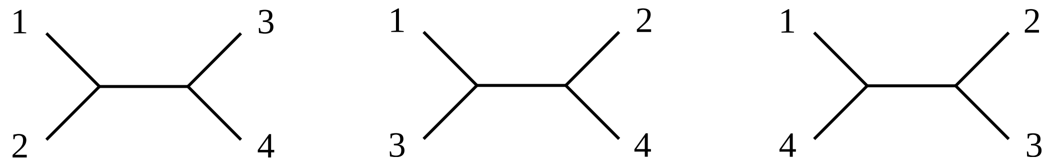
of Theorem 4.4 (but this may require a lot more character states). A general lower bound on the number of characters required to construct a misleading sequence would therefore be of interest. We provide such a bound for the 4-taxa case in the following chapter.

4.2 A Characterization of Misleading Sequences for Four Taxa

We now introduce an approach on how to characterize all misleading sequences on four taxa. Here, we explicitly consider all misleading sequences, i.e., including those which do not conform to an ultrametric. Our main objective is to state equations and inequalities that need to be fulfilled by a misleading sequence.

On four taxa, there are 15 characters (when for instance $\alpha\alpha\beta\beta$, $\beta\beta\alpha\alpha$, $\gamma\gamma\delta\delta$ etc. are assumed to be equal as they correspond to the same partitioning of the leaf set X). These 15 characters are listed in Table 3. In the following, we denote by x_i the number of characters of type i, for some $i \in \{1, \ldots, 15\}$ used in a sequence.

As depicted in Figure 4.5, there are three different unrooted binary phylogenetic trees on four taxa: 12|34, 13|24 and 14|23.



(a) (b) (c)

Figure 4.5: Illustration of the three different unrooted trees on 4 taxa.

We now analyze the case where the sequence is convex only on 12|34 and additive only on 13|24. The other cases can be analyzed analogously. In order to make a sequence convex on 12|34 and no other tree, it has to employ the binary informative character $\alpha\alpha\beta\beta$, i.e., $x_1 \geq 1$, and no other binary informative character may be used (i.e. $x_{12} = x_{13} = 0$).

By analyzing the characters 1 – 11 and 14 – 15 given by Table 3, one can easily see which character contributes to the distance between taxa i and j for $i, j \in \{1, 2, 3, 4\}, i \neq j$. Then, the Hamming distances derived by a sequence employing only characters 1 – 11 and 14 – 15 can be calculated as follows:

$$\begin{aligned}
d_{12} &= x_2+x_3+x_7+x_8+x_9+x_{10}+x_{11}+x_{14}+x_{15}\\
d_{13} &= x_1+x_2+x_4+x_6+x_8+x_9+x_{10}+x_{11}+x_{14}+x_{15}\\
d_{14} &= x_1+x_2+x_5+x_6+x_7+x_9+x_{10}+x_{11}+x_{14}+x_{15}\\
d_{23} &= x_1+x_3+x_4+x_6+x_7+x_8+x_{10}+x_{11}+x_{14}+x_{15}\\
d_{24} &= x_1+x_3+x_5+x_6+x_7+x_8+x_9+x_{11}+x_{14}+x_{15}\\
d_{34} &= x_4+x_5+x_6+x_7+x_8+x_9+x_{10}+x_{14}+x_{15}
\end{aligned}$$

By [38], Theorem 7.2.6, the 4-point condition is a necessary and sufficient criterion for a sequence to be a tree metric. In order to be a tree metric on 13|24 and no other tree, the derived distances thus have to fulfill the 4-point condition on 13|24: we require

$$d_{12} + d_{34} = d_{14} + d_{23} > d_{13} + d_{24}.$$

Note that in order to avoid a star tree, the last inequality has to be strict. Furthermore, we require all triangle inequalities to hold (note that these are induced by the general form of the 4-point condition as the four points need not be all distinct):

$$d_{ij} + d_{jk} - d_{ik} \geq 0 \text{ for all } i, j, k \in \{1, 2, 3, 4\}.$$

So altogether we have the following conditions on d to make it a tree metric (exclusively) on $13|24$:

$$\begin{aligned}
d_{12} + d_{34} = d_{14} + d_{23} &> d_{13} + d_{24} && (14)\\
d_{12} + d_{23} &\geq d_{13} && (15)\\
d_{13} + d_{23} &\geq d_{12} && (16)\\
d_{12} + d_{13} &\geq d_{23} && (17)\\
d_{12} + d_{24} &\geq d_{14} && (18)\\
d_{14} + d_{24} &\geq d_{12} && (19)\\
d_{12} + d_{14} &\geq d_{24} && (20)\\
d_{13} + d_{34} &\geq d_{14} && (21)\\
d_{14} + d_{34} &\geq d_{13} && (22)\\
d_{13} + d_{14} &\geq d_{34} && (23)\\
d_{23} + d_{34} &\geq d_{24} && (24)\\
d_{24} + d_{34} &\geq d_{23} && (25)\\
d_{23} + d_{24} &\geq d_{34} && (26)
\end{aligned}$$

This leads directly to the following corollary.

Corollary 4.9. *Any choice of $x_1, \ldots, x_{15}$ with $x_1 \geq 1$ and $x_{12} = x_{13} = 0$ such that Inequalities (14) – (26) are fulfilled induces a misleading sequence S on four taxa. More precisely, S is homoplasy-free only on tree $12|34$ and additive only on tree $13|24$. Moreover, for any sequence S which is homoplasy-free only on tree $12|34$ and additive only on tree $13|24$ we have $x_1 \geq 1$ and $x_{12} = x_{13} = 0$. Additionally, such a sequence fulfills Inequalities (14) – (26).*

Proof. The claim follows directly from the preceding arguments. □

Remark: Note that the roles of the trees can be swapped by modifying the inequalities accordingly. Thus, a full characterization of all misleading sequences on 4 taxa can be achieved with this approach.

We are now in the position to prove a lower bound for the sequence length required to construct misleading sequences on four taxa.

Proposition 4.10. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two distinct binary phylogenetic X-trees, where $|X| = 4$. Then, any sequence S of characters on X such that*

1. *$\mathcal{T}_1$ is the unique MP-tree and*

2. *the Hamming distances derived from S are perfectly additive on $\mathcal{T}_2$*

has length at least six.

Proof. We assume wlog. that $\mathcal{T}_1 = 12|34$ and $\mathcal{T}_2 = 13|24$. Then, we know that in order for $\mathcal{T}_1$ to be the unique MP-tree, we require $x_{12} = x_{13} = 0$. Note that the unitary non-informative character $\alpha\alpha\alpha\alpha$ and the quaternary non-informative character $\alpha\beta\gamma\delta$ are not only non-informative regarding parsimony, but also regarding distances. Therefore, in order to determine the minimum length of a misleading sequence S, we can furthermore assume wlog. $x_{14} = x_{15} = 0$. Then, by Inequality (14), we need $d_{12} + d_{34} = d_{14} + d_{23} > d_{13} + d_{24}$. As above we can express the distances as follows:

1. $d_{12} + d_{34} = x_2 + x_3 + x_4 + x_5 + x_6 + 2x_7 + 2x_8 + 2x_9 + 2x_{10}$

2. $d_{14} + d_{23} = 2x_1 + x_2 + x_4 + x_5 + 2x_6 + 2x_7 + x_8 + x_9 + 2x_{10} + 2x_{11}$

3. $d_{13} + d_{24} = 2x_1 + x_2 + x_3 + x_4 + x_5 + 2x_6 + x_7 + 2x_8 + 2x_9 + x_{10} + 2x_{11}$

This leads to $x_7+x_{10} > x_8+x_9 = 2x_1+x_6+2x_{11}$. As $x_1 \geq 1$ as $\mathcal{T}_1$ is the unique MP-tree, this means that $x_8+x_9 \geq 2$ and $x_7+x_{10} \geq 3$. Therefore, we have $x_1+x_7+x_8+x_9+x_{10} \geq 6$. This completes the proof. □

Corollary 4.11. *Let S be a misleading sequence on $|X| = n \geq 4$ taxa, such that S contains all binary characters induced by the non-trivial splits of its perfect phylogeny. Then, S has length at least six. Moreover, the function $f : \mathbb{N} \to \mathbb{N}$ defined by $n \mapsto \min_S \text{length}(S)$, which maps the number of taxa to the minimum length of misleading sequences on n taxa with the additional property of containing all binary characters induced by the non-trivial splits of their perfect phylogeny, is non-decreasing.*

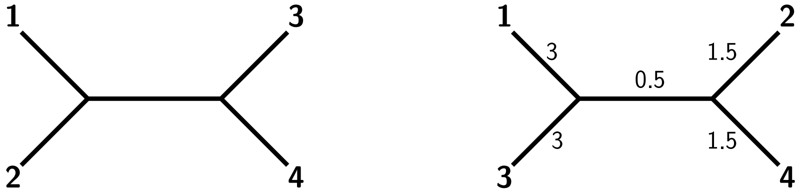
Proof. We will prove the claim by induction on $n = |X|$. For $|X| = 4$, the result follows from Proposition 4.10. Now let n be a number of taxa such that for misleading sequences (with the additional property of containing all binary characters induced by the non-trivial splits of their perfect phylogeny) on fewer than n taxa we already know that they have at least length k for some $k \geq 6$, and that $f(i)$ is non-decreasing over the range $i = 4, \ldots, n-1$. Let S be a character sequence which has $\mathcal{T}_1$ as its unique perfect phylogeny and which contains all $n-3$ binary characters induced by its non-trivial splits and whose derived Hamming distances are treelike only on $\mathcal{T}_2$, where $\mathcal{T}_1 \neq \mathcal{T}_2$ are X-trees with $|X| = n$. Let additionally the length of S be $f(n)$ (i.e., S is a misleading sequence containing all binary characters induced by splits of $\mathcal{T}_1$ of minimum length). Now for an arbitrary taxon $x \in X$, we consider $S^x := S_{|X-\{x\}}$, $\mathcal{T}_1^x := \mathcal{T}_{1|X-\{x\}}$ and $\mathcal{T}_2^x := \mathcal{T}_{2|X-\{x\}}$, i.e., the restrictions of S, $\mathcal{T}_1$ and $\mathcal{T}_2$ on $X - \{x\}$, respectively. Note that S^x is convex on $\mathcal{T}_1^x$. This is due to the fact that the removal of a row in the alignment S cannot create homoplasies as some informative characters may become non-informative, but non-informative characters remain non-informative. Therefore, no conflicting signals are added. Furthermore, note that the Hamming distances derived from S^x are additive on $\mathcal{T}_2^x$, as deleting a row of S cannot destroy additivity: since the distances derived by S fit on $\mathcal{T}_2$, by disregarding x we do not change the fact that the remaining distances fit on the remaining edges. Now, S contains by assumption all binary informative characters such that all non-trivial splits of $\mathcal{T}_1$ are represented. If taxon x does not belong to a cherry of $\mathcal{T}_1$, for all splits $A|B$ of $\mathcal{T}_1$ with $x \in A$ we have $|A| \geq 3$. Therefore, deleting the row corresponding to x in the alignment cannot turn any informative characters into non-informative ones in this case. If, on the other hand, taxon x belongs to a cherry of $\mathcal{T}_1$, there is a taxon $y \in X$, $y \neq x$, such that the split $xy|X - \{x, y\}$ is part of $\mathcal{T}_1$ and represented in S by a binary character of the kind $\alpha\alpha\beta \ldots \beta$. The restriction of this character on $X - \{x\}$ is non-informative, as now one of the states (here α) remains only once. But all other binary characters representing non-trivial splits of $\mathcal{T}_1$ remain informative when restricted to $X - \{x\}$. In particular, all non-trivial splits of $\mathcal{T}_1^x$ appear in S^x. Thus, S^x is convex *only* on $\mathcal{T}_1^x$ and contains all binary characters induced by non-trivial splits of $\mathcal{T}_1^x$. Now assume that there is no $x \in X$ such that S^x is misleading. Since S^x is convex only on T_1^x and additive on T_2^x, S^x is not misleading only if $\mathcal{T}_1^x = \mathcal{T}_2^x$ for all x, i.e. all subtrees of $\mathcal{T}_1$ and $\mathcal{T}_2$ of size $n-1$ are equal. Then, all quartet subtrees of $\mathcal{T}_1$ and $\mathcal{T}_2$ are also equal (as they all belong to some subtree on $n-1$ taxa), and thus, by Theorem 6.3.9 of [38], $\mathcal{T}_1$ and $\mathcal{T}_2$ are equal. This is a contradiction. So there must be an $x \in X$ for which S^x is misleading, i.e., with $\mathcal{T}_1^x \neq \mathcal{T}_2^x$. Since S^x is a sequence on $n-1$ taxa, by induction it has length at least $f(n-1) \geq k$. Since S has the same length as S^x (as only a row was removed but not a column), this completes the proof. □

We now demonstrate with the following example that the lower bound on the sequence length given in Proposition 4.10 is a tight bound, i.e., can be achieved. The example also shows how misleading sequences on four taxa can be constructed using Corollary 4.9.

Example 4.12. Let $x_1 = x_8 = x_9 = 1$, $x_{10} = 3$, $x_2 = \ldots = x_7 = x_{11} = x_{12} = x_{13} = x_{14} = x_{15} = 0$. It can be easily checked that this choice fulfills the requirements of Corollary 4.9, and it corresponds to the following sequence S:

$$S: \quad \begin{array}{llllll l} 1: & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\ 2: & \alpha & \beta & \beta & \beta & \beta & \beta \\ 3: & \beta & \gamma & \beta & \gamma & \gamma & \gamma \\ 4: & \beta & \alpha & \gamma & \beta & \beta & \beta \end{array}$$

Trivially, this sequence is homoplasy-free only on tree 12|34, as the only informative character in the sequence is $\alpha\alpha\beta\beta$. The corresponding distances derived from the sequence are $d_{12} = d_{14} = d_{23} = d_{34} = 5$, $d_{13} = 6$ and $d_{24} = 3$. These distances fit on tree 13|24 as illustrated in Figure 4.6 (but they do not conform to a molecular clock).



(a) (b)

Figure 4.6: The tree illustrated in (a) is the perfect phylogeny for S, whereas (b) depicts the tree on which S is additive.

Note that Example 4.12 provides a shorter misleading sequence than achieveable with the construction given in the proof of Theorem 4.4, even if we disregard the non-informative binary characters that make the sequence clocklike. Using Corollary 4.5, for $n = 4$ the sequence length of the (non-clocklike) misleading sequence which has $\mathcal{T}_1 = 12|34$ as its perfect phylogeny and whose derived Hamming distances are additive only on $\mathcal{T}_2 = 13|24$ as suggested by the proof of Theorem 4.4 can be calculated as follows:

$$(4-3) + |\{1,2\}| \cdot |\{3,4\}| + \left[\binom{|\{1,3\}|}{2} + \binom{|\{2,4\}|}{2} \right] = 1 + 2 \cdot 2 + (1+1) = 7,$$

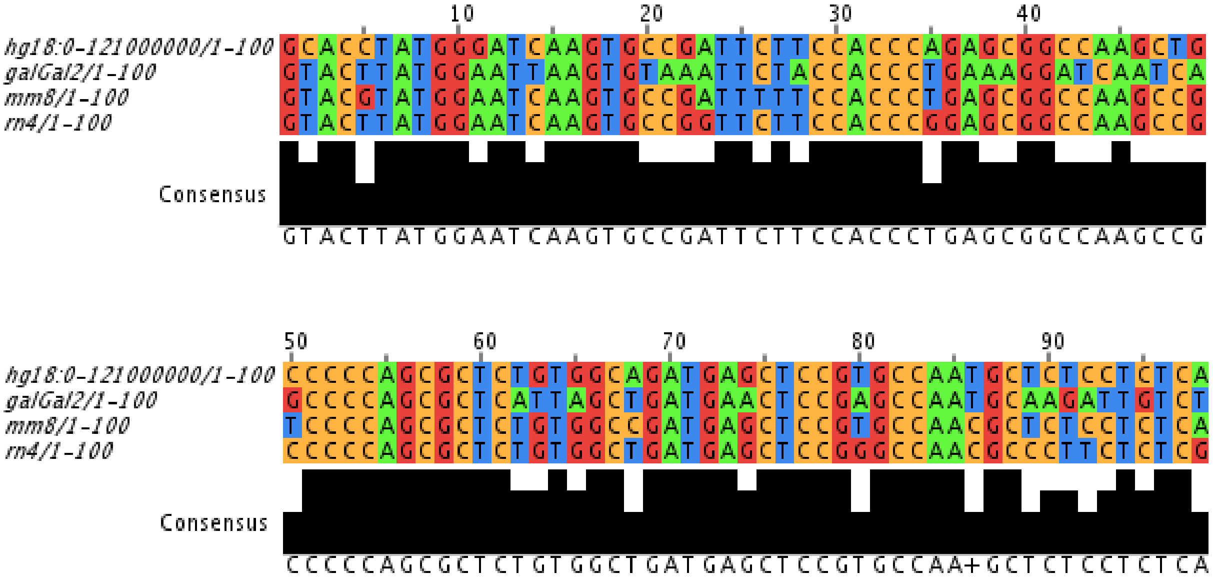
which proves that this sequence is longer than the one of length 6 given in the above example.

The sequence constructed in Example 4.12 will be analyzed further in Chapter 5.2, as for certain model assumptions, the ML-tree for this sequence coincides with $\mathcal{T}_2$, i.e., with the distance-wise best choice, rather than with MP.

4.3 Relevance of Misleading Sequences in Practice

For various reasons it is often argued that misleading sequences are not too relevant for practical purposes. In fact, even homoplasy-free sequences are rare, and homoplasy-free sequences fulfilling an additional property are even rarer. We will discuss this issue further in the next chapter. But first we show with an example that misleading sequences indeed do occur in real data, even when the human genome is considered.

Using the UCSC Genome Browser, Wolfgang Fischl and I aligned the following genes: the human gene *hg18* located on Chromosome 1, positions 37720932 – 37721032, the corresponding genes of chicken (*galGal2*), mouse (*mm8*) and rat (*rn4*). The resulting alignment is depicted in Figure 4.7.



```
                                  10        20        30        40
hg18:0-121000000/1-100  GCACCTATGGGATCAAGTGCCGATTCTTCCACCCAGAGCGGCCAAGCTG
galGal2/1-100           GTACTTATGGAATTAAGTGTAAATTCTACCACCCTGAAAGGATCAATCA
mm8/1-100               GTACGTATGGAATCAAGTGCCGATTTTTCCACCCTGAGCGGCCAAGCCG
rn4/1-100               GTACTTATGGAATCAAGTGCCGGTTCTTCCACCCGGAGCGGCCAAGCCG
             Consensus
                        GTACTTATGGAATCAAGTGCCGATTCTTCCACCCTGAGCGGCCAAGCCG

                        50        60        70        80        90
hg18:0-121000000/1-100  CCCCCAGCGCTCTGTGGCAGATGAGCTCCGTGCCAATGCTCTCCTCTCA
galGal2/1-100           GCCCCAGCGCTCATTAGCTGATGAACTCCGAGCCAATGCAAGATTGTCT
mm8/1-100               TCCCCAGCGCTCTGTGGCCGATGAGCTCCGTGCCAACGCTCTCCTCTCA
rn4/1-100               CCCCCAGCGCTCTGTGGCTGATGAGCTCCGGGCCAACGCCCTTCTCTCG
             Consensus
                        CCCCCAGCGCTCTGTGGCTGATGAGCTCCGTGCCAA+GCTCTCCTCTCA
```

Figure 4.7: DNA alignment built with the UCSC Gene Browser: human gene *hg18* (Chromosome 1, positions 37720932 – 37721032), chicken gene *galGal2*, mouse gene *mm8* and rat gene *rn4*. This whole sequence contains only one informative binary character *TTCC* at the 86th position.

After aligning these genes, we removed all unitary non-informative characters, as these are also distance-wise uninformative in the sense that they induce the same distance, namely 0, between all pairs of species for all possible underlying trees. For the same reason, we would have disregarded all quaternary non-informative characters, but the alignment shown in Figure 4.7 does not contain any of these, and it also does not contain any gaps. The sequence resulting from removing the unitary characters is shown in Figure 4.8, where the only binary informative character is marked with an asterix. Note that with regards to parsimony, i.e., in particular to convexity, as well as concerning the induced distances, the explicite character states do not matter, only the patterns or partitions induced by them are relevant. Table 4 summarizes this information of the alignment given in Figure 4.8.

```
                                                                *
1    (Human):  C C G C C C G A C T A G C C C A G C T G C T G G A G T T T C T C C C A
2  (Chicken):  T T A T T A A A C A T A A A T C A T C A G A T A T A A T A A G A T G T
3    (Mouse):  T G A C C C G A T T T G C C C A G C C G T T G G C G T C T C T C C C A
4      (Rat):  T T A C C C G G C T G G C C C A G C C G C T G G T G G C C C T T C C G
```

Figure 4.8: The alignment resulting from the alignment given in Figure 4.7 after removing all characters which are not required. The only informative character is marked with the asterix symbol.

Pattern	α α β β	α β β β	α β α α	α α β α	α α α β	α β α γ	α β γ α	α β β γ	α β γ β
Frequency	1	3	21	1	1	4	1	1	2

Table 4: Summary of the information provided by the alignment of Figure 4.8.

Since the alignment given in Figure 4.8 contains only one informative binary character, the sequence is homoplasy-free only on the corresponding tree, namely 12|34. Moreover, the sequence is also treelike, which can be seen in Figure 4.9(b). Therefore, this sequence is misleading.

Unsurprisingly, in the distance tree the chicken branch is by far the longest. Therefore, the sequence does not conform to a molecular clock. However, it was necessary to include a less related species in the analysis to achieve better results. In [13], Fischl

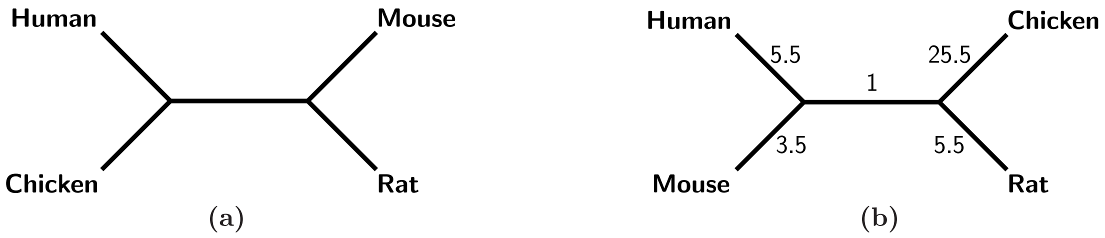


Figure 4.9: Tree 12|34 is the perfect phylogeny for the sequence given in Figure 4.8, whereas the tree 13|24 is the tree on which the derived distances are additive.

conducted the same analysis including the rhesus gene *rheMac2* instead of the chicken gene. He found that because of the relatedness of human and rhesus, there were no misleading sequences in the whole genome. However, as soon as the chicken genome was included, the situation changed. In order to analyze his findings, Fischl distinguishes between sequences which deliver incongruent trees and sequences which are misleading. Incongruent trees are trees for which the best parsimony tree does not coincide with the best distance-derived tree, but the first does not have to be homoplasy-free and the sequence does not need to be additive on the latter. We could also call such sequences 'weakly misleading' as they lead to different trees as different tree inference methods are applied. Fischl found that the number of misleading sequence parts in the whole genome alignment varies strongly from chromosome to chromosome. Chromosome 19, for example, returned about ten times more sequences of length 1000 with 'incongruent' trees than the other chromosomes. Only when searching for short misleading sequences of length 20, Fischl found no significant differences between the chromosomes.

Fischl's approach was to search the alignments with the help of a computer program that allows for different 'window' sizes. A window corresponds to the length of the alignment part under investigation. His results for the whole genome alignment of the species human, chicken, mouse and rat are summarized in Table 5.

Table 5 shows that long misleading sequences (> 250 bp) are not to be found. However, 1297 misleading short sequences show that misleading sequences not only

window size	incongruent	additive	convex	misleading
20	80021 (2.134%)	8329	34402	1297 (0.0346%)
50	35761 (2.384%)	4319	9580	512 (0.0341%)
100	11745 (1.567%)	1170	1267	80 (0.0107%)
250	1577 (0.526%)	79	14	0 (0%)
500	378 (0.252%)	7	0	0 (0%)
1000	110 (0.147%)	2	0	0 (0%)

Table 5: Results of the whole genome alignment: no misleading sequences longer than 250 basepairs were found, and the length of the whole alignment (disregarding all gaps) was roughly 75,000,000 bp.

exist in theory, and the 80021 sequences with incongruent trees even more underline the discrepancies between distance-based and parsimony-based methods. However, the percentages show that the phenomenon of misleading sequences is not a major issue when reconstructing trees from sequence data.

4.4 Interpretation

Tree reconstruction from character data, such as DNA sequences, is a common approach to estimate phylogenetic trees. Often, however, the given data are not additive, i.e., not treelike, for any tree but rather represent a network, which usually requires further examination. Treelike data, on the other hand, intuitively seem uncomplicated and unequivocal. But our results show that even treelike DNA sequences on any number of species larger than three may be positively misleading, as they may be treelike on a tree different from the perfect phylogeny.

In contrast to the result of Huson and Steel, which required $n-1$ character states, our approach is completely independent of the number n of taxa and can therefore be realized with the four DNA character states on any set of species. With regard to the number of character states employed, Theorem 4.4 is best possible, that is, compatible binary characters alone would not generate the contrast between parsimony-based and distance-based trees. This is because the splits induced by these characters would be faithfully reconstructed from the derived distance function [38, Prop. 7.1.9].

Note that correcting for so-called ‘hidden changes’ under some model of sequence evolution would not eliminate the problem. This is due to the fact that for clocklike data, the corrections will not change the relative size of distances and the corrected sequence would still be treelike and conform to a molecular clock on the tree which is not the perfect phylogeny (for more details, see [38], Proposition 7.5.2).

Although ternary characters come up naturally with DNA data, Theorem 4.4 is of rather theoretical nature because such an amount of ternary characters necessary for split switching is out of reach with natural data (particularly when n is large). Especially in mitochondrial DNA data sets, ternary characters are sparse. Nonetheless, the presence of ternary characters with their different contributions to parsimony scores and distance values can yield a minor effect that would come on top of other noise patterns, which could adversely influence the reconstruction of distance trees and their interpretation as phylogenetic trees. Moreover, for small n, e.g. $n = 4$, we have shown in Chapter 4.3 that misleading sequences can be found in DNA data, so they do not only exist in theory. However, the misleading sequences found were relatively small and would therefore nowadays not be used for tree inference.

The purpose of Theorem 4.4 is not to state the importance of misleading sequences for tree reconstruction, but rather to show that even in the perfect – and unlikely – case of a homoplasy-free sequence whose derived distances are additive, the corresponding trees can differ. If such a scenario is possible even if the sequence is ‘best possible’ according to both the convexity as well as the additivity criterion, it is intuitively clear that tree reconstruction from DNA data that does incluce some homoplasies or corresponds to a network rather than a tree has to be conducted even more carefully. As Fischl has shown [13], the number of sequences giving rise to contradictory (‘incongruent’) trees is, as could be expected, far higher than that of misleading sequences.

In Section 4.2, we introduced an approach which allows for sequences on four taxa to be checked for misleadingness by just considering some equations and inequalities, but without the necessity of inferring the corresponding trees. We believe that this approach can easily be expanded to five or more taxa, but the number of equations and inequalities to consider will increase correspondingly. We conjecture that our characterization can be used to explicitly calculate the probability of the occurrence of misleading sequences for a given model of nucleotide substitution, which would help answer the question of how relevant misleading sequences really are. Moreover, the sequence constructed in Example 4.12 also has some interesting properties concerning the relationship of MP and ML, which will be highlighted in Chapter 5.2. In fact, the idea used in this chapter (namely constructing a disagreement between MP and distance methods by using the binary characters induced by the splits of the tree supposed to be the perfect phylogeny and then adding only parsimoniously uninformative characters) can also be used to construct a disagreement between MP and ML under certain model assumptions. This is due to the fact that ML, just like distance methods but opposed to MP, does not ignore parsimoniously uninformative characters.

Chapter 5

Equivalence of Maximum Parsimony and Maximum Likelihood

> You can't be suspicious of a tree (or accuse a bird or a squirrel of subversion or challenge the ideology of a violet).
>
> Hal Borland

In the previous chapter, the differences of Maximum Parsimony and distance-based tree reconstruction methods were analyzed. In the present chapter, MP will be compared to Maximum Likelihood (ML), which is another frequently used tree inference method. As explained in Chapter 2, a basic difference between these two methods is that MP, unlike ML, is not based on a specific nucleotide substitution model, and thus the methods may perform differently under different models: if the sequences under consideration are related by a specific model of substitution, the results of MP and ML coincide [17], but there are also examples, such as the famous 'Felsenstein Zone', for which this is not the case [10].

In 1997, Tuffley and Steel carried the analysis of MP and ML an important step further [44]: they showed that the N_r-model with 'no common mechanism' is sufficient for MP and ML to be equivalent when applied to a sequence of characters. Here, we will present a very elementary and short proof of the Tuffley-Steel result in order to make it more widely accessible. Along the way, we exploit some useful properties of the likelihood function, such as its multilinearity.

Moreover, the main purpose of this chapter is to analyze this equivalence of MP and ML further by considering slightly modified model assumptions that are of biological relevance. For instance, MP is often assumed to be justified whenever the nucleotide substitution probabilities are small (see, e.g., [12], p. 101). Therefore, we restrict the model by placing an upper bound on these probabilities, and find that under no common mechanism, MP and ML are no longer equivalent. Moreover, the equivalence of MP and ML under a 'no common mechanism model' also fails under the constraint of a molecular clock, even without a bound on the substitution probabilities. These two claims will be established by constructing counterexamples that are minimal with respect to the number of taxa. To construct our examples, we exploit a useful property of the likelihood function for a 'no common mechanism' model, namely that it is multilinear in the substitution probabilities. This fact underlies Equation 18 and Lemma 2 in [44], which we use in our arguments. We will show that some ideas which are useful to construct a disagreement of MP and ML are related to the ideas used in Chapter 4.

Additionally, we prove bounds on the ML-value of a given sequence of characters on a tree, and use them to show that it is possible to choose sufficiently small substitution probabilities (depending on the number of taxa, the number of characters and the number of states) so that every tree chosen by ML is also a most parsimonious tree.

5.1 Elementary Proof of the Tuffley-Steel Result

We begin with explicitly stating the equivalence of MP and ML in the following theorem. We first prove it for a single character, and then generalize the proof to a sequence of characters, assuming no common mechanism.

Theorem 5.1. (Tuffley and Steel 1997) *Let $m \in \mathbb{N}$ and $S := f_1 \dots f_m$ be a sequence of r-state characters on a phylogenetic X-tree $\mathcal{T}$. Then, under the N_r-model with no common mechanism, the maximum likelihood of S and its parsimony score are related by*

$$\max P(S|\mathcal{T}) = r^{-l_{\mathcal{T}}(S)-m}, \tag{27}$$

which implies that Maximum Likelihood and Maximum Parsimony both choose the same tree(s).

We first consider the case $m = 1$, i.e., a single character f. Thus, we first show

$$\max P(f) = r^{-l_{\mathcal{T}}(f)-1}. \tag{28}$$

Note that in the N_r-model, $P(f) = \frac{1}{r} \cdot P(f|f(1) = c_1)$, where c_1 is the character state assigned to leaf taxon 1 (for details on the so-called reversibility of the N_r-model, which is the reason why the actual root position does not matter, the reader is referred to [11]). Thus, in order to show (28) it is sufficient to show

$$\max P(f|f(1) = c_1) = r^{-l_{\mathcal{T}}(f)}. \tag{29}$$

We prove (29) by first showing with Lemma 5.2 that $r^{-l_{\mathcal{T}}(f)}$ is a lower bound for $\max P(f|f(1) = c_1)$. Then, Lemma 5.5 completes the proof by demonstrating that $r^{-l_{\mathcal{T}}(f)}$ is additionally an upper bound for $P(f|f(1) = c_1)$.

Lemma 5.2.

$$\max\ P(f|f(1) = c_1) \geq r^{-l_{\mathcal{T}}(f)}$$

Proof. Consider a most parsimonious extension g of f. Then, g requires exactly $l_{\mathcal{T}}(f)$ substitutions. We assign substitution probability $\frac{1}{r}$ to those edges on which a substitution occurs in g, and substitution probability 0 to all other edges. Then, $P(g|f(1) = c_1) = r^{-l_{\mathcal{T}}(f)}$. The lower bound for $\max P(f|f(1) = c_1)$ now follows from the fact that $P(f|f(1) = c_1) = \sum\limits_{g} P(g|f(1) = c_1)$. □

Lemma 5.2 states that $r^{-l_{\mathcal{T}}(f)}$ is a lower bound for $\max P(f|f(1) = c_1)$. In order to show that it is also an upper bound, we need some preliminaries concerning the likelihood function. These will be provided by Lemma 5.3 and Corollary 5.4.

Lemma 5.3. *Let h be a function from the k-dimensional box $B^k = [0,t]^k$ to the real numbers, i.e., $h : B^k \to \mathbb{R}$. If h is multilinear, then there is a corner p of B^k such that $h(p) \geq h(x)$ for every point x in B^k.*

Proof. We use an arbitrary point $x := (x_1, x_2, \ldots, x_k)$ in B^k as the initial value of a greedy hill climbing algorithm. Suppose we fix the values x_j for all j other than i. Since h is multilinear, $h(x)$ then takes the form $ax_i + b$. Now we distinguish between two cases:

1. $a \geq 0$. In this case, we replace x_i by t.
2. $a < 0$. In this case, we replace x_i by 0.

In either case, the value of $h(x)$ cannot decrease. Repeating this step for all $i = 1, \ldots, k$, one eventually arrives at a corner p of the box B^k, where $h(p) \geq h(x)$. The particular corner p obtained by the above procedure depends on the initial choice of x. Thus, as a last step we select a corner of B^k that maximizes h among all corners. This completes the proof. □

We are now in the position to make first statements on the likelihood function.

Corollary 5.4. *Let f be a character on a phylogenetic X-tree $\mathcal{T}$. Then under the N_r-model with all substitution probabilities bounded by u, where $0 \leq u \leq \frac{1}{r}$, the likelihood $P(f|\mathcal{T})$ can be maximized at a point where all substitution probabilities are either 0 or u.*

Proof. First, recall that

$$P(f|f(1) = c_1) = \sum_g P(g|f(1) = c_1), \tag{30}$$

where the summation is over all possible extensions g of f. Now let e be the edge $(1, v)$. Then the likelihood may be computed by the recursion

$$P(f|f(1) = c_1) = \sum_{g:g(v)=c_1} P(g|f(1) = c_1)q_e + \sum_{\substack{g:g(v)=s\neq c_1,\\ s\in\mathcal{C}}} P(g|f(1) = c_1)p_e, \qquad (31)$$

where p_e is the probability of a substitution on edge e, and $q_e = 1 - (r-1)p_e$ is the probability of no substitution on edge e. Clearly, Equation (31) is linear in each p_e. Therefore, the likelihood function $P(f|f(1) = c_1)$ is multilinear, and the claim follows from Lemma 5.3 and the fact that $0 \leq p_e \leq u$. □

Note that Corollary 5.4 is the same as Lemma 2 in [44] except that Tuffley and Steel stated their result only for $u = \frac{1}{r}$. However, this assumption is not used in their proof and therefore not required for the corollary to hold. We keep the corollary more general in order to enable an analysis of bounded substitution probabilities later in this chapter.

Lemma 5.5. *Let f be an r-state character on a phylogenetic X-tree $\mathcal{T}$. Then,*

$$P(f|f(1) = c_1) \leq r^{-l_{\mathcal{T}}(f)}.$$

Proof. In view of Corollary 5.4, there is always an ML-tree with the property that some of the substitution probabilities on its edges are $\frac{1}{r}$ and all other substitution probabilities are 0. Therefore, for an ML-tree $\mathcal{T}$ of a character f we can assume wlog. that $\mathcal{T}$ has this property. We partition the edge set $E(\mathcal{T})$ of this ML-tree into two sets E_1 and E_0, such that edges in E_1 have substitution probability $\frac{1}{r}$ and edges in E_0 have substitution probability 0. Let $k := |E_1|$.

If an extension g of f has a substitution on an edge e in E_0, then $P(g|f(1) = c_1) = 0$, i.e., g does not contribute to the likelihood calculation. Therefore, $k \geq l_{\mathcal{T}}(f)$, since any extension g of f has, by definition of the parsimony score $l_{\mathcal{T}}(f)$, at least $l_{\mathcal{T}}(f)$

substitutions, all of which must occur on edges in E_1.

If $P(g|f(1) = c_1) \neq 0$, then $P(g|f(1) = c_1) = (\frac{1}{r})^k$ by definition of E_1, as on these edges the substitution probabilities p_e as well as the probability q_e for no substitution are all $\frac{1}{r}$. Therefore, $P(f|f(1) = c_1) = \frac{N}{r^k}$, where N is the number of extensions g that have a non-zero likelihood. We now show that $N \leq r^{k-l_{\mathcal{T}}(f)}$. Figure 5.1 illustrates E_1 and E_0 by solid and dotted edges, respectively. The groups of vertices that are connected by edges of E_0 must be assigned the same state by any extension g of f that contributes to the likelihood, because there the substitution probabilities are 0. Note that for such extensions, substitutions can only occur on edges of E_1, but it is not required that on all such edges there is a substitution.

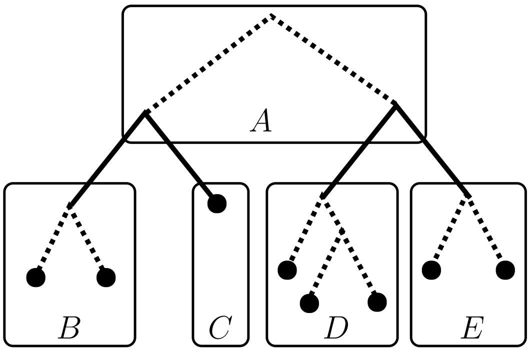


Figure 5.1: Maximal blocks that induce subtrees with dotted edges. Block A does not contain any leaf and is thus called 'unlabeled'. All other blocks are labeled.

Maximal subtrees consisting only of edges of E_0 are enclosed in boxes in Figure 5.1. We call the vertex sets of such subtrees *blocks*, and we call a block *labeled* whenever it contains a leaf. As explained before, any extension g of f that contributes to the likelihood $P(g|f(1) = c_1)$ only allows for changes on edges of E_1. Therefore, whenever a block contains a leaf vertex i, all vertices in this labeled block must be assigned the same state $f(i)$ by such an extension g.

Note that there are exactly $k+1$ blocks as $|E_1| = k$ (for instance, in Figure 5.1 there are four bold edges separating the vertex set into five blocks), and we define M to be the number of labeled ones. Then, for the parsimony score $l_{\mathcal{T}}(f)$ we know $l_{\mathcal{T}}(f) \leq M - 1$. This is true because even if all labeled blocks are in different states, MP chooses one of the leaf states to be the root state. Therefore, on the way from the root to at most all but one labeled blocks a change is required. Re-writing this condition gives a lower bound for M: $M \geq l_{\mathcal{T}}(f) + 1$. Thus, at least $l_{\mathcal{T}}(f) + 1$ of the $k + 1$ blocks are labeled and therefore at most $(k + 1) - (l_{\mathcal{T}}(f) + 1) = k - l_{\mathcal{T}}(f)$ are unlabeled. This implies that there are at most $r^{k-l_{\mathcal{T}}(f)}$ extensions g of f that contribute to the likelihood, i.e., $N \leq r^{k-l_{\mathcal{T}}(f)}$. This is due to the fact that each unlabeled block can be assigned one of the r different character states, whereas the state of the labeled blocks is fixed as f is given. Therefore, altogether we have $P(f|f(1) = c_1) = Nr^{-k} \leq r^{-l_{\mathcal{T}}(f)}$, which completes the proof. □

We are now in a position to prove Theorem 5.1.

Proof of Theorem 5.1. Combining Lemmas 5.2 and 5.5 yields Equation (29). Thus, for a single character f, MP and ML are equivalent. We now generalize this result for a character sequence S of no common mechanism. By definition of 'no common mechanism', the likelihood of each character can be maximized independently. Therefore,

$$\max\ P(S = f_1 \ldots f_m) = \prod_{i=1}^{m} \max\ P(f_i) \overset{\text{L.5.2 \& 5.5}}{=} \prod_{i=1}^{m} r^{-l_{\mathcal{T}}(f_i)-1} = r^{-l_{\mathcal{T}}(S)-m}.$$

This completes the proof. □

In the following, we show that small changes to the assumptions of the N_r-model with no common mechanism may suffice to let the equivalence of MP and ML fail. In particular, we analyze two settings of biological interest: first, we consider bounded substitution probabilities, and second, we investigate the case of a molecular clock. Both cases allow us to explicitly construct examples in which MP and ML choose different trees even under no common mechanism.

5.2 Bounded Substitution Probabilities

In this chapter, we consider a modification of the N_r-model with no common mechanism in which the substitution probabilities on all edges are bounded above by some $u < \frac{1}{r}$. We construct sequences of characters for which MP and ML choose different sets of trees.

Proposition 5.6. *Under the N_r-model with no common mechanism, for $r \geq 2$, there exist values of u such that if the substitution probabilities are bounded above by u, where $0 < u < \frac{1}{r}$, MP and ML choose different sets of trees. In particular, we have:*

1. *For $r = 2$, for all values of $u \in \left(0, 1 - \frac{1}{\sqrt{2}}\right)$, there exist sequences of characters for which MP and ML choose different sets of trees.*

2. *For $r > 2$, for all values of $u \in (0, \frac{1}{r})$ there exist sequences of characters for which MP and ML choose unique and distinct trees.*

Consider again Corollary 5.4. Recall that Tuffley and Steel stated it only for the case $u = \frac{1}{r}$ and used it to explicitly maximize the likelihood of a character on a given tree under the N_r-model: for a given character f and tree $\mathcal{T}$ with a most parsimonious extension g of f, assigning substitution probability $\frac{1}{r}$ to edges where a substitution is induced by g and 0 elsewhere gives $\max_{\bar{p}} P(f|\mathcal{T}, \bar{p})$, where $\bar{p}$ is the vector containing all substitution probabilities (cf. Theorem 3 of [44]).

But it turns out that an ML solution cannot be similarly related to an MP solution when $u < \frac{1}{r}$. That is, if g is a most parsimonious extension of a character f we may not be able to maximize the likelihood by simply assigning the substitution probability u to edges on which there is a substitution in g and 0 to edges on which there is no substitution in g. The likelihood may actually be maximized at some other corner of the feasibility region of $\bar{p}$. This is the idea of the following construction.

Proof of Proposition 5.6. We provide examples of sequences of characters for which MP and ML may choose different sets of trees. We first prove the case $r = 2$ with an example on five taxa, and show that in this case, there are no such examples on fewer than five taxa. Then we explicitly prove the case $r = 3$ with an example on four taxa and show how this example can be generalized for $r > 3$.

Case $r = 2$:

Let the set of character states be $\{\alpha, \beta\}$. Consider the two trees $\mathcal{T}_1$ and $\mathcal{T}_2$ shown in Figure 5.2 alongside the characters $f_1 = \alpha\alpha\beta\beta\beta$ and $f_2 = \alpha\beta\alpha\beta\beta$. We consider the character sequence $S := f_1 f_2$.

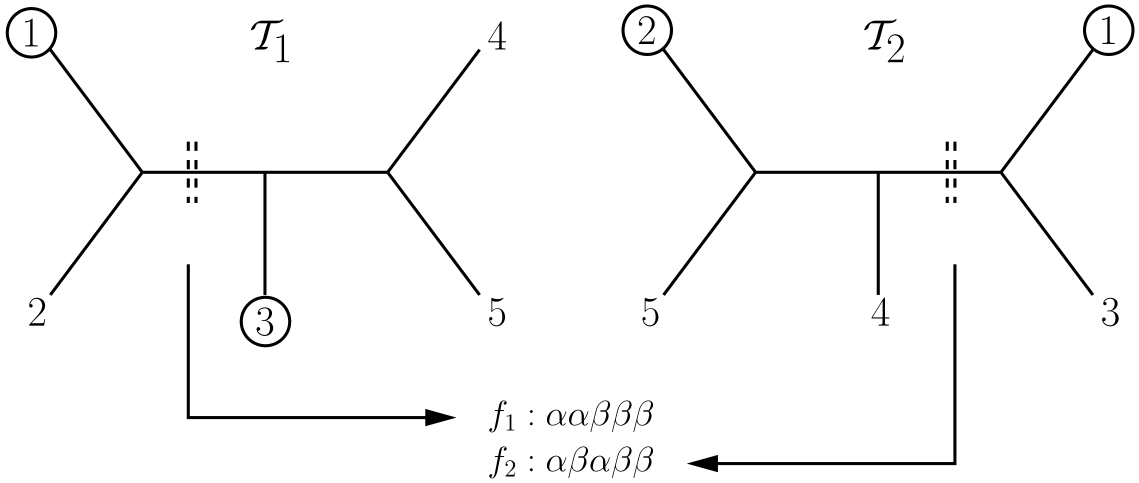

1
$\mathcal{T}_1$
4
2
3
5
2
$\mathcal{T}_2$
1
5
4
3
$f_1 : \alpha\alpha\beta\beta\beta$
$f_2 : \alpha\beta\alpha\beta\beta$

Figure 5.2: The characters f_1 and f_2 both correspond to a split on an interior edge of $\mathcal{T}_1$ or $\mathcal{T}_2$, respectively. But, as highlighted by the circled leaves, the assignment of f_1 on $\mathcal{T}_2$ differs from the assignment of f_2 to $\mathcal{T}_1$.

Note that $l_{\mathcal{T}_1}(f_1) = l_{\mathcal{T}_2}(f_2) = 1$ and $l_{\mathcal{T}_1}(f_2) = l_{\mathcal{T}_2}(f_1) = 2$. Therefore, $l_{\mathcal{T}_1}(S) = l_{\mathcal{T}_2}(S) = 3$, which means that MP will not favor either of the two trees $\mathcal{T}_1$, $\mathcal{T}_2$ over the other one. Moreover, as f_1 and f_2 are incompatible with one another, it can easily be seen that both trees are actually MP-trees: the minimal score of either character is 1, as two states are employed, and this score is achieved when the character corresponds to a split on an edge of the underlying tree – but because of the incompatibility, the other character will have a score of at least 2. So for S, a score of 3 is best possible, and thus both $\mathcal{T}_1$ and $\mathcal{T}_2$ are MP-trees.

For ML, the situation is different. This is because the assignments of f_1 on $\mathcal{T}_2$ and f_2 on $\mathcal{T}_1$ differ, as highlighted by Figure 5.2. In fact, character f_1 has a unique most parsimonious extension on $\mathcal{T}_2$, whereas f_2 has two most parsimonious extensions on $\mathcal{T}_1$. As we show in the following, for a sufficiently small upper bound u, the likelihood function is maximized when these extensions both contribute to the likelihood. We use the symbolic computer algebra system MAXIMA to evaluate $P(f_i|\mathcal{T},\bar{p})$ for $i = 1, 2$, for all trees on five taxa and at all corners of the feasibility region of $\bar{p}$ (see Corollary 5.4). More specifically, for the five-leaf-trees under investigation, there are seven edges to which either 0 or u can be assigned, which gives $2^7 = 128$ possible parameter vectors $\bar{p}$ at which the likelihood might be maximized. We observe that $\max P(f_1|\mathcal{T}_1) = \max P(f_2|\mathcal{T}_2) = \frac{1}{2}u$, but $\max P(f_1|\mathcal{T}_2) = \frac{1}{2}u^2$ and $\max P(f_2|\mathcal{T}_1) = \max(\frac{1}{2}u^2, u^2(1-u)^2)$. So there are choices of u, namely all $u < 1 - \frac{1}{\sqrt{2}}$, for which $\max P(f_1|\mathcal{T}_2) < \max P(f_2|\mathcal{T}_1)$. In these cases, even though both $\mathcal{T}_1$ and $\mathcal{T}_2$ are MP-trees, ML will favor tree $\mathcal{T}_1$ over $\mathcal{T}_2$. Therefore, MP and ML are not equivalent in this case.

Now let sequence $\tilde{S}$ contain k copies of character f_1 and $k+1$ copies of character f_2 for some integer $k > 0$. Then, clearly $l_{\mathcal{T}_1}(\tilde{S}) = 3k+2$, but $l_{\mathcal{T}_2}(\tilde{S}) = 3k+1$. Therefore, MP will favor tree $\mathcal{T}_2$ over $\mathcal{T}_1$. Moreover, $\mathcal{T}_2$ is an MP-tree (by the same incompatibility argument concerning f_1 and f_2 as above). On the other hand, we have $\max P(\tilde{S}|\mathcal{T}_1) = (\frac{1}{2}u)^k \cdot \left(u^2(1-u)^2\right)^{k+1}$ and $\max P(\tilde{S}|\mathcal{T}_2) = \frac{u^{3k+1}}{2^{2k+1}}$ (provided $u < 1 - \frac{1}{\sqrt{2}}$). We choose a sufficiently large value of k so that the former value is larger than the latter. For such choices of k, ML will favor tree $\mathcal{T}_1$ over $\mathcal{T}_2$, even though MP favors $\mathcal{T}_2$. It is important to note, however, that for the sequence $\tilde{S}$, the tree $\mathcal{T}_1$ is not an ML-tree. It can be easily verified for the tree $\mathcal{T}_3$ in Figure 5.3 that $\max P(\tilde{S}|\mathcal{T}_3) = \left(\frac{u}{2}\right)^{k+1}\left(u^2(1-u)^2\right)^k$, which is more than $\max P(\tilde{S}|\mathcal{T}_1)$. In fact, $\max P(\tilde{S}|\mathcal{T}_3) > \max P(\tilde{S}|\mathcal{T}_1)$ for all $u \leq \frac{1}{2}$. So $\mathcal{T}_3$ is the unique ML-tree. Moreover, $\mathcal{T}_3$ is also an MP-tree. So for $r = 2$, it remains unclear whether MP and ML can make strictly conflicting choices.

Note that when $r = 2$, examples demonstrating the inequivalence of MP and ML

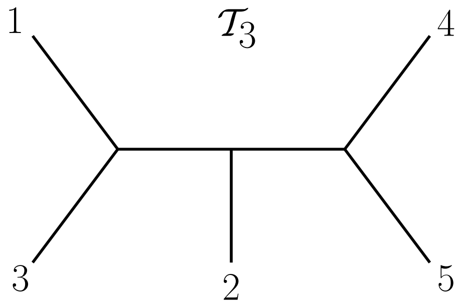


Figure 5.3: Tree $\mathcal{T}_3$ is both an MP- and an ML-tree for sequence $\tilde{S}$

cannot be constructed with fewer than five taxa. This is because given at most one interior edge, it can be easily checked that all non-informative binary characters have the same maximum probability on all trees, whereas informative binary characters on four taxa have a higher probability on the tree where they have parsimony score 1.

Case $r = 3$:

Let the set of character states be $\{\alpha, \beta, \gamma\}$. We consider four taxa and the characters $f_1 := \alpha\alpha\beta\beta$ and $f_2 := \alpha\beta\gamma\beta$, as well as the sequence S of characters defined by $S := f_1 \underbrace{f_2 \dots f_2}_{k \text{ times}}$ for some $k \in \mathbb{N}$. Two of the three possible trees on four taxa are shown in Figure 5.4: the tree $\mathcal{T}_4 = 12|34$ and the tree $\mathcal{T}_5 = 13|24$.

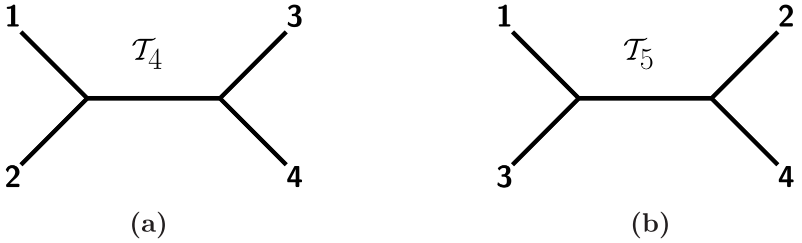


Figure 5.4: Tree $\mathcal{T}_4$ illustrated in (a) is the unique MP-tree for S, whereas (b) depicts tree $\mathcal{T}_5$, which is the unique ML-tree for S when, say, $u = \frac{1}{8}$ is chosen.

Tree $\mathcal{T}_4$ is clearly the unique MP-tree of S, as the only informative character in S is $f_1 = \alpha\alpha\beta\beta$.

As before, we use the symbolic algebra system MAXIMA to evaluate $P(f|\mathcal{T}, \bar{p})$ for all characters f in the sequence, for all trees on four taxa and at all corners of the feasibility region of $\bar{p}$. We observed that $\max P(f_2|\mathcal{T}_4) = \frac{u^2}{3}$ and $\max P(f_2|\mathcal{T}_5) = u^2(1-2u)$. Therefore, for all $u < \frac{1}{r}$, we have $\max P(f_2|\mathcal{T}_5) > \max P(f_2|\mathcal{T}_4)$. Now for any $u < \frac{1}{r}$, a sufficiently large value of k may be chosen such that $\frac{\max P(S|\mathcal{T}_5)}{\max P(S|\mathcal{T}_4)} > 1$. We do not analyze the character f_1, although the actual choice of k will depend on the ratio $\frac{\max P(f_1|\mathcal{T}_5)}{\max P(f_1|\mathcal{T}_4)}$ and on u. Therefore, MP and ML choose different trees in this three-state setting. Moreover, it can be easily verified that for the third topology on four taxa, namely $\mathcal{T}_6 = 14|23$, we have $\max P(S|\mathcal{T}_6) < \max P(S|\mathcal{T}_5)$ for all choices of $u \leq \frac{1}{3}$. So, $\mathcal{T}_5$ is the unique ML-tree, whereas $\mathcal{T}_4$ is the unique MP-tree in this setting. So MP and ML make strictly conflicting choices.

Case $r > 3$:

Let the set of states be $\mathcal{C} := \{\alpha, \beta, \gamma, \delta_1, \delta_2, \ldots, \delta_{r-3}\}$. Let $\mathcal{D} := \{\delta_1, \delta_2, \ldots, \delta_{r-3}\}$. We analyze four taxa and the same characters $f_1 := \alpha\alpha\beta\beta$ and $f_2 := \alpha\beta\gamma\beta$ that were analyzed in the case $r = 3$, but this time under the N_r-model with $r > 3$. Again we consider the sequence of characters $S := f_1 \underbrace{f_2 \ldots f_2}_{k \text{ times}}$.

We only sketch the proof in this case. In particular, we indicate how the expressions for the likelihood function may be written regardless of the number of states.

The expressions for $P(f_i|\mathcal{T}_j)$ for $i = 1, 2$ and $j = 1, 2, 3$ can be written in a simple manner since the states δ_i do not occur in S. For example, let the substitution probabilities on the edges of a four-taxa tree $\mathcal{T}$ be $\bar{p} = (p_i, i = 1, 2, \ldots, 5)$, where p_i (for $i \in \{1, 2, 3, 4\}$) are the substitution probabilities on the pending edges adjacent to taxa 1, 2, 3, 4, respectively, and p_5 is the substitution probability on the internal edge. Let u and v be the internal vertices of $\mathcal{T}$. We write $P(f_i|\mathcal{T}, \bar{p}) = \sum_g P(g|\mathcal{T}, \bar{p})$, where the summation is over all extensions g of f_i.

Now observe that if g and h are two extensions of a character f, then we have $P(g|\mathcal{T}, \bar{p}) = P(h|\mathcal{T}, \bar{p})$ if $g(u), h(u) \in \mathcal{D}$ and $g(v) = h(v) = s \notin \mathcal{D}$ (or vice versa with the roles of u and v interchanged).

Therefore:

$$\sum_{\substack{g\,:\,g(u)\,\in\,\mathcal{D},\\ g(v)\,=\,s\,\notin\,\mathcal{D}}} P(g|\mathcal{T},\bar{p}) = (r-3)P(h|\mathcal{T},\bar{p}),$$

where h is an extension of f for which $h(u) = \delta_1$ and $h(v) = s$.

Similarly:

$$\sum_{\substack{g\,:\,g(u)\,=\,s\,\notin\,\mathcal{D},\\ g(v)\,\in\,\mathcal{D}}} P(g|\mathcal{T},\bar{p}) = (r-3)P(h|\mathcal{T},\bar{p}),$$

where h is an extension of f for which $h(u) = s$ and $h(v) = \delta_1$.

Finally:

$$\sum_{\substack{g\,:\,g(u)\,\in\,\mathcal{D},\\ g(v)\,\in\,\mathcal{D}}} P(g|\mathcal{T},\bar{p}) = (r-3)(1-3p_5)p_1p_2p_3p_4,$$

With these observations, it is possible to write the expressions for computing $P(f_i|T_j)$ in a computer algebra system like MAXIMA. As in the case $r = 3$, we analyzed only $P(f_2|\mathcal{T}_4)$ and $P(f_2|\mathcal{T}_5)$, and verified that $\max P(f_2|\mathcal{T}_5) \geq \frac{u^2(3-2ru)}{r}$ and $\max P(f_2|\mathcal{T}_4) = \frac{u^2}{r}$. Since $(3-2ru) > 1$ for all $u < \frac{1}{r}$, there is a k for which $\max P(S|\mathcal{T}_5) > \max P(S|\mathcal{T}_4)$. This means that ML will favor $\mathcal{T}_5$ over $\mathcal{T}_4$, even though $\mathcal{T}_4$ is the unique MP-tree in this setting. □

Remark 1. It is important to state that in the examples for $r \geq 3$ introduced in the proof of Proposition 5.6, where the number of taxa is bounded (in fact, it is only 4), as u approaches $\frac{1}{r}$, we require the sequence length $k+1$, and thus k, to tend to infinity for ML and MP to make different choices. However, this is a necessary property of any such example for which the number of taxa is bounded: For any fixed character sequence S, the continuity of the likelihood function and the Tuffley-Steel result (Theorem 5.1) imply that there is a positive real number $\epsilon(S)$ such that if $u > \frac{1}{r} - \epsilon(S)$, then ML and MP choose the same sets of trees. Therefore, for a bounded number of taxa, since there are only finitely many sequences of length at most $k+1$, we set $\epsilon := \min_S(\epsilon(S))$, where the minimization takes place over all character sequences of length at most $k+1$, and conclude that MP and ML would be equivalent (in the sense of the Tuffley-Steel result) for all $u > \frac{1}{r} - \epsilon$, for all sequences of length at most $k+1$. Therefore, as u approaches $\frac{1}{r}$, the sequence length $k+1$ of sequences for which MP and ML make conflicting choices has to tend to infinity.

Note that the role of the N_2-model with no common mechanism in Proposition 5.6 is different than that of the N_r-model with no common mechanism for $r > 2$: for $r = 2$, we only state that the sets of trees chosen by MP and ML may be different for certain choices of values for the upper bound u. In particular, our proof shows that some MP-trees may not be ML-trees in this setting, but it is still unknown if the opposite can also happen. The intuitive reason why the binary case might indeed be different is due to the relationship of sequence $S = f_1 f_2 \dots f_2$ as used in Propositon 5.6 and the (distance-wise) misleading sequences as introduced in Chapter 4. In fact, $S = f_1 f_2 f_2 f_2$ (i.e. the particular case with three copies of character f_2) is a shorter version of the sequence constructed in Chapter 4.2; only two characters are missing, namely $\alpha\beta\gamma\alpha$ and $\alpha\beta\beta\gamma$ – and trivially, the ML-values of these characters on $\mathcal{T}_4$ and $\mathcal{T}_5$ (depicted in Figure 5.4) are equal (as the pattern of both characters is equal on both trees). It can be easily verified that the ML-value of each of these characters on both trees is $\frac{u^2}{3}$. So adding these characters to the sequence used in Proposition 5.6 does not change the example: in fact, this sequence, which was constructed in Chapter 4.2 as a case in which the perfect phylogeny and the tree induced by the derived distances differ, is also an example for a sequence where a certain choice of an upper bound u will make MP and ML differ (but note that the shorter version of this sequence, which for certain values of u causes a disagreement of MP and ML, is *not* a misleading sequence concerning the Hamming distance as it is not even treelike). In fact, for these choices of u, e.g. $u = \frac{1}{8}$ as shown in the proof of Proposition 5.6, ML will agree with the tree on which the Hamming distances are treelike, namely $\mathcal{T}_2$, whereas MP will naturally favor $\mathcal{T}_1$ as the only informative character in the sequence is $f_1 = \alpha\alpha\beta\beta$.

We now complement the above inequivalence results by showing that for sufficiently small choices of u, all ML-trees are also MP-trees. To prove this result, we first establish lower and upper bounds for the maximum probability of observing a character given a tree.

Proposition 5.7. *Let $\mathcal{T}$ be a phylogenetic X-tree, where $|X| = n$. Let f be a character on X. Then under the N_r-model with no common mechanism and with all substitution probabilities bounded by u, where $0 \leq u < \frac{1}{r}$, we have*

$$\left(\frac{1}{r}\right) u^{l_{\mathcal{T}}(f)} \leq \max P(f|\mathcal{T}) \leq r^{m-3} u^{l_{\mathcal{T}}(f)},$$

where $\max P(f|\mathcal{T})$ *is the maximum likelihood of f on $\mathcal{T}$ and $l_{\mathcal{T}}(f)$ is the parsimony score of f on $\mathcal{T}$.*

Proof. For the lower bound, just as in the Tuffley-Steel approach explained above, we take a most parsimonious extension g of f and assign substitution probability u to each edge that has a substitution in g and 0 to all other edges. Considering the r possible root states (for an arbitrarily chosen root), this gives the lower bound for $\max_{\bar{p}} P(f|\mathcal{T}, \bar{p})$.

To prove the upper bound, we observe that there are exactly r^{n-2} extensions of f, where $n-2$ is the number of internal vertices, each of which may be assigned any of the r states. We will now analyze these extensions. Let g be any extension of f. For any assignment of substitution probabilities to the edges of the tree, the value of $P(g|\mathcal{T}, \bar{p})$ for an assignment of probabilities that maximizes $P(f|\mathcal{T}, \bar{p})$ is either 0 (if one of the edges on which there is a substitution in g has been assigned a substitution probability 0) or is given by

$$\max_{\bar{p}} P(g|\mathcal{T}, \bar{p}) = \frac{1}{r} u^{k_1} (1 - (r-1)u)^{k_2} \leq \frac{1}{r} u^{k_1} \leq \frac{1}{r} u^{l_{\mathcal{T}}(f)}, \tag{32}$$

where $k_1 \geq l_{\mathcal{T}}(f)$ is the number of edges on which there is a substitution in g and k_2 is the number of edges which require no substitution in g but have been assigned substitution probability u. The factor $\frac{1}{r}$ is due to the r possible choices for the root state.

By Equation (32), each extension of f has a likelihood of at most $\frac{1}{r}u^{l_{\mathcal{T}}(f)}$. The upper bound now follows by summing the likelihoods of all extensions. □

Now we will use the bounds for the maximum likelihood to derive the desired conclusion on ML-trees.

Theorem 5.8. *Let $S = f_1 f_2 \dots f_m$ be a character sequence and let $\mathcal{T}_a$ and $\mathcal{T}_b$ be two phylogenetic X-trees, where $|X| = n$. Then under the N_r-model with no common mechanism and with all substitution probabilities bounded by u, where $0 \leq u < \frac{1}{r}$, we have for sufficiently small choices of u*

$$l_{\mathcal{T}_b}(S) < l_{\mathcal{T}_a}(S) \Rightarrow \max P(S|\mathcal{T}_b) > \max P(S|\mathcal{T}_a),$$

i.e., if $\mathcal{T}_b$ is parsimoniously better than $\mathcal{T}_a$, it also has a better ML-value.

Proof. By Proposition 5.7 we have

$$\max P(S|\mathcal{T}_a) \leq r^{(n-3)m} u^{\sum_i l_{\mathcal{T}_a}(f_i)} \tag{33}$$

and

$$\max P(S|\mathcal{T}_b) \geq \left(\frac{1}{r}\right)^m u^{\sum_i l_{\mathcal{T}_b}(f_i)}. \tag{34}$$

Note that for any positive integers a and b such that $b < a$ and any positive constant c, for sufficiently small values of u we have $u^a < cu^b$. Now by assumption we have $b := \sum_i l_{\mathcal{T}_b}(f_i) < \sum_i l_{\mathcal{T}_a}(f_i) =: a$. With $c := \min(1, (r^{n+(n-3)n})^{-1})$, using this argument for Equations (33) and (34), we get for sufficiently small values of u that $\max P(S|\mathcal{T}_b) > \max P(S|\mathcal{T}_a)$. □

Corollary 5.9. *Let S be sequence of m characters on a set of n taxa. Then there is an $\epsilon = \epsilon(n, m, r)$ such that under the N_r-model with no common mechanism and with all substitution probabilities subject to an upper bound $u \in [0, \epsilon)$, all ML-trees of S are also MP-trees.*

Proof. Let ϵ be 'sufficiently small' with respect to Theorem 5.8, and let $\mathcal{T}_b$ be any MP-tree for S. Let $\mathcal{T}_a$ be another tree that is not a most parsimonious tree. That is, $l_{\mathcal{T}_b}(S) < l_{\mathcal{T}_a}(S)$. Therefore, by Theorem 5.8, for $u \in [0, \epsilon)$, we have $\max P(S|\mathcal{T}_b) > \max P(S|\mathcal{T}_a)$. Thus, $\mathcal{T}_a$ cannot be an ML-tree, implying that all ML-trees are also MP-trees. □

5.3 Molecular Clock

We now prove a statement similar to Proposition 5.6, but instead of considering bounded substitution probabilities, we analyze substitution probabilities which conform to a molecular clock. Moreover, we consider only the three-state symmetric model. Under the N_3-model, we consider two cases (for any phylogenetic X-tree rooted at some root ρ and conforming to a molecular clock). In the first case, the probability of a substitution from the root to any of its leaves is bounded above by $p_{max} \in [0, \frac{1}{3})$. In the second case, the probability of a substitution from the root to any of its leaves is bounded above by $p_{max} = \frac{1}{3}$ (as suggested by the N_3-model).

Proposition 5.10. *Under the N_3-model with no common mechanism, with the substitution probabilities constrained by a molecular clock, MP and ML are not equivalent for any bound $p_{max} \in [0, \frac{1}{3}]$.*

Proof. We show the inequivalence first for $p_{max} \in [0, \frac{1}{3})$, and then for $p_{max} = \frac{1}{3}$.

Consider the two rooted trees $\mathcal{T}_1$ and $\mathcal{T}_2$ along with substitution probabilities p_i and $\tilde{p}_i$, respectively, on their edges as shown in Figure 5.5. The trees have the same shape but different leaf labels, and possibly different probabilities of a substitution from the root to any of its leaves. Under a molecular clock, we have $p_1 = p_2$ and $p_3 = p_4$ in $\mathcal{T}_1$, and $\tilde{p}_1 = \tilde{p}_3$ and $\tilde{p}_2 = \tilde{p}_4$ in $\mathcal{T}_2$. Let $p, \tilde{p} \in [0, p_{max}]$ be the probabilities of a substitution from the root ρ to any leaf in $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively.

Then under the N_3-model, we write p and $\tilde{p}$ in terms of the substitution probabilities on the edges of the trees as follows:

$$\begin{aligned} p &= (1-2p_5)p_1 + p_5(1-2p_1) + p_5p_1 = p_1 + p_5 - 3p_1p_5 \\ &= (1-2p_6)p_3 + p_6(1-2p_3) + p_6p3 = p_3 + p_6 - 3p_3p_6. \end{aligned}$$

Thus $p_5 = \frac{p-p_1}{1-3p_1}$ and $p_6 = \frac{p-p_3}{1-3p_3}$. Similarly, on $\mathcal{T}_2$, we have $\tilde{p}_5 = \frac{\tilde{p}-\tilde{p}_1}{1-3\tilde{p}_1}$ and $\tilde{p}_6 = \frac{\tilde{p}-\tilde{p}_2}{1-3\tilde{p}_2}$.

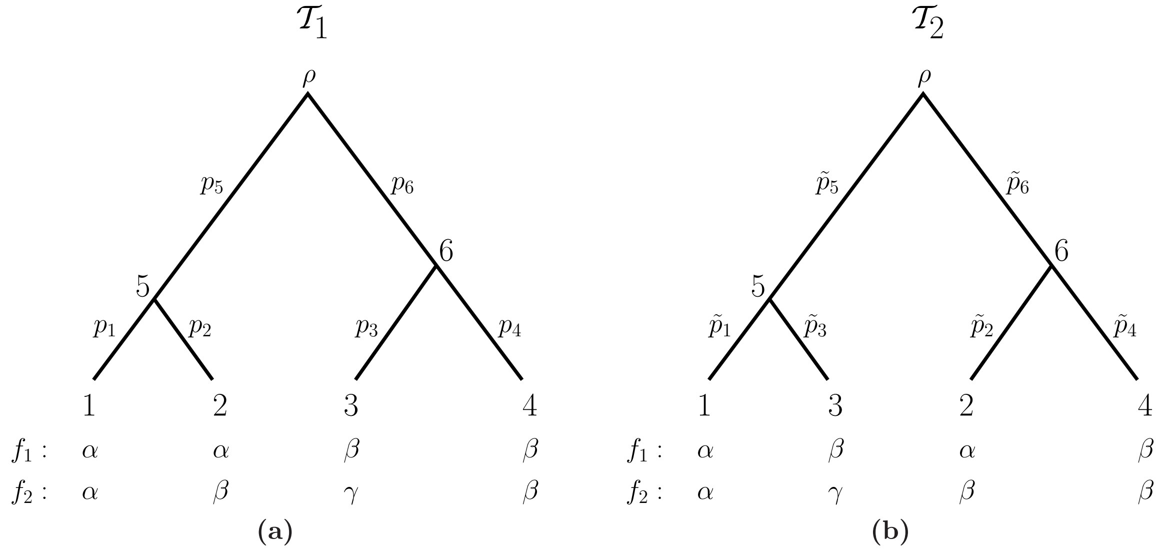


Figure 5.5: Rooted binary trees $\mathcal{T}_1$ and $\mathcal{T}_2$, which conform to a molecular clock, and the assignment of characters $f_1 = \alpha\alpha\beta\beta$ and $f_2 = \alpha\beta\gamma\beta$, respectively.

As in the proof of Proposition 5.6, we consider the N_3-model with state space $\mathcal{C} := \{\alpha, \beta, \gamma\}$. Consider the characters $f_1 := \alpha\alpha\beta\beta$ and $f_2 := \alpha\beta\gamma\beta$, and a sequence of characters $S := f_1 \underbrace{f_2 \ldots f_2}_{n \text{ times}}$, where n is a positive integer.

As before, $\mathcal{T}_1$ is the unique MP-tree of S. We claim that $\mathcal{T}_1$ is not an ML-tree if n is sufficiently large. In order to show this, we show that $\max P(S|\mathcal{T}_2) > \max P(S|\mathcal{T}_1)$ for a suitable choice of n.

Since we assume no common mechanism, we have

$$\frac{\max P(S|\mathcal{T}_2)}{\max P(S|\mathcal{T}_1)} = \frac{(\max P(f_1|\mathcal{T}_2))(\max P(f_2|\mathcal{T}_2))^n}{(\max P(f_1|\mathcal{T}_1))(\max P(f_2|\mathcal{T}_1))^n}.$$

We now demonstrate, without calculating $\max P(f_1|\mathcal{T}_1)$ and $\max P(f_1|\mathcal{T}_2)$ explicitly (but using the fact that these values are positive), that $\max P(f_2|\mathcal{T}_2) > \max P(f_2|\mathcal{T}_1)$ for all values of p_{max}. This allows us to choose a sufficiently large value of n so that the ratio above is more than 1.

First we seek to maximize:

$$P(f_2|\mathcal{T}_1, \bar{p}) = \sum_{c \in \mathcal{C}} P(f_2|\mathcal{T}_1, \bar{p}, \rho = c)P(\rho = c) = \frac{1}{3} \sum_{c \in \mathcal{C}} P(f_2|\mathcal{T}_1, \bar{p}, \rho = c).$$

Using the computer algebra system MAXIMA, we expand the right-hand side of this equation by summing the probabilities over all possible assignments of states to the internal nodes 5 and 6, and substitute $p_5 = \frac{p-p_1}{1-3p_1}$ and $p_6 = \frac{p-p_3}{1-3p_3}$ to obtain:

$$P(f_2|\mathcal{T}_1, \bar{p}) = \frac{p_1 p_3 (3p_1 p_3 - 2p_1 - 2p_3 + 1 + 2p - 3p^2)}{3}.$$

Observe that for any fixed values of $p_1 + p_3$ and p, the expression above is maximized when $p_1 p_3$ is maximized, i.e. when $p_1 = p_3$. Therefore, we can substitute p_1 for p_3 and maximize the resulting expression given by:

$$\frac{p_1^2(1 - p - p_1)(1 + 3p - 3p_1)}{3}.$$

Under the constraint $p \in [0, p_{max}]$, straightforward arguments show that the expression shown above has a maximum at $p_1 = p = p_{max}$, and the maximum is given by $\frac{p_{max}^2(1-2p_{max})}{3}$. Therefore:

$$\max P(f_2|\mathcal{T}_1) = \frac{p_{max}^2(1-2p_{max})}{3}. \tag{35}$$

Similar calculations show that:

$$\max P(f_2|\mathcal{T}_2) = p_{max}^2(1-2p_{max}), \tag{36}$$

where the maximum is obtained by setting $\tilde{p}_2 = \tilde{p}_4 = \tilde{p}_5 = 0$ and $\tilde{p}_1 = \tilde{p}_3 = \tilde{p}_6 = \tilde{p} = p_{max}$.

Equations (35) and (36) imply that $\max P(f_2|\mathcal{T}_2) > \max P(f_2|\mathcal{T}_1)$ for all $p_{max} \in (0, \frac{1}{3}]$. Now we can select a sufficiently large value of n so that $\max P(S|\mathcal{T}_1) < \max P(S|\mathcal{T}_2)$, where the actual choice of n will depend on the ratio $\frac{\max P(f_1|\mathcal{T}_2)}{\max P(f_1|\mathcal{T}_1)}$ and p_{max}.

This analysis does not show that $\mathcal{T}_2$ is an ML-tree, but it shows that $\mathcal{T}_1$, which is a unique MP-tree, is not an ML-tree. Therefore, the two methods are not equivalent under the constraint of a molecular clock, even when we assume no common mechanism.

Note that the relationship $\max P(f_2|\mathcal{T}_1) < \max P(f_2|\mathcal{T}_2)$ holds even when $p_{max} = \frac{1}{3}$. In particular, we observe that $P(f_2|\mathcal{T}_2, \tilde{p}_i, i = 1, 2, \ldots, 6)$ takes a maximum value of $\frac{1}{27}$ in the limit as $p \to \frac{1}{3}$, and $P(f_2|\mathcal{T}_1, p_i, i = 1, 2, \ldots, 6)$ takes a maximum value of $\frac{1}{81}$ in the limit as $\tilde{p} \to \frac{1}{3}$. □

5.3.1 Analysis of Binary Characters under a Molecular Clock

Just as in the previous chapter, the question arises whether an example for which MP and ML make strictly conflicting choices can also be found for binary characters. Using the same arguments as before, in the following we provide an example which shows that in N_2-model, there are sequences for which some MP-trees are not ML-trees, but as before we do not know if an example vice versa does also exist.

Example 5.11. Consider the two rooted trees shown in Figure 5.6. As before, for all $i = 1, \ldots, 8$, we denote the substitution probabilities on edge e_i with p_i and we define $P_i = 1 - 2p_i$. Note that due to the clock restriction, $P_2 = P_6$, $P_5 = P_8$ and $P_7 = P_4 P_5$. In this notation, the probability for a substitution on edge i can be written as $p_i = \frac{1-P_i}{2}$ and the probability q_i for no substitution on edge i is $q_i = \frac{1+P_i}{2}$. Furthermore, we call $P := P_1 P_2 = P_3 P_4 P_5$ the height of the tree. Thus the probability of a state change from the root to any leaf is $\frac{1-P}{2}$.

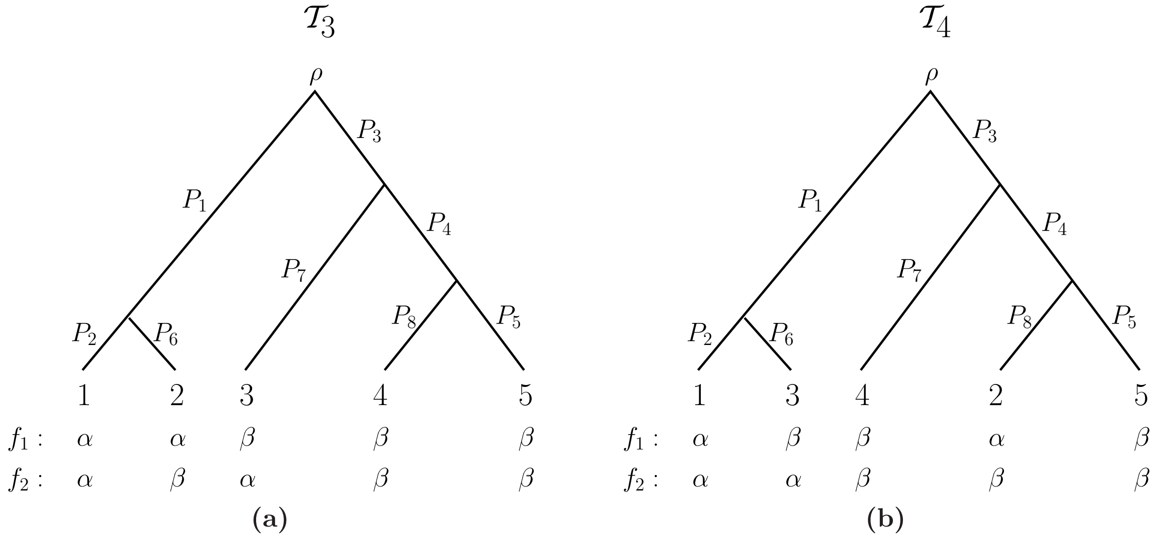

$\mathcal{T}_3$
ρ
P_1
P_2
P_3
P_4
P_5
P_6
P_7
P_8
1
2
3
4
5
f_1 : α α β β β
f_2 : α β α β β
(a)
$\mathcal{T}_4$
ρ
P_1
P_2
P_3
P_4
P_5
P_6
P_7
P_8
1
3
4
2
5
f_1 : α β β α β
f_2 : α α β β β
(b)

Figure 5.6: Rooted binary trees $\mathcal{T}_3$ and $\mathcal{T}_4$, which conform to a molecular clock, and the assignment of characters $f_1 = \alpha\alpha\beta\beta\beta$ and $f_2 = \alpha\beta\alpha\beta\beta$, respectively.

As in the proof of Proposition 5.6, we consider the characters $f_1 = \alpha\alpha\beta\beta\beta$ and $f_2 = \alpha\beta\alpha\beta\beta$ as well as the sequence $S := f_1 f_2$ in a 2-state setting, i.e., $r = 2$. As before, we get $l_{\mathcal{T}_3}(S) = l_{\mathcal{T}_4}(S) = 3$, which shows that MP will not favor either of the two trees $\mathcal{T}_3$, $\mathcal{T}_4$ over the other one. Also, both $\mathcal{T}_3$ and $\mathcal{T}_4$ are among the most parsimonious trees, since each of the characters f_1 and f_2 has a parsimony score of at least 1 on any tree, and as they are incompatible with one another, S must have a parsimony score of at least 3 on any tree.

We claim that ML favors tree $\mathcal{T}_3$ over $\mathcal{T}_4$. In order to show this, we first compute $\max P(f_1|\mathcal{T}_4)$ and then show that $\max P(f_2|\mathcal{T}_3) > \max P(f_1|\mathcal{T}_4)$.

Let P denote a given height and let f_1^l be the restriction of f_1 on the left subtree of

$\mathcal{T}_4$, and f_1^r the restriction of f_1 on the right subtree of $\mathcal{T}_4$. We have

$$\begin{aligned} P(f_1|\mathcal{T}_4) &= P(f_1|\mathcal{T}_4, \rho = \alpha)P(\rho = \alpha) + P(f_1|\mathcal{T}_4, \rho = \beta)P(\rho = \beta) \\ &= \frac{1}{2}P(f_1^l|\mathcal{T}_4, \rho = \alpha)P(f_1^r|\mathcal{T}_4, \rho = \alpha) + \frac{1}{2}P(f_1^l|\mathcal{T}_4, \rho = \beta)P(f_1^r|\mathcal{T}_4, \rho = \beta) \\ &= \frac{1}{2}P(f_1^l|\mathcal{T}_4, \rho = \alpha)\left(P(f_1^r|\mathcal{T}_4, \rho = \alpha) + P(f_1^r|\mathcal{T}_4, \rho = \beta)\right), \end{aligned} \tag{37}$$

since $P(f_1^l|\mathcal{T}_4, \rho = \alpha) = P(f_1^l|\mathcal{T}_4, \rho = \beta)$. Moreover, to obtain the ML value, the two factors $P(f_1^l|\mathcal{T}_4, \rho = \alpha)$ and $(P(f_1^r|\mathcal{T}_4, \rho = \alpha) + P(f_1^r|\mathcal{T}_4, \rho = \beta))$ can be maximized independently.

Evaluating the left factor $P(f_1^l|\mathcal{T}_4, \rho = \alpha)$ yields

$$\begin{aligned} P(f_1^l|\mathcal{T}_4, \rho = \alpha) &= \left(\frac{1+P_1}{2}\right)\left(\frac{1-P_2}{2}\right)\left(\frac{1+P_2}{2}\right) + \left(\frac{1-P_1}{2}\right)\left(\frac{1-P_2}{2}\right)\left(\frac{1+P_2}{2}\right) \\ &= \frac{1-P_2^2}{4}, \end{aligned}$$

which is maximized when P_2 takes the minimum possible value. Since $P = P_1P_2$ is given (and fixed), P_2 cannot be less than P (otherwise, P_1 would have to be more than 1, which is not possible). Therefore,

$$\max P(f_1^l|\mathcal{T}_4, \rho = \alpha) = \frac{1-P^2}{4}, \tag{38}$$

and the maximum is obtained when $P_1 = 1$ (since $P_1P_2 = P$). In other words, the substitution probability $p_1 = 0$, so the left subtree is rooted at ρ by a 0-length edge.

Analyzing the right factor $(P(f_1^r|\mathcal{T}_4, \rho = \alpha) + P(f_1^r|\mathcal{T}_4, \rho = \beta))$ yields

$$\begin{aligned}
& P(f_1^r|\mathcal{T}_4, \rho = \alpha) + P(f_1^r|\mathcal{T}_4, \rho = \beta) \\
= & \left(\frac{1+P_3}{2}\right)\left(\frac{1-P_4P_5}{2}\right)\left(\left(\frac{1+P_4}{2}\right)\left(\frac{1-P_5^2}{4}\right)+\left(\frac{1-P_4}{2}\right)\left(\frac{1-P_5^2}{4}\right)\right) \\
+ & \left(\frac{1-P_3}{2}\right)\left(\frac{1+P_4P_5}{2}\right)\left(\left(\frac{1-P_4}{2}\right)\left(\frac{1-P_5^2}{4}\right)+\left(\frac{1+P_4}{2}\right)\left(\frac{1-P_5^2}{4}\right)\right) \\
+ & \left(\frac{1-P_3}{2}\right)\left(\frac{1-P_4P_5}{2}\right)\left(\left(\frac{1+P_4}{2}\right)\left(\frac{1-P_5^2}{4}\right)+\left(\frac{1-P_4}{2}\right)\left(\frac{1-P_5^2}{4}\right)\right) \\
+ & \left(\frac{1+P_3}{2}\right)\left(\frac{1+P_4P_5}{2}\right)\left(\left(\frac{1-P_4}{2}\right)\left(\frac{1-P_5^2}{4}\right)+\left(\frac{1+P_4}{2}\right)\left(\frac{1-P_5^2}{4}\right)\right) \\
= & \left(\frac{1-P_5^2}{4}\right), \qquad (39)
\end{aligned}$$

which is maximized when P_5 is minimum (that is $P_5 = P$) and $P_3 = P_4 = 1$. Combining Equations (37), (38) and (39) yields

$$\max P(f_1|\mathcal{T}_4) = \frac{1}{2}\left(\frac{1-P^2}{4}\right)^2, \qquad (40)$$

obtained by setting $P_2 = P_5 = P$ and $P_1 = P_3 = P_4 = 1$, that is, by attaching all leaves to the root ρ like in a star tree.

It can easily be seen that the right-hand side of Equation (40) is maximized, namely equal to $\frac{1}{32}$, when $P = 0$.

Similarly, maximizing $P(f_2|\mathcal{T}_3)$ yields

$$\max P(f_2|\mathcal{T}_3) = \left(\frac{1-P^2}{4}\right)^2, \qquad (41)$$

and this value is obtained by setting $P_5 = 1$ (which corresponds to making the corresponding edges e_5 and e_8 short) and as before setting $P_2 = P$, $P_1 = 1$, $P_3 = 1$ and $P_4 = P$. The right-hand side of Equation (41) assumes its maximum $\frac{1}{16}$ at $P = 0$.

Equations (40) and (41) imply $\max P(f_2|\mathcal{T}_3) > \max P(f_1|\mathcal{T}_4)$ for all fixed values of P as well as in the limit as $P \to 0$. Note that $\max P(f_1|\mathcal{T}_3) = \max P(f_2|\mathcal{T}_4)$. So altogether we have for the sequence $S = f_1 f_2$: $\max P(S|\mathcal{T}_3) > \max P(S|\mathcal{T}_4)$. Thus, ML favors tree $\mathcal{T}_3$ over tree $\mathcal{T}_4$, whereas both trees are most parsimonious. Therefore, MP and ML are not equivalent under the constraints of a molecular clock, even when there is no common mechanism.

As before, we can construct a longer sequence $\tilde{S}$ comprised of k copies of character f_1 and $k + c$ copies of character f_2 for some integers $k, c > 0$. Then k can be chosen sufficiently large such that ML will favor tree $\mathcal{T}_3$ over $\mathcal{T}_4$, while $\mathcal{T}_4$ is an MP-tree and $\mathcal{T}_3$ is not. Thus, even under the assumption of no common mechanism, MP and ML choose different sets of trees in this setting.

As in Proposition 5.6, the above example for the binary case subject to a molecular clock only provides an MP-tree, namely $\mathcal{T}_4$, which is not an ML-tree, but it is unclear whether or not the opposite can occur. However, as both our ternary as well as binary examples show, it is often possible to modify the examples for bounded substitution probabilities such that they also provide examples for the clock case. Therefore, we assume that if such an example exists for the bounded case described in Chapter 5.2, this might also lead to an example for the restriction induced by a molecular clock.

5.4 Interpretation

Our main objective was to present examples demonstrating that MP and ML may choose different sets of trees under the N_r-model with no common mechanism when the substitution probabilities are bounded above by $u < \frac{1}{r}$ or when a molecular clock is assumed. Another intention was to give an elementary proof for the equivalence of MP and ML in the N_r-setting with no common mechanism, as shown by Tuffley and Steel [44], and thus to make this important result more easily accessible. The main idea of the proof is basically Corollary 5.4 (which corresponds to Lemma 2 of [44]), which states that the maximum likelihood can be obtained at a corner of the k-dimensional box $[0, \frac{1}{r}]^k$. But as we have shown, for an upper bound $u < \frac{1}{r}$ on the substitution probabilities, the approach of assigning probabilities u to those edges on which a most parsimonious extension g for a given character requires a change and 0 elsewhere, does not work. This is due to the fact that the maximum of the likelihood function might occur at a different corner of the box in this case.

Our 4-taxa example employing three character states shows that even if the upper bound u is arbitrarily close to $\frac{1}{r}$, we can find a sequence of characters which causes MP and ML to make conflicting choices. The motivation for our 4-taxa examples was based on the idea of misleading sequences as defined in Chapter 4. In particular, the construction idea of the sequences of the present chapter is based on the fact that MP ignores parsimoniously non-informative characters in any sequence, whereas ML (just as distance-based methods) does not. We exploited this fact to cause a discrepancy between MP and ML by taking sufficiently many non-informative characters.

It has been known that there are no binary misleading sequences: If the pairwise Hamming distances among a set of binary sequences perfectly fit a tree then this tree is also the perfect phylogeny (see [38], Proposition 7.1.9). But for MP and ML it is still unknown if under the N_2-model with no common mechanism there is a sequence of binary characters for which these methods make strictly conflicting choices when the substitution probabilities are bounded above by $u < \frac{1}{r}$. Our 5-taxa example, which uses binary characters only, shows that in this setting there are some MP-trees that are not ML-trees, but we observed that the ML-trees in our examples are also MP-trees – which means that the equivalence of MP and ML failed but we did not find an example for strictly conflicting choices in the binary case.

A particularly surprising aspect of our results regards small substitution probabilities, where MP is normally assumed to be justified in the sense of agreement with ML (see, e.g., [12]): small substitution probabilities or, more precisely, substitution probabilities subject to an upper bound, are one of the settings that make the equivalence fail – however, we also showed that for a sufficiently small choice of the upper bound, at least every ML-tree is also an MP-tree (but not vice versa). Moreover, although MP has been proven to agree with ML in the N_r-model under the assumption of no common mechanism (under no further constraints) our examples show that this equivalence may fail when the model is changed slightly. Therefore, we conclude that neither the presence nor the absence of a common mechanism alone can justify MP in the sense of an MP-ML equivalence. More research could be done on other models of nucleotide substitution in order to analyze conditions under which ML and MP may give conflicting results. This might highlight even more differences between MP and ML.

Chapter 6

Sequence Length Bounds for Resolving a Deep Phylogenetic Divergence

> There is, after all, one true tree of life, the unique pattern of evolutionary branchings that actually happened. It exists.
>
> Richard Dawkins

When sequence sites evolve independently under a Markov process along the branches of a tree $\mathcal{T}$, the sequences observed at the tips contain information concerning the underlying tree. This allows for the tree $\mathcal{T}$ to be reconstructed accurately from sufficiently long sequences, which is the basis of modern molecular systematics [12]. The number of sites required to accurately reconstruct $\mathcal{T}$ depends on how long the edges of the tree are. More precisely, it depends on the expected number of substitutions on each branch (edge) e of the tree – which we refer to as the *branch length* of e (this is the product of the temporal duration of the branch and the substitution rate).

A number of authors (e.g. [6; 21; 35; 43; 46; 47; 49]) have considered various ways to quantify the phylogenetic signal in aligned DNA sequences, and to estimate the sequence length required to reconstruct a phylogenetic tree. Most of these studies have involved simulations or heuristic approaches, although some analytical bounds have also been obtained [28; 41]. Typically, these bounds state that if an interior branch length is very short, or if a terminal (external) branch length is long, then a large number of sites will be required.

In this chapter, we explore these results further by obtaining bounds that are expressed purely in terms of the relative sizes of the branch lengths, not their absolute values. One motivation for our approach is that different genes are known to evolve at different rates, so that any particular branch length will depend on which gene is considered; however, the ratios of the branch lengths will be unchanged if the gene-specific rate applies uniformly across the tree.

A particularly difficult tree reconstruction problem, requiring long sequences to resolve, arises when one has an interior edge with a short branch length incident with edges (or subtrees) having large branch lengths. Such a scenario happens, for example, when speciation events occurred in rapid succession (leading to short branch lengths) in the distant past (leading to the large branch lengths for the incident edges). Several examples of this have been highlighted in the literature [23; 33] and include the origin of metazoa and the origin of photosynthesis.

In this chapter, we analyze a scenario which, although somewhat idealized, nevertheless captures the essence of this problem – a 4-taxa tree, where the terminal edges have equal branch lengths that are $\lambda > 1$ times the branch lengths of the interior edge, and a simple symmetric model of site evolution (specifically, we assume sites evolve according to the N_2-model as described earlier).

We provide a mathematical analysis to the question of how many sites are required to resolve the tree correctly (from the three possible resolved topologies on four taxa). We are particularly interested in how the growth of the sequence length k required to reconstruct the true tree with high probability depends on λ, independent of the absolute value of a particular edge length. We establish that k must grow at the rate λ^2, which implies that regardless of how fast (or slow) any particular sequence is evolving, we can set explicit lower bounds on the length of sequences required to resolve the tree. We then show that for our setting, the growth in k does not have to be any worse than this quadratic growth in λ, because an existing method (namely, MP) achieves this growth rate. This does not imply that MP is the 'best' method for tree reconstruction; we chose it simply because we can analytically calculate tree reconstruction probabilities for this method. Our results complement an earlier simulation-based analysis [49]. We contrast our results by considering a quite different model of site evolution (the infinite state model) and establishing that order λ growth in k can sometimes suffice for this model.

We also extend the approach to more general Markov processes on trees, obtaining exact, but less explicit lower bounds on k and which involve absolute (rather than relative) branch lengths. Our arguments are based on standard techniques from probability theory, such as a Central Limit approximation, and information-theoretic arguments based on the properties of the Hellinger distance.

6.1 Preliminaries

Consider an unrooted binary phylogenetic tree on four taxa, say 12|34, with branch length x for the interior edge e_5 and λx for the terminal edges $e_1, \ldots, e_4$, where $\lambda > 1$. This is illustrated in Figure 6.1(a), and the topology of the tree is shown at the top of Figure 6.1(b). The other two competing topologies (13|24 and 14|23) are also shown in Figure 6.1(b). Here, branch length refers to the expected number of substitutions under some continuous time substitution process.

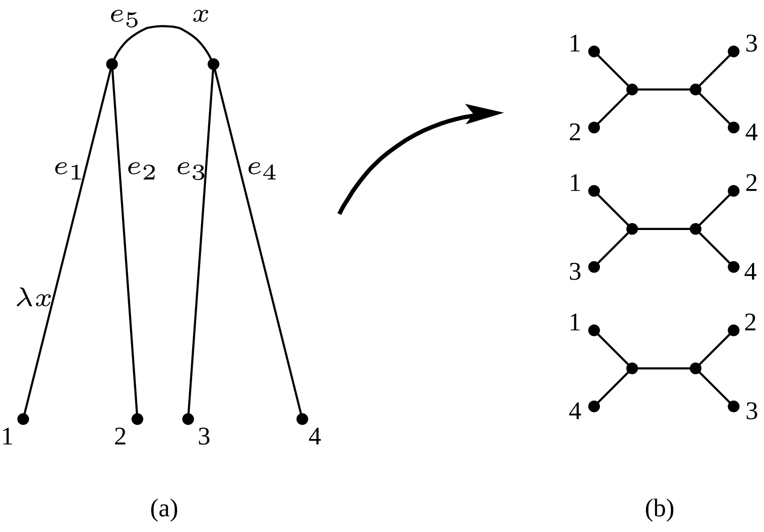


Figure 6.1: (a) The generating tree with interior branch length x and all four terminal branch lengths equal to λx. (b) This tree has the topology 12|34, while the other two binary topologies are 13|24 and 14|23.

Suppose that a sequence of binary characters are generated independently and identically (i.i.d.) under the N_2-model. Although it is the simplest non-trivial Markov process on a tree, it allows for an exact analysis. Moreover, stochastic results for this model typically extend to more general finite-state models where an exact analysis is usually more complex [28], and in Section 6.5 we show how some of our approaches extend to more general Markov processes.

If we denote the substitution probability on edge e_i by $P(e_i)$, then for each terminal edge we have $P(e_i) = \frac{1}{2}(1 - \exp(-2\lambda x))$ while for the central edge e_5, we have $P(e_5) = \frac{1}{2}(1 - \exp(-2x))$. Let $\theta_i = 1 - 2P(e_i)$ for $i = 1, \ldots, 5$. Then we can express these five θ_i-values in terms of $\theta := e^{-2x}$ as follows:

$$\theta_i = \theta^{\lambda} \text{ for } i = 1, \ldots, 4; \text{ and } \theta_5 = \theta.$$

Now, if we fix x and let λ grow, or, alternatively, if we fix λx and let x tend to zero, then the sequence length k required to accurately reconstruct the topology of the generating tree tends to infinity. Informally, this is because under either of the two limiting situations described, the three trees in Figure 6.1(b) will (in the limit) give the same probability distribution on site patterns, and so the three trees will describe any data equally well. This holds for any tree reconstruction method that treats all three topologies fairly (if, on the other hand, a method has an a priori preference for a particular topology, it will perform worse on an alternative topology). Moreover, if λx is fixed, then k grows at the rate $\frac{1}{x^2}$ as x tends to zero (by Theorem 4.1 of [41]). However, if we do not fix x or λx in advance, two fundamental questions arise: what is the slowest rate that k can possibly grow as a function of λ, and does some value of x (dependent on λ) achieve this rate of growth for a certain tree reconstruction method? We will see that for the simple scenario described, the answer is that the slowest rate, namely λ^2, can be achieved (up to a constant factor), for example by Maximum Parsimony.

6.2 Lower Bounds

The main result of this section is the following theorem.

Theorem 6.1. *Suppose k sites evolve i.i.d. under a symmetric 2-state model on some (unknown) 4-taxa tree that has branch length x on the interior edge and λx on each terminal edge. Then any method that is able to correctly identify the underlying tree topology with probability at least $1-\epsilon$ requires:*

$$k \geq c_\epsilon \cdot \lambda^2$$

for any x, where $c_\epsilon = \frac{1}{2}(1-\frac{3}{2}\epsilon)^2$.

To establish this result we require some preliminaries. We begin with a general information-theoretic bound on the number of i.i.d. observations required to reconstruct a discrete parameter in a general setting.

Suppose one has a finite set A (e.g., the set of all phylogenetic trees on n taxa), and each element $a \in A$ has an associated probability distribution on a finite set U (e.g., the set of all site patterns on n taxa). Suppose we have k observations from U that are generated independently by the same unknown element $a \in A$. Suppose, furthermore, that some method M estimates the element of A that generated our observations and does so correctly with probability at least $1-\epsilon$ (regardless of which element a actually generated the data). Then we can set a lower bound on k in terms of a stochastic distance between elements of A.

Definition 6.2 (Hellinger distance)**.** *We define the* Hellinger distance *of two elements $a, a' \in A$ for a finite set A with associated probability distribution on a finite set U as follows. If $p : U \rightarrow [0,1]$, such that $p(u) = p_u$, and $q : U \rightarrow [0,1]$, such that $q(u) = q_u$,* *denote the probability distributions induced by a and a', respectively, then let:*

$$d_H^2(a,a') := \sum_{u\in U}(\sqrt{p_u}-\sqrt{q_u})^2 = 2\left(1-\sum_{u\in U}\sqrt{p_u q_u}\right). \tag{42}$$

Note that the latter equality holds as $\sum\limits_{u\in U} p_u = \sum\limits_{u\in U} q_u = 1$.

The following result corresponds to Theorem 3.1 and Equation (2.7) in [41].

Lemma 6.3. *If there is a subset A' of A of size $m \geq 2$ for which $d_H(a,a') \leq d$ for all $a, a' \in A'$, and some method M correctly identifies each element of A' with probability at least $1-\epsilon$ from k independently generated elements in some set U, then:*

$$k \geq \frac{1}{4}(1-\frac{m}{m-1}\epsilon)^2 d^{-2}.$$

In our setting, A will consist of the three binary 4-taxa trees on leaf set $\{1,2,3,4\}$, U will consist of the assignment of states to the elements of this leaf set, and m will be 3 (in this chapter) or 2 (in Chapter 6.5).

Let S be the set of possible binary site patterns on $\{1,2,3,4\}$. These consist of the site patterns $s_1 := \alpha\alpha\beta\beta, s_2 := \alpha\beta\alpha\beta$ and $s_3 := \alpha\beta\beta\alpha$, and five non-informative ones $s_4,\ldots,s_8$. Note that pairs of complementary site patterns – for example $\alpha\alpha\beta\beta$ and $\beta\beta\alpha\alpha$ – are regarded as equivalent. For any site pattern $s\in S$, let $p_s = P(s|\mathcal{T}_1)$ (and $q_s = P(s|\mathcal{T}_2)$, respectively) be the probability that the site pattern s is generated on $\mathcal{T}_1$ (or $\mathcal{T}_2$, respectively). We can express the probabilities p_{s_1} and p_{s_2} in terms of $\theta = e^{-2x}$ by using the Hadamard representation of [16] (see [38], Section 8.6):

$$p_{s_1} = \frac{1}{8}\cdot\left(1+\theta_1\theta_2+\theta_3\theta_4-\theta_1\theta_3\theta_5-\theta_2\theta_3\theta_5-\theta_1\theta_4\theta_5-\theta_2\theta_4\theta_5+\theta_1\theta_2\theta_3\theta_4\right),$$

$$p_{s_2} = \frac{1}{8}\cdot\left(1-\theta_1\theta_2-\theta_3\theta_4+\theta_1\theta_3\theta_5-\theta_2\theta_3\theta_5-\theta_1\theta_4\theta_5+\theta_2\theta_4\theta_5+\theta_1\theta_2\theta_3\theta_4\right).$$

So here, we have:

$$p_{s_1} = \frac{1}{8} \cdot \left(1 + 2 \cdot \theta^{2\lambda} - 4 \cdot \theta^{2\lambda+1} + \theta^{4\lambda}\right), \tag{43}$$

and

$$p_{s_2} = \frac{1}{8} \cdot \left(1 - 2 \cdot \theta^{2\lambda} + \theta^{4\lambda}\right) = \frac{1}{8}\left(1 - \theta^{2\lambda}\right)^2. \tag{44}$$

To obtain an upper bound on the Hellinger distance, we require a further technical lemma.

Lemma 6.4. *Let $\gamma > 1$ and let $h(x) = \frac{x^{\gamma}(1-x)}{(1-x^{\gamma})}$. Then the supremum of $h(x)$ for x in the half-open interval $[0,1)$ equals $\frac{1}{\gamma}$.*

Proof. Since $\gamma > 1$, it can be checked that $h'(x) > 0$ for all x in $(0,1)$, and so $\sup_{x\in[0,1)} h(x) = \lim_{x\uparrow 1} h(x)$. By L'Hôpital's rule, we have $\lim_{x\uparrow 1} h(x) = \frac{1}{\gamma}$. □

We are now in a position to prove the theorem.

Proof of Theorem 6.1. If any method has a probability of at least $1-\epsilon$ of correctly reconstructing each of the three binary trees on four taxa from i.i.d. sequences of length k then, by Lemma 6.3 with $m = 3$ we have:

$$k \geq \frac{(1 - \frac{3}{2}\epsilon)^2}{4} \cdot d_H^{-2}, \tag{45}$$

where d_H is the maximum Hellinger distance between any two of the three trees. Now, if each of the three trees has the $x, \lambda x$ combination of branch lengths (for interior and pending branches, respectively) then, by symmetry, all three of these pairwise Hellinger distances are equal. Moreover, we claim that:

$$d_H^{-2} \geq 2\lambda^2, \tag{46}$$

which together with (45) requires $k \geq c_\epsilon \lambda^2$ for $c_\epsilon = \frac{1}{2}(1 - \frac{3}{2}\epsilon)^2$. Thus, it remains to establish (46).

Wlog. $\mathcal{T}_1 = 12|34$ and $\mathcal{T}_2 = 13|24$. Now, for all $i = 3, \ldots, 8$, we have $p_{s_i} = q_{s_i}$. Furthermore, $p_{s_1} = q_{s_2}$ and $p_{s_2} = q_{s_1}$ as the given trees are identical except for their leaf labeling. Consequently, Equation (42) can be simplified as follows:

$$\begin{aligned} d_H^2(\mathcal{T}_1, \mathcal{T}_2) &= 2\left(1 - \sum_{i=1}^{8} \sqrt{p_{s_i} q_{s_i}}\right) = 2\left(1 - \sum_{i=3}^{8} p_{s_i} - 2\sqrt{p_{s_1} p_{s_2}}\right), && (47)\\ &= 2\left(1 - (1 - p_{s_1} - p_{s_2}) - 2\sqrt{p_{s_1} p_{s_2}}\right), && (48)\\ &= 2\left(p_{s_1} + p_{s_2} - 2\sqrt{p_{s_1} p_{s_2}}\right). && (49)\end{aligned}$$

Let $\delta = \frac{1}{2}\theta^{2\lambda}(1 - \theta)$. Then $p_{s_1} = p_{s_2} + \delta$, and so Equation (49) can be re-written as:

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) = 4p_{s_2}\left(1 + \frac{\delta}{2p_{s_2}} - \sqrt{1 + \frac{\delta}{p_{s_2}}}\right). \quad (50)$$

Applying the inequality $\sqrt{1+y} \geq 1 + \frac{y}{2} - \frac{y^2}{4}$, for any $y > 0$, to $y = \frac{\delta}{p_{s_2}}$ in (50), gives:

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) \leq \frac{\delta^2}{p_{s_2}} = 2\left[\frac{\theta^{2\lambda}(1-\theta)}{1 - \theta^{2\lambda}}\right]^2 \leq \frac{1}{2\lambda^2},$$

where the last inequality follows by invoking Lemma 6.4 with $\gamma = 2\lambda, x = \theta$. This establishes (46) and thereby completes the proof of the theorem. □

6.3 An Upper bound: The Performance of Maximum Parsimony

We now show that the lower bound described above is essentially 'best possible' (up to a constant factor) for the given model, as it can be achieved for a certain choice of x by a simple tree reconstruction method, namely Maximum Parsimony.

The probability that MP correctly reconstructs the true tree 12|34 will be called the *MP reconstruction probability*. Let $f(\epsilon)$ denote the one-sided ϵ-critical value for the standard normal distribution, defined by:

$$f(\epsilon) = z \Leftrightarrow \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \epsilon.$$

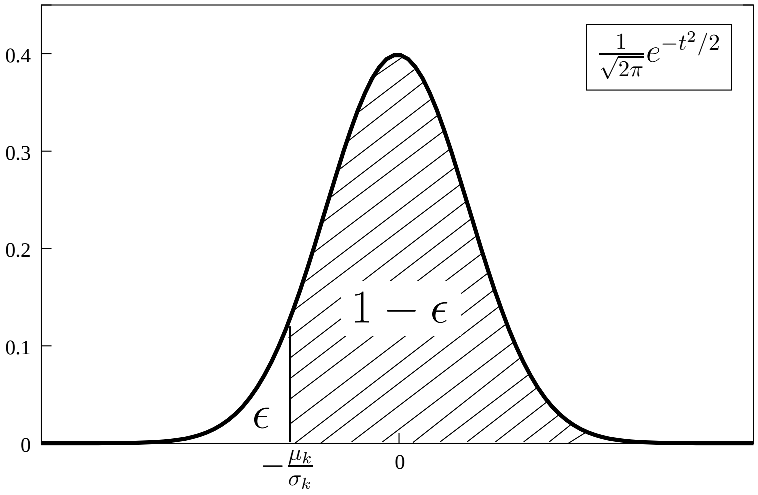


Figure 6.2: The probability density function of the Standard Normal Distribution. The shaded area shows the reconstruction probability.

Theorem 6.5. *Suppose k sites evolve i.i.d. under a symmetric 2-state model on some (unknown) 4-taxa tree that has branch length x on the interior edge and λx on each terminal edge. Then for a sequence c'_λ with $\lim_{\lambda\to\infty} c'_\lambda = 4e^2$, the following holds: If $k \geq c'_\lambda f(\frac{\epsilon}{2})^2 \cdot \lambda^2$, an interior branch length x exists for which the MP reconstruction probability is at least $1-\epsilon$.*

In order to prove this theorem, some preliminary work is required. Suppose we generate a sequence $\mathcal{C}$ of k i.i.d. sites under the symmetric 2-state model. Define the random variables X_i and Y_k as follows. Let:

$$X_i = \begin{cases} 1, & \text{if } i^{th} \text{ character in } \mathcal{C} \text{ is of the kind } (\alpha,\alpha,\beta,\beta); \\ -1, & \text{if } i^{th} \text{ character in } \mathcal{C} \text{ is of the kind } (\alpha,\beta,\alpha,\beta); \\ 0, & \text{else.} \end{cases}$$

and let:

$$Y_k = \sum_{i=1}^{k} X_i. \tag{51}$$

The probability that MP will favor the tree 12|34 over 13|24 is then $P(Y_k > 0)$. We will exploit the fact that the random variables X_i are i.i.d., and so Y_k can be approximated for large k by a normal distribution with a mean μ_k and a standard deviation σ_k. These two parameters can be easily described (just) in terms of θ, λ and k as follows.

Lemma 6.6.

1. $\mu_k = k \cdot [P(X_1 = 1) - P(X_1 = -1)]$

2. $\sigma_k^2 = k \cdot \left[P(X_1 = 1) + P(X_1 = -1) - [P(X_1 = 1) - P(X_1 = -1)]^2\right]$

Proof.

1. $\mu_k = E(Y_k) = E\left(\sum_{i=1}^{k} X_i\right) \overset{add.}{=} \sum_{i=1}^{k} E(X_i) \overset{i.i.d.}{=} k \cdot E(X_1) =$
$k \cdot [P(X_1 = 1) \cdot 1 + P(X_1 = -1) \cdot (-1) + P(X_1 = 0) \cdot 0] = k \cdot [P(X_1 = 1) - P(X_1 = -1)]$

2. $\sigma_k^2 = Var(Y_k) = Var\left(\sum_{i=1}^{k} X_i\right) \overset{i.i.d.}{=} \sum_{i=1}^{k} Var(X_i) \overset{i.i.d.}{=} k \cdot Var(X_1) =$
$k \cdot \left[(1 - E(X_1))^2 \cdot P(X_1 = 1) + (-1 - E(X_1))^2 \cdot P(X_1 = -1) + (0 - E(X_1))^2 \cdot P(X_1 = 0)\right]$
$= k \cdot \left[P(X_1 = 1) + P(X_1 = -1) - [P(X_1 = 1) - P(X_1 = -1)]^2\right]$

□

Using this lemma, we can establish the following.

Lemma 6.7.

1. $\mu_k = k \cdot \frac{1}{2}\theta^{2\lambda}(1 - \theta)$.

2. $\sigma_k^2 = k \cdot \frac{1}{4}(1 + 2\theta^{4\lambda+1} - 2\theta^{2\lambda+1} - \theta^{4\lambda+2})$.

3. $\frac{\mu_k}{\sigma_k} \geq \sqrt{k} \cdot \theta^{2\lambda}(1 - \theta)$.

Proof. In the N_2-model and the generating tree in Figure 6.1(a), we have:

$$P(X_1 = 1) = p_{s_1}, \text{ and } P(X_1 = -1) = p_{s_2},$$

where p_{s_1}, p_{s_2} were given above in Equations (43) and (44), respectively. Parts (1) and (2) of the lemma now follow by substitution of the expressions for p_{s_1}, p_{s_2} into the formulas given by Lemma 6.6(1) and (2), respectively. For Part (3), note that Parts (1) and (2) imply that

$$\frac{\mu_k}{\sigma_k} = \sqrt{k} \cdot \frac{N_\theta}{D_\theta} \tag{52}$$

where $N_\theta = \theta^{2\lambda}(1-\theta); D_\theta = \sqrt{1 + 2\theta^{4\lambda+1} - 2\theta^{2\lambda+1} - \theta^{4\lambda+2}}$. We now show that $D_\theta \leq 1$. We have $1 + 0.5\theta^{2\lambda+1} \geq \theta^{2\lambda}$ and so $2\theta^{2\lambda+1}(1 - \theta^{2\lambda} + 0.5\theta^{2\lambda+1}) \geq 0$. Consequently $1 - 2\theta^{2\lambda+1}(1 - \theta^{2\lambda} + 0.5\theta^{2\lambda+1}) \leq 1$, which implies that $D_\theta^2 \leq 1$. Part (3) now follows from (52) by the inequality $D_\theta \leq 1$. □

Proof of Theorem 6.5. Note that the MP reconstruction probability is the probability that MP will favor the true tree 12|34 over both alternative trees on four taxa, namely 13|24 and 14|23. Recall that the event of the tree 12|34 being favored over 13|24 can be expressed as $P(Y_k > 0)$. The event of 12|34 being favored over 14|23 can be expressed similarly by defining the random variables $\tilde{X}_i$ and $\tilde{Y}_k$ which are analogous to X_i and Y_k, using the character $(\alpha, \beta, \beta, \alpha)$ instead of $(\alpha, \beta, \alpha, \beta)$. Then, the MP reconstruction probability can be written as $P\Big((Y_k > 0) \cap (\tilde{Y}_k > 0)\Big)$. Let:

$$Z_k = \frac{Y_k - \mu_k}{\sigma_k}.$$

Thus, Z_k is the normalized difference of the parsimony score between tree 13|24 and 12|34 for a k i.i.d. characters generated by the tree in Figure 6.1(a). By Lemma 6.7(3) we have

$$P(Y_k \leq 0) = P(Z_k \leq -\frac{\mu_k}{\sigma_k}) \leq P\Big(Z_k \leq -\sqrt{k}\theta^{2\lambda}(1-\theta)\Big). \tag{53}$$

Now, by symmetry of the branch length of the generating tree in Figure 6.1(a), we have

$P(Y_k \leq 0) = P(\tilde{Y}_k \leq 0)$. Moreover, by Boole's inequality:

$$P\left((Y_k > 0) \cap (\tilde{Y}_k > 0)\right) \geq 1 - P(Y_k \leq 0) - P(\tilde{Y}_k \leq 0),$$

which, combined with (55), furnishes the following inequality for the MP reconstruction probability:

$$P\left((Y_k > 0) \cap (\tilde{Y}_k > 0)\right) \geq 1 - 2P(Y_k \leq 0) \geq 1 - 2P(Z_k \leq -\sqrt{k}\theta^{2\lambda}(1-\theta)). \quad (54)$$

Now, $\theta^{2\lambda} \cdot (1-\theta)$ has a unique local maximum in $[0,1]$, namely at $\theta' := 1 - \frac{1}{2\lambda+1}$, at which it takes the value $\frac{\alpha_\lambda}{\lambda}$, where $\alpha_\lambda = \left(1 - \frac{1}{1+2\lambda}\right)^{2\lambda} \cdot \frac{\lambda}{(1+2\lambda)} \to \frac{1}{2}e^{-1}$ as $\lambda \to \infty$. Moreover, the difference between the distribution of Z_k and a standard normal distribution tends uniformly to zero as λ (and hence k) grows. This follows by applying standard bounds on the Central Limit Theorem approximation (see, for example, [51]; one cannot directly apply the usual form of the central limit theorem as the distribution of the X_i's is changing with increasing λ). Thus we have $P(Z_k \leq -\sqrt{k}\frac{\alpha_\lambda}{\lambda}) \leq \frac{\epsilon}{2}$ provided that k grows at the rate $c'_\lambda \lambda^2 f(\frac{\epsilon}{2})^2$ for a sequence $c'_\lambda \to 4e^2$ as $\lambda \to \infty$.

In summary, by (54), a value for θ exists, namely $\theta' = 1 - \frac{1}{1+2\lambda}$, and thus a value for $P(e_5) = \frac{1}{2}(1-\theta') = \frac{1}{2(1+2\lambda)} \sim \frac{1}{4\lambda}$ also exists, for which the MP reconstruction probability is at least $1 - \epsilon$. This completes the proof. □

Interestingly, the optimal choice of x of (approximately) $\frac{1}{4\lambda}$ for MP has already been observed in a slightly different setting by Townsend in [43].

Note that while Theorem 6.1 provides a general lower bound for any tree reconstruction method, such a bound can be established for MP independently. We present this result in the following proposition.

Proposition 6.8. *Suppose k sites evolve i.i.d. under the N_2-model on some (unknown) 4-taxa tree that has branch length x on the interior edge and λx on each terminal edge.*

Then, in order to correctly identify the underlying tree topology with probability at least $1-\epsilon$*, Maximum Parsimony requires*

$$k \geq f(\epsilon)^2 \cdot \lambda^2$$

for any x*, where* $f(\epsilon)$ *is the one-sided* ϵ*-critical value of the standard normal distribution.*

Proof. By definition of $f(\epsilon)$ and with μ_k and σ_k as in Equation (52), we have

$$k \cdot \frac{N_\theta^2}{D_\theta^2} = \frac{\mu_k^2}{\sigma_k^2} \geq f(\epsilon)^2 \Leftrightarrow P(Y_k > 0) = P\left(\frac{Y_k - \mu_k}{\sigma_k} \geq - \quad \frac{\mu_k}{\sigma_k}\right) \geq 1-\epsilon. \qquad (55)$$

Let the reconstruction probability $P\Big((Y_k \geq 0) \cap (\tilde{Y}_k \geq 0)\Big) \geq 1-\epsilon$.
Since $P\Big((Y_k \geq 0) \cap (\tilde{Y}_k \geq 0)\Big) \leq P(Y_k > 0)$ by definition, this implies
$P(Y_k > 0) \geq 1-\epsilon$. Using Equation (55), this leads to $k \cdot \frac{N_\theta^2}{D_\theta^2} \geq f(\epsilon)^2$. Furthermore,
$1 \geq D_\theta^2 \geq (1-\theta^{2\lambda})^2$. In order to establish the last inequality, we need to examine D_θ^2:

$$D_\theta^2 \geq (1-\theta^{2\lambda})^2$$

$$\Leftrightarrow 2\theta^{4\lambda+1} - 2\theta^{2\lambda+1} - \theta^{4\lambda+2} \geq -2\theta^{2\lambda} + \theta^{4\lambda}$$

$$\Leftrightarrow 2\theta^{2\lambda+1} - 2\theta - \theta^{2\lambda+1} \geq -2 + \theta^{2\lambda}$$

$$\Leftrightarrow 2 \geq 2\theta + \theta^{2\lambda}(1-\theta)^2$$

Note that we defined $\theta = \theta_5 = 1 - 2P(e_5)$, where $P(e_5)$ is the substitution probability of the interior edge e_5, and thus $0 \leq P(e_5) \leq 1$. Substituting θ accordingly leads to:

$$\Leftrightarrow 2 \geq 2(1-2P(e_5)) + (1-2P(e_5))^{2\lambda}(2P(e_5))^2$$

$$\Leftrightarrow 4P(e_5) \geq (2P(e_5))^2(1-2P(e_5))^{2\lambda}$$

$$\Leftrightarrow 2 \geq 2 \underbrace{P(e_5)}_{\leq 1} \underbrace{(1-2P(e_5))^{2\lambda}}_{\leq 1}$$

Since the last inequality obviously holds, this gives $1 \geq D_\theta^2 \geq (1-\theta^{2\lambda})^2$.

Now let $\hat{\lambda} := \lfloor \lambda \rfloor$, i.e. $\hat{\lambda} = \lambda$ if λ is an integer, else $\hat{\lambda}$ is the largest integer smaller than λ. Note that then by definition $\lambda < \hat{\lambda} + 1 \leq 2\hat{\lambda}$ for $\lambda > 1$. Then,

$$\Rightarrow f(\epsilon)^2 \leq k \cdot \frac{N_\theta^2}{(1-\theta^{2\lambda})^2} = k \cdot \left(\frac{\theta^{2\lambda}(1-\theta)}{1-\theta^{2\lambda}}\right)^2 \leq k \cdot \left(\frac{\theta^{2\hat{\lambda}}(1-\theta)}{1-\theta^{2\hat{\lambda}}}\right)^2$$

$$= k \cdot \left(\frac{\theta^{2\hat{\lambda}}}{\underbrace{1+\theta^{\hat{\lambda}}}_{>2\theta^{\hat{\lambda}}}} \cdot \frac{1}{1+\theta+\theta^2+\ldots+\theta^{\hat{\lambda}-1}} \right)^2 \leq k \cdot \left(\frac{\theta^{2\hat{\lambda}}}{2\theta^{\hat{\lambda}} \cdot \left(1+\theta+\theta^2+\ldots+\theta^{\hat{\lambda}-1}\right)} \right)^2$$

$$= k \cdot \left(\frac{\theta^{\hat{\lambda}}}{2 \cdot \left(1+\theta+\theta^2+\ldots+\theta^{\hat{\lambda}-1}\right)} \right)^2 = k \cdot \left(\frac{1}{2 \cdot \left(\frac{1}{\theta^{\hat{\lambda}}} + \frac{1}{\theta^{\hat{\lambda}-1}} + \frac{1}{\theta^{\hat{\lambda}-2}} + \ldots + \frac{1}{\theta}\right)} \right)^2$$

$$\leq k \cdot \left(\frac{1}{2\hat{\lambda}}\right)^2 \leq k \cdot \frac{1}{\lambda^2}$$

$$\Rightarrow k \geq f(\epsilon)^2 \lambda^2.$$

□

Regarding Theorem 6.5 and Proposition 6.8, other tree reconstruction methods have a similar performance to MP when k grows at the rate λ^2. Indeed it is possible that such methods will require shorter sequences and have better statistical properties on trees with different tree shapes. This is due to the fact that MP is statistically inconsistent under some combinations of branch lengths (but these lie outside those considered in the scenario of Figure 6.1). We have chosen to consider MP here, because the analysis is relatively straightforward and it suffices to prove the matching lower bound of λ^2.

To conclude this chapter, we now show that one can also derive a form of Proposition 6.8 using Azuma's inequality [1] in a slightly modified form as introduced by McDiarmid [26].

Theorem 6.9 (McDiarmid, 1989)**.** *Let $X_1, \ldots, X_k$ be independent random variables taking values in a set A, and assume that $f : A^k \to \mathbb{R}$ satisfies*

$$|f(x_1, \ldots, x_i, \ldots, x_k) - f(x_1, \ldots, x_i', \ldots, x_k)| < c$$

for some constant c and for all $x_1, \ldots, x_k, x_i' \in A$. Let Y be the random variable $f(X_1, \ldots, X_k)$. Then for any $t > 0$, the following inequality holds:

$$P(Y - E(Y) < -t) < e^{-2\frac{t^2}{kc^2}}.$$

Corollary 6.10. *Suppose k sites evolve i.i.d. under the N_2-model on some (unknown) 4-taxa tree that has branch length x on the interior edge and λx on each terminal edge. Then, in order to correctly identify the underlying tree topology with probability at least $1 - \epsilon$, Maximum Parsimony requires for any x that*

$$k \geq 32e^2 \ln\left(\frac{1}{\epsilon}\right)\lambda^2.$$

Proof. Let $X_1, \ldots, X_k$ and Y_k be as defined in Equation (51). Let $Y := Y_k$ and $Y_k' := \sum_{j=1}^{k} X_j - X_i + X_i'$ for some $i \in \{1, \ldots, k\}$. This means that Y_k and Y_k' differ only by one summand. Then, as $X_i' \in \{-1, 0, 1\}$ by definition, we note that $|Y_i - Y_i'| \leq 2$. Therefore, we can set $c := 2$ and apply Theorem 6.9. Thus, $P(Y_k < 0) \Leftrightarrow P(Y - \mu_k < -\mu_k) < e^{-2\frac{\mu_k^2}{kc^2}}$. By Lemma 6.7, we get $e^{-2\frac{\mu_k^2}{2^2 k}} = e^{-k\frac{\theta^{4\lambda}(1-\theta)^2}{8}}$. For the reconstruction probability to be at least $1 - \epsilon$, we now require $e^{-k\frac{\theta^{4\lambda}(1-\theta)^2}{8}} \leq \epsilon$ and thus

$$k \geq \frac{8\ln\left(\frac{1}{\epsilon}\right)}{\theta^{4\lambda}(1-\theta)^2} \geq \boxed{\frac{8\ln\left(\frac{1}{\epsilon}\right)}{\underbrace{\left(1 - \frac{1}{1+2\lambda}\right)^{4\lambda}}_{\leq 1}\underbrace{\left(\frac{1}{1+2\lambda}\right)^2}_{\leq\left(\frac{1}{2\lambda}\right)^2}}} \geq 32\ln\left(\frac{1}{\epsilon}\right)\lambda^2.$$

This completes the proof. □

Note that as $\lambda \to \infty$, the denominator in the boxed fraction goes to $\frac{1}{4e^2\lambda^2}$, which gives $k \geq 32e^2 \ln\left(\frac{1}{\epsilon}\right)\lambda^2$. So in the limit, the term in place of c'_λ is larger by a factor of 8 than suggested by the Central Limit Theorem, but the advantage of the Azuma approach as opposed to the CLT is that it is non-asymptotic and therefore valid for any given finite k.

6.4 The Performance of MP under the Random Cluster Model

We now consider Markov processes in which the state space is countably infinite, and where a substitution is always to a new state. In particular, we consider the following random process on a phylogenetic tree $\mathcal{T}$. For each edge $e \in E(\mathcal{T})$, let a map $p : e \mapsto p(e)$, where $p(e)$ denotes a probability, be given (and let all $p(e)$'s be independent). Then cut this edge with probability $p(e)$ (or leave it intact with probability $1 - p(e)$). The resulting disconnected graph partitions the vertex set $V(\mathcal{T})$ into non-empty components according to the equivalence relation that $u \sim v$ if u and v are in the same component. This process thus generates random partitions of $V(\mathcal{T})$, and thereby of X. We call the resulting probability distribution on partitions of X the *random cluster model* for homoplasy-free evolution with parameters $(\mathcal{T}, p)$. For more information on this model, the reader is referred to [27].

For the random cluster model, the situation regarding sequence length requirements is quite different from that of the N_2-model described in the previous chapter. In this case, the required sequence length need only grow at the rate λ (not λ^2), as the following result shows.

Proposition 6.11. *Suppose k sites evolve i.i.d. under a random cluster model on some (unknown) 4-taxa tree that has branch length x on the interior edge and λx on each terminal edge. Then, for a constant C'_ϵ which depends just on ϵ, the following holds: If $k \geq C'_\epsilon \cdot \lambda$, then an x exists for which the MP reconstruction probability is at least $1 - \epsilon$.*

Proof. In the random cluster model, the probability of a substitution event on an edge e

can be written as $P(e) = 1 - \exp(-l)$ where l is the expected number of changes on the edge (the branch length). Now, the random cluster model generates only characters that are homoplasy-free on the generating tree; thus MP will return the generating tree from a sequence of characters, provided this tree is the only one on which those characters are homoplasy-free. For a tree with topology $12|34$, this will occur precisely if at least one of the k characters generated assigns taxa $1, 2$ a shared state, and taxa $3, 4$ a second shared state that is different to that assigned to $1, 2$. The probability Q that any given character generated by the tree in Figure 6.1(a) has this property is given by:

$$Q = P(e_5)\prod_{i=1}^{4}(1 - P(e_i)) = (1 - e^{-x})e^{-4\lambda x}.$$

Moreover, if $k \geq \frac{\log\left(\frac{1}{\epsilon}\right)}{Q}$ then $1-(1-Q)^k \geq 1-\epsilon$ (using the inequality $-\log(1-Q) \geq Q$). Consequently, MP will correctly reconstruct the generating tree with probability at least $1 - \epsilon$ provided that:

$$k \geq \log(\epsilon^{-1}) \cdot (1 - e^{-x})^{-1}e^{4\lambda x}. \qquad (56)$$

Taking $x = \frac{1}{4\lambda}$ we have $(1 - e^{-x})^{-1}e^{4\lambda x} \sim 4e\lambda$, which, in view of Equation (56), establishes the result. □

6.5 Lower Bounds for More General Models

In this chapter we derive a lower bound on the sequence length required for tree reconstruction, for a much wider range of Markov processes. However, unlike the previous chapters our bound is expressed in terms of the absolute branch lengths (or bounds on these) rather than in terms of ratios, and it involves constants that depend on the details of the model.

We first derive a general lemma. Consider any continuous-time, stationary and reversible Markov process. Let $\mathcal{S}$ denote its state space (thus, in previous sections $\mathcal{S} = \{\alpha, \beta\}$), and let $S = \mathcal{S}^4$. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two topologically distinct 4-taxa trees.

Suppose that the branch lengths of $\mathcal{T}_1$ are arbitrary, and that each edge of $\mathcal{T}_2$ has the corresponding interior or pendant branch length specified by $\mathcal{T}_1$ (where the pendant edge incident with leaf i in $\mathcal{T}_1$ corresponds to the pendant edge incident with leaf i in $\mathcal{T}_2$). For $s = (s_1, s_2, s_3, s_4) \in S$, let p_s (respectively q_s) denote the probability of generating s at the tips of $\mathcal{T}_1$ (respectively $\mathcal{T}_2$). Let p'_s (respectively q'_s) denote the conditional probability of generating s at the tips of $\mathcal{T}_1$ (respectively $\mathcal{T}_2$) given that a substitution has occurred on the central edge of $\mathcal{T}_1$ (respectively $\mathcal{T}_2$), and let $D_s := q'_s - p'_s$. Then we have the following result.

Lemma 6.12.

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) \leq l^2 \cdot \sum_{s \in S} \frac{D_s^2}{p_s}$$

where l denotes the branch length of the interior edge of $\mathcal{T}_1$.

Proof. Let τ denote the probability that a substitution occurs on the interior edge of $\mathcal{T}_1$, and let p_s^0 (respectively q_s^0) denote the conditional probability of generating s on $\mathcal{T}_1$ (respectively $\mathcal{T}_2$) given that no substitution occurs on the interior edge of $\mathcal{T}_1$ (respectively $\mathcal{T}_2$). By the law of total probability we have:

$$p_s = (1 - \tau) \cdot p_s^0 + \tau \cdot p'_s$$

and

$$q_s = (1 - \tau) \cdot q_s^0 + \tau \cdot q'_s.$$

Moreover, the assumptions on the correspondence between branch lengths of $\mathcal{T}_1$ and $\mathcal{T}_2$ imply that $p_s^0 = q_s^0$ for all $s \in S$ and so:

$$q_s - p_s = \tau(q'_s - p'_s) = \tau D_s.$$

Now,

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) = 2(1 - \sum_{s \in S} \sqrt{p_s q_s}) = 2\left(1 - \sum_{s \in S} p_s \sqrt{1 + \frac{\tau D_s}{p_s}}\right).$$

Applying the inequality $\sqrt{1+y} \geq 1 + \frac{y}{2} - \frac{y^2}{2}$ (for all $y \geq -1$) to $y = \frac{\tau D_s}{p_s}$ (and observing that $y \geq -1$ since $q_s \geq 0$), we obtain:

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) \leq 2\left(1 - \sum_s p_s \left(1 + \tau\frac{D_s}{2p_s} - \tau^2 \frac{D_s^2}{2p_s}\right)\right).$$

Now, $\sum_s p_s = 1$, and $\sum_s D_s = 0$ (since $\sum_s q_s' = \sum_s p_s' = 1$) and so this last inequality reduces to:

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) \leq \tau^2 \cdot \sum_{s \in S} \frac{D_s^2}{p_s}. \tag{57}$$

Furthermore, $\tau = P(N > 0)$, where N is the number of substitutions occurring on the interior edge of $\mathcal{T}_1$. However, $P(N > 0) \leq \mathbb{E}(N)$; that is, $\tau \leq l$, which, together with (57), provides the inequality stated in the lemma. □

We now apply this lemma to a slightly more restricted class of Markov processes to obtain the main result of this section.

Theorem 6.13. *Suppose k sites evolve i.i.d. under a finite-state, stationary and reversible continuous-time Markov process in which each state is accessible from any other state. Let l_0 be any strictly positive value. Consider this process on some (unknown) 4-taxa tree that has branch length at most l on the interior edge and at least $L \geq l_0$ on each terminal edge. Then any method that is able to correctly identify with probability at least $1 - \epsilon$ the underlying tree topology given these restriction requires:*

$$k \geq \frac{C}{4}(1 - 2\epsilon)^2 \cdot \frac{e^{cL}}{l^2}$$

where c and C are positive constants that depend only on R (the rate matrix for the process) and l_0.

Proof. We exploit the fact that any Markov process of the type described converges to its unique stationary distribution at an exponential rate (see, for example, Theorem 8.3 of [34]). Let $\pi(s)$ denote the stationary probability of s under the model. For $j = 1, \ldots, 4$, let $p(j) \in \{u, v\}$ be the end of the interior edge uv of $\mathcal{T}_1$ that is adjacent to leaf j (we may assume $p(1) = p(2) = u$; $p(3) = p(4) = v$), and let $S_{p(j)}$ denote the random state present at that vertex under the model. Then for any $s_j, s'_j \in \mathcal{S}$ there exist positive constants A, a (dependent on R) for which:

$$|P(S_j = s_j | S_{p(j)} = s'_j) - \pi(s_j)| \leq Ae^{-aL_j} \tag{58}$$

([34], Theorem 8.3), where L_j denotes the branch length of the edge incident with leaf j. For $s = (s_1, s_2, s_3, s_4) \in S = \mathcal{S}^4$, let

$$\pi_s = \prod_{j=1}^{4} \pi(s_j).$$

For $s's'' \in \mathcal{S}$ let $p'(s', s'')$ denote the probability of generating state s' at u and the state s'' at v given that at least one substitution occurs on the edge uv. Then, by the Markov assumption, and recalling the definition of p'_s from Lemma 6.12, we have:

$$p'_s = \sum_{(s',s'')\in\mathcal{S}^2} p'(s', s'') \cdot \prod_{j=1}^{2} P(S_j = s_j | S_u = s') \cdot \prod_{j=3}^{4} P(S_j = s_j | S_v = s''). \tag{59}$$

Combining (58) and (59), there exist positive constants B, b (dependent only on R) such that:

$$|p'_s - \pi_s| \leq Be^{-bL} \tag{60}$$

for all $s \in S$ (recall that $L \leq L_j$ for all j). Now, consider tree $\mathcal{T}_2$ which has branch lengths that correspond to those in $\mathcal{T}_1$ (as in Lemma 6.12). Then we also have:

$$|q'_s - \pi_s| \leq Be^{-bL} \tag{61}$$

for all $s \in S$. Combining (60) and (61) using the triangle inequality gives:

$$|D_s| = |q_s - p_s| \leq 2Be^{-bL}. \tag{62}$$

Moreover, since $L_j \geq l_0$ (for all j) and each state is accessible from any other state, we have $p_s \geq \delta$ (for some $\delta > 0$ dependent only on R and l_0). Combining this with (62) gives the following inequality, for all $s \in S$:

$$\frac{D_s^2}{p_s} \leq (4B^2/\delta)e^{-2bL}. \tag{63}$$

The theorem now follows from Lemma 6.12 and Lemma 6.3 (with $m = 2$). □

6.6 Interpretation

In this chapter we have provided precise results for a specific and simple model (the N_2-model), along with less explicit results for more general Markov processes (and phrased in terms of absolute rather than relative branch lengths). The aim was to determine rigorous bounds on the sequence length required for resolving a deep divergence, which may shed light on debates as to whether some early radiations might be fundamentally unresolvable on the basis of current models and data.

Of course, in applications, other phenomena (such as lineage sorting, sequencing errors, substitution model mis-specification, misalignment of sequences and alignment artifacts [32] and so forth) may further impede phylogenetic reconstruction. These errors are unlikely to help the tree reconstruction if our bound shows it is impossible even when the ideal model assumptions hold. We have seen that some models require significantly fewer characters for resolving a tree – in particular this holds for the random cluster model, and it is possible that new types of genomic data (involving rare genomic events where homoplasy is unlikely) can be described by these and related processes that preserve more phylogenetic signal regarding distant evolutionary divergences.

One limitation concerning our bounds is that they apply to pure Markov processes, in which each character evolves according to the same process. In molecular biology, a common assumption is that there is a distribution of rates across sites, in which each site evolves at a rate (selected independently from some fixed distribution) that acts as a multiplier for all the branch lengths in the tree (see e.g. [12; 38]). It would be interesting to extend the analysis in the last section to these models to obtain a lower bound on k analogous to Theorem 6.13.

Chapter 7

Expected Anomalies in the Fossil Record

> Why has not anyone seen that fossils alone gave birth to a theory about the formation of the earth, that without them, no one would have ever dreamed that there were successive epochs in the formation of the globe?
>
> Georges Cuvier

Since Darwin's book *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* [7], there has been much debate about the evidence for continuous evolution from a universal common ancestor. Initially, Darwin only assumed the relatedness of the majority of species, not of all of them. Later, however, he came to the view that because of the similarities of all existing species, there could only be one 'root' and one 'tree of life' [39]. All species have descended from this common ancestor and indications for their gradual evolution have been sought in the fossil record ever since. Usually, the improbability of fossilization or of finding existing fossils was put forward as the standard answer to the question of why there are so many 'gaps' in the fossil record. Such gaps have become popularly referred to as 'missing links', i.e., missing intermediates between taxa existing either today or as fossils.

Of course, the existence of gaps is in some sense inevitable: every new link gives rise to two new gaps, since evolution is generally a continuous process whereas fossil discovery will always remain discontinuous. Moreover, a patchy fossil record is not necessarily evidence against evolution from a common ancestor through a continuous series of intermediates – indeed, in a recent approach, Elliott Sober applied simple probabilistic arguments to conclude that the existence of some intermediates provides a stronger support for evolution than the non-existence of any (or some) intermediates could ever provide for a hypothesis of separate ancestry [39]. Moreover, some lineages appear to be densely sampled, whereas of others only few fossiliferous horizons are known [37]. This problem has been well investigated and statistical models have been developed to master it (see e.g. [25], [24]), [42]).

In this chapter, we suggest a further argument that may help explain missing links in the fossil record. Suppose that three fossils can be dated back to three different times. Can we really expect that a fossil from the intermediate time will appear (morphologically) to be an 'intermediate' of the other two fossils? We will explore this question via a simple stochastic model.

In order to develop this model, we first state some assumptions: firstly, we will consider that we are sampling fossil taxa of closely related organisms and which differ in a number of morphological characteristics. We assume this group of taxa has evolved in a 'treelike' fashion from some common ancestor; that is, there is an underlying phylogenetic tree, and the taxa are randomly sampled from points on the branches of this tree for given times. It is also necessary to say how morphological divergence might be related to time, as this is important for deciding whether a taxon is an intermediate or not. We will introduce two different interpretations for the degree of morphological divergence: in Chapter 7.2, we assume that fossils connected by short paths in the underlying phylogenetic tree are closely related, whereas fossils connected by a longer path are not as closely related. In Chapter 7.3, we will explain a biological drawback of this approach and analyze a more cladistic setting to overcome this problem. We will show, however, that for both approaches the main result does not change; namely, that it is possible that fossils from an intermediate time can, for certain trees and times, be expected to be less related to the first and the last sample than the latter two to one another. This surprising and counterintuitive result may help explain some of the 'gaps' in the fossil record.

7.1 Preliminaries

For both approaches, we need the following preliminaries. We begin with some notation. Throughout this chapter, we assume a rooted binary phylogenetic tree to be given with an associated time scale $0 < T_1 < T_2 < T_3$. The number of T_i-lineages (of lineages extant at time T_i) is denoted by n_i. For instance, in Figure 7.1, the number n_1 of T_1-lineages is 3, whereas the numbers n_2 and n_3 of T_2- and T_3-lineages are both 5. If not stated otherwise, extinction may occur in the tree. Every bifurcation in the tree is denoted by b_i, where b_0 is the root. Note that in a tree without extinction, the total number of bifurcations up to time T_3 (including the root) is $n_3 - 1$. For every b_i let t_i denote the time of the occurrence of bifurcation b_i. We may assume that the root is at time $t_0 = 0$.

Now, for every b_i, we make the following definitions:

$$P_i^{j,k} := n_{j,i}^l \cdot n_{k,i}^r + n_{j,i}^r \cdot n_{k,i}^l \quad \text{for all} \quad j,k \in \{1,2,3\}, j \neq k$$

where $n_{j,i}^l$ denotes the number of descendants the subtree with root b_i has at time T_j to the left of its root b_i, and $n_{j,i}^r$ is defined analogously for the descendants on the right hand side of b_i.

It can be seen that bifurcations for which at least one branch of offspring dies out in the same interval where the bifurcation lies always have $P_i^{j,k}$-value 0. This means, if either $t_0 < t_i < T_1$ or $T_1 < t_i < T_2$ or $T_2 < t_i < T_3$ and one of b_i's branches becomes extinct in the same interval, respectively, then $P_i^{j,k}$ is 0 for all j,k. If one regards the species present at times T_i, $i \in \{1,2,3\}$, as vertices, the number $P_i^{j,k}$ denotes the number of different paths in the tree from species present at time T_j to those present at time T_k in the subtree with root b_i (and in which, according to the common definition of 'path' in graph theory, no edge is taken twice). This is illustrated in Figure 7.1 and in the following example.

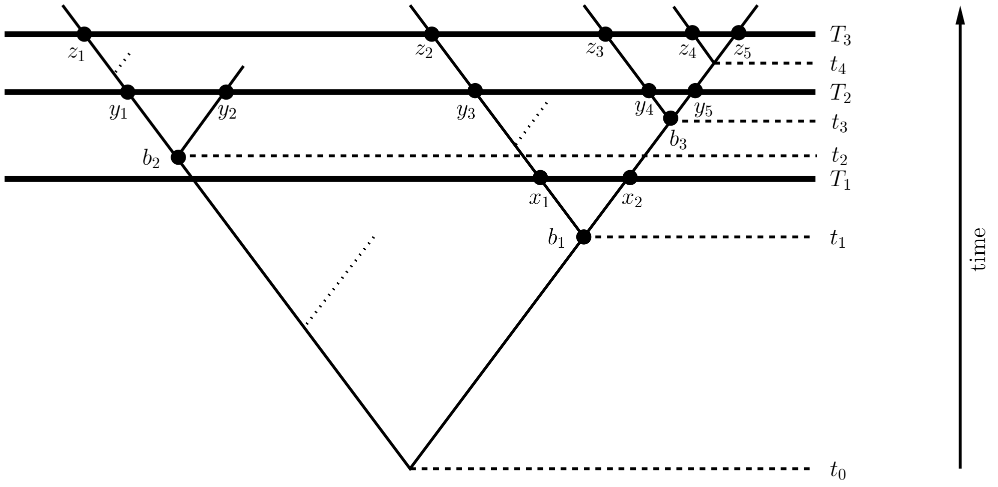


Figure 7.1: A rooted binary phylogenetic tree with three times T_1, T_2, T_3 at which taxa have been sampled. The dotted branches refer to taxa that do not contribute to the expected distances from one of these times to another and thus are not taken into account. On the other hand, bifurcation b_2 at time t_2 shows that extinction may have an impact on the expected values. Such branches have to be considered.

Example 7.1. Consider the tree given in Figure 7.1. Here, the paths using root b_1 (corresponding to time t_1) to connect species from time T_1 with species of time T_2 are $(x_1, b_1, x_2, b_3, y_4)$, $(x_1, b_1, x_2, b_3, y_5)$ and (x_2, b_1, x_1, y_3). We get $P_1^{1,2} = n_{1,1}^l \cdot n_{2,1}^r + n_{1,1}^r \cdot n_{2,1}^l = 1 \cdot 2 + 1 \cdot 1 = 3$, which equals the number of paths. Similarly, the paths along root b_1 connecting species from time T_1 with those of time T_3 are $(x_1, b_1, x_2, b_3, y_4, z_3)$, $(x_1, b_1, x_2, b_3, y_5, z_4)$, $(x_1, b_1, x_2, b_3, y_5, z_5)$ and $(x_2, b_1, x_1, y_3, z_2)$, and accordingly we have $P_1^{1,3} = 1 \cdot 3 + 1 \cdot 1 = 4$. Similarly, the paths along b_1 connecting species from times T_2 with those of time T_3 are $(y_3, x_1, b_1, x_2, b_3, y_4, z_3)$, $(y_3, x_1, b_1, x_2, b_3, y_5, z_4)$, $(y_3, x_1, b_1, x_2, b_3, y_5, z_5)$, $(y_4, b_3, x_2, b_1, x_1, y_3, z_2)$ and $(y_5, b_3, x_2, b_1, x_1, y_3, z_2)$, and we get $P_1^{2,3} = 1 \cdot 3 + 2 \cdot 1 = 5$.

Note that if $P_i^{j,k}$ is 0, there is a branch descending from b_i which does not contribute to the expected distance from one time to another (cf. Figure 7.1). We can therefore assume without loss of generality that all bifurcations b_i have at least one descendant on their left-hand side and at least one on their right-hand side, each in at least one of the times T_1, T_2, T_3. This means $n_{j,i}^l > 0$ and $n_{k,i}^r > 0$ for at least one $j \in \{1, 2, 3\}$ and at least one $k \in \{1, 2, 3\}$. In Figure 7.1, branches that therefore need not be considered are represented with dotted lines.

7.2 An Approach based on Evolutionary History

In this chapter, we make the additional simplifying assumption that, within the limited group of taxa under consideration (and over the limited time period being considered), the expected degree of morphological divergence between two taxa is proportional to the total amount of evolutionary history separating those two taxa. This evolutionary history is simply the time obtained by adding together the two time periods from the most recent common ancestor of the two taxa until the times from which each was sampled (in the case where one taxon is ancestral to the other, this is simply the time between the two samples). This assumption on morphological diversity would be valid (in expectation) if we view morphological distance as being proportional to the number of discrete characters that two species differ on, provided that two conditions hold: (i) each character has a constant rate of character state change (substitution) over the time T represented by the tree (i.e., the height of the tree), and (ii) T is short enough that the probability of a reverse or convergent change at any given character is low. We require these conditions to hold in the proofs of the following results.

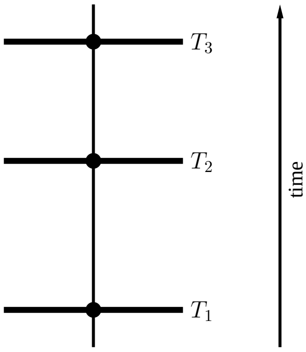


Figure 7.2: When the tree consists of only one lineage from which samples are taken at times T_1, T_2 and T_3, then clearly the distance $d_{1,3}$ is always larger than $d_{1,2}$ and $d_{2,3}$. Consequently, $E_{1,3} > \max\{E_{1,2}, E_{2,3}\}$.

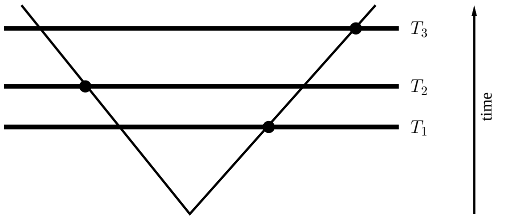


Figure 7.3: For samples taken from different lineages of a tree, the distance $d_{1,3}$ of one particular sample from time T_1 to the one of T_3 can be smaller than the distance of either of them to the sample taken at time T_2. Yet in expectation we always have $E_{1,3} > \max\{E_{1,2}, E_{2,3}\}$ for two-branch trees. For more complex trees this can fail as we show in Example 7.7.

The simplest scenario is the case where the samples from the three different times all lie on the same lineage, so that the evolutionary tree can be regarded as a path (cf. Figure 7.2). In this case, the path distance (and hence expected morphological distance) between the outer two fossils is always larger than the distance that either of them has from the fossil sampled from an intermediate time. But for samples that straddle bifurcations in a tree, it is quite easy to imagine how this intermediacy could fail; for example, if the two outer taxa lie on one branch of the tree and the fossil from the intermediate time lies on another branch far away (cf. Figure 7.3). But this example might be unlikely to occur, and indeed we will see that if sampling is uniform across the tree at any given time, in expectation the morphological distances remain intermediate even for this case (cf. Figure 7.3). Yet for more complex trees, this expected outcome can fail, and perhaps most surprisingly, the distance between the earliest and latest sample can, in expectation, be the *smallest* of the three distances in certain extreme cases.

Thus, in order to make general statements, we will consider the expected degree of relatedness of fossils sampled randomly from given times. Our results will depend solely on the tree shape (including branch lengths) of the underlying tree and the chosen times.

In the sampling, select uniformly at random one of the T_i-lineages as well as one of the T_j-lineages to get the expected length $E_{i,j}$ of the path connecting a lineage at time T_i with one at time T_j in the underlying phylogenetic tree. Then, the expectation that a fossil from the intermediate time T_2 also will be an intermediate taxon of two taxa taken from T_1 and T_3, respectively, refers to the assumption that $E_{1,3} > \max\{E_{1,2}, E_{2,3}\}$. We will show in the following lemma that this last inequality can fail and describe the precise condition for this to occur. Moreover, we later show that $E_{1,3}$ can be strictly smaller (!) than both $E_{1,2}$ and $E_{2,3}$ - that is the temporally most distant samples can, on average, be more similar than the temporally intermediate sample is to either of the two.

In order to simplify the statement of our results, for all bifurcations b_i set

$$Q_i^{j,k} := \frac{2 \cdot P_i^{j,k}}{n_j n_k} \quad \text{for all} \quad j,k \in \{1,2,3\}, j \neq k.$$

We are now in the position to state the first lemma.

Lemma 7.2. *Given a rooted binary phylogenetic tree with times* $0 < T_1 < T_2 < T_3$ *and the root at time* $t_0 = 0$. *Then,* $E_{1,3} \leq E_{1,2}$ *if and only if*

$$T_3 - T_2 \leq \sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i.$$

Proof.

$$E_{1,3} = \frac{1}{n_1 n_3} \left(\underbrace{n_3 (T_3 - T_1)}_{\substack{\text{every } T_3\text{-lineage} \\ \text{has an ancestor} \\ \text{in } T_1}} + \sum_{i:0<t_i<T_1} \left[P_i^{1,3} (T_3 - T_1 + 2(T_1 - t_i)) \right] + \underbrace{P_0^{1,3} (T_3 + T_1)}_{\text{ways along the root}} \right) \tag{64}$$

In the above bracket, the three summands refer to different paths from time T_1 to time T_3. The first summand belongs to those paths that go directly from T_1 to T_3 and thus have length $T_3 - T_1$. There are n_3 such ways as every T_3-lineage has an ancestor in T_1. The second summand sums up all paths going along one of the bifurcations b_i for $i \neq 0$. For every i, there are by definition exactly $P_i^{1,3}$ such paths. Similarly, the third summand refers to all paths along the root b_0, whose lengths are determined by taking the distance from T_1 to the root plus the distance from there to T_3.

As there are altogether $n_1 n_3$ different paths from T_1 to T_3 in the tree, we have:

$$n_3 + \sum_{i:0<t_i<T_1} P_i^{1,3} + P_0^{1,3} = n_1 n_3. \tag{65}$$

Then, by (64) and (65), we get

$$E_{1,3} = \frac{1}{n_1} \cdot \frac{1}{n_3} \cdot \left(n_1 n_3 T_3 + (n_1 n_3 - 2n_3) T_1 - 2 \cdot \sum_{i:0<t_i<T_1} P_i^{1,3} t_i \right),$$

and thus

$$E_{1,3} = T_3 + \frac{n_1 - 2}{n_1} T_1 - \sum_{i:0<t_i<T_1} Q_i^{1,3} t_i. \tag{66}$$

Analogously,

$$E_{1,2} = T_2 + \frac{n_1 - 2}{n_1} T_1 - \sum_{i:0<t_i<T_1} Q_i^{1,2} t_i. \tag{67}$$

Thus, with (66) and (67), we can conclude:

$$\begin{aligned} E_{1,3} \leq E_{1,2} \quad \Leftrightarrow \quad T_3 - \sum_{i:0<t_i<T_1} Q_i^{1,3} t_i \quad &\leq \quad T_2 - \sum_{i:0<t_i<T_1} Q_i^{1,2} t_i, \\ \Leftrightarrow \qquad T_3 - T_2 \qquad &\leq \quad \sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i. \end{aligned}$$

□

Corollary 7.3. *For a given tree there exist times* $0 < T_1 < T_2 < T_3$ *such that* $E_{1,3} \leq E_{1,2}$ *if and only if* $\sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i > 0$.

Proof. If $\sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i \leq 0$, then by Lemma 7.2 we need $T_2 \geq T_3$ in order to get $E_{1,3} \leq E_{1,2}$. Hence, there are no values $0 < T_1 < T_2 < T_3$ such that $T_3 - T_2$ fulfills the required condition, and so $E_{1,3} > E_{1,2}$ for all choices of T_i. Conversely, suppose $\sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i > 0$. Then, select T_1, T_2 with $0 < T_1 < T_2$ and set

$$T_3 := \frac{1}{2} \cdot \sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i + T_2$$

Then, $T_3 > T_2$ and

$$T_3 - T_2 = \frac{1}{2} \cdot \sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i \leq \sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i.$$

By Lemma 7.2, this choice of $0 < T_1 < T_2 < T_3$ leads to $E_{1,3} \leq E_{1,2}$. □

Corollary 7.4. *If either (i) $n_1 = 2$ or (ii) no extinction occurs in the tree and $n_2 = n_3$, then $E_{1,3} > E_{1,2}$.*

Proof. (i) Note that if $n_1 = 2$, obviously only one bifurcation, say $b_{\hat{\imath}}$ (for some $\hat{\imath}$ such that $0 \leq t_{\hat{\imath}} < T_1$), contributes to the number n_1 of lineages at time T_1, all the branches added by additional bifurcations become extinct before T_1. Thus: $P_{\hat{\imath}}^{1,3}, P_{\hat{\imath}}^{1,2} \neq 0$ and $P_i^{1,3}, P_i^{1,2} = 0$ for all $i \neq \hat{\imath}$.

Analogously to the proof of Lemma 7.2 we have for $n_1 = 2$: $n_1 n_3 = 2n_3 = n_3 + P_{\hat{\imath}}^{1,3}$ and $n_1 n_2 = 2n_2 = n_2 + P_{\hat{\imath}}^{1,2}$. Thus, $n_2 = P_{\hat{\imath}}^{1,2}$ and $n_3 = P_{\hat{\imath}}^{1,3}$. Therefore, $Q_{\hat{\imath}}^{1,2} = Q_{\hat{\imath}}^{1,3} = \frac{2}{n_1}$ and $Q_i^{1,2} = Q_i^{1,3} = 0$ for all $i \neq \hat{\imath}$. Thus, $\sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i = 0$ and it follows with Corollary 7.3 that $E_{1,3} > E_{1,2}$.

(ii) In this case, obviously $Q_i^{1,2} = Q_i^{1,3}$ for all $i : 0 < t_i < T_1$ and therefore $\sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2}) t_i = 0$. Thus, by Corollary 7.3, $E_{1,3} > E_{1,2}$.

□

Lemma 7.2 essentially states that the expected degree of relatedness from taxa of time T_1 to taxa of time T_3 can be larger than the one to taxa of time T_2, but it requires the distance from T_2 to T_3 to be 'small enough'. Whether such a solution is feasible can be checked via Corollary 7.3. Lemma 7.2 shows already how the role of intermediates depends on the times the fossils are taken from. Corollary 7.4(i) on the other hand shows how the tree itself has an impact on the expected values: if the tree shape (including branch lengths) is such that at time T_1 only two taxa exist, then the just mentioned scenario cannot happen as the condition of Corollary 7.3 is not fulfilled.

However, we can prove an even stronger result, namely that not only $E_{1,3} < E_{1,2}$ is possible, but $E_{1,3} < \min\{E_{1,2}, E_{2,3}\}$ can be obtained for a suitable choice of times T_1, T_2, T_3. For this, we need the following lemma.

Lemma 7.5. *Given a rooted binary phylogenetic tree with times* $0 < T_1 < T_2 < T_3$ *and the root at time* $t_0 = 0$. *Then* $E_{1,3} \leq E_{2,3}$ *if and only if*

$$\frac{n_2-2}{n_2}T_2 - \frac{n_1-2}{n_1}T_1 \;\geq\; \sum_{i:0<t_i<T_1}(Q_i^{2,3} - Q_i^{1,3})t_i + \sum_{i:T_1<t_i<T_2} Q_i^{2,3}t_i$$

Proof. As in the proof of Lemma 7.2, we have (cf. (66))

$$E_{1,3} = T_3 + \frac{n_1-2}{n_1}T_1 - \sum_{i:0<t_i<T_1} Q_i^{1,3}t_i. \tag{68}$$

$$\text{Analogously,}\quad E_{2,3} = T_3 + \frac{n_2-2}{n_2}T_2 - \sum_{i:0<t_i<T_2} Q_i^{2,3}t_i. \tag{69}$$

$$\text{Thus,}\quad E_{1,3} \leq E_{2,3}\quad \text{if and only if}$$

$$\frac{n_1-2}{n_1}T_1 - \sum_{i:0<t_i<T_1} Q_i^{1,3}t_i \;\leq\; \frac{n_2-2}{n_2}T_2 - \sum_{i:0<t_i<T_2} Q_i^{2,3}t_i,$$

which holds precisely if

$$\frac{n_2-2}{n_2}T_2 - \frac{n_1-2}{n_1}T_1 \;\geq\; \sum_{i:0<t_i<T_1}(Q_i^{2,3} - Q_i^{1,3})t_i + \sum_{i:T_1<t_i<T_2} Q_i^{2,3}t_i.$$

□

With the help of the two lemmas we can now state the following theorem.

Theorem 7.6. *Given a rooted binary phylogenetic tree with times* $0 < T_1 < T_2 < T_3$ *and the root at time* 0. *Then,* $E_{1,3} \leq \min\{E_{1,2}, E_{2,3}\}$ *if and only if the following two conditions hold:*

$$\text{(i) } T_3 - T_2 \leq \sum_{i:0<t_i<T_1} (Q_i^{1,3} - Q_i^{1,2})t_i,$$

$$\text{(ii) } \tfrac{n_2-2}{n_2}T_2 - \tfrac{n_1-2}{n_1}T_1 \ \geq \sum_{i:0<t_i<T_1} (Q_i^{2,3} - Q_i^{1,3})t_i + \sum_{i:T_1<t_i<T_2} Q_i^{2,3}t_i.$$

Proof. The Theorem follows directly from Lemmas 7.2 and 7.5. □

The following example demonstrates the influence of times $0 < T_1 < T_2 < T_3$ according to the above theorem.

Example 7.7. *Consider again Figure 7.1.*

1. *Assume* $t_1 = 15, T_1 = 100, t_2 = 107, t_3 = 109, T_2 = 110, T_3 = 130$. *Then,* $E_{1,2} = 137.33$, $E_{2,3} = 155.28$ *and* $E_{1,3} = 155.33$. *Hence, for this choice of times, we have* $E_{1,3} > \max\{E_{1,2}, E_{2,3}\}$.

2. *Consider the same times as in the previous case, but choose* $T_2 = 129$ *instead of* $T_2 = 110$. *This means to move* T_2 *further away from* T_1 *and closer to* T_3. *This change is enough to give completely different expected values:* $E_{1,2} = 156.33$, $E_{2,3} = 166.68$ *and* $E_{1,3} = 155.33$. *Hence, for this choice of times, we have* $E_{1,3} < \min\{E_{1,2}, E_{2,3}\}$.

7.3 Cladistic Approach

In the previous section, we regarded the degree of morphological divergence between two taxa to be proportional to the amount of evolutionary history separating these taxa, or, in other words, to the sum of the distances of each of the taxa to their most recent common ancestor. However, even in settings where this relationship holds, it is sometimes biologically more appropriate to define the degree of relatedness of different taxa according to the clades of the tree on which they are located. A clade is a so-called monophyletic group of taxa, i.e., a single common ancestor and all its descendants (cf. [5]). It often can be assumed that taxa from the same clade are more related to one another than to taxa on different clades. The disadvantage of the distance definition based on evolutionary history as given in the previous section is that it disregards such cladistic relationships. This is illustrated by Figure 7.4: here, the branches of the cherry (1,2) are very long, so in terms of evolutionary history, taxa 1 and 2 are not as closely related as either one of them is to taxon 3. The cladistic view, however, is that since taxa 1 and 2 are in one clade, which is even a cherry, they should be more closely related to one another than to taxon 3, as taxon 3 is in a different clade.

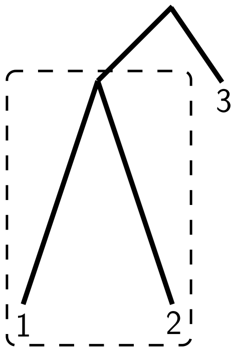


Figure 7.4: Taxa 1 and 2 form a cherry, i.e., a clade of size 2 (as indicated by the dotted box), whereas taxon 3 is on a different clade. In the cladistic view, taxa 1 and 2 are regarded more closely related than either of them to taxon 3. However, because of the branch lengths of the two cherry branches, regarding the amount of evolutionary history, taxon 3 is more related to taxon 1 and 2 than the latter two are to one another.

So, naturally the question arises whether anomalies in the fossil record can also be expected if the degree of relatedness of the different taxa is defined in a more cladistic way. We will therefore introduce a cladistic distance definition and show that it is in fact a tree metric. Then, just as in the previous section, we will provide exact formulas for the cladistic expected values $E^c_{1,2}, E^c_{1,3}$ and $E^c_{2,3}$ and show that there are indeed cases in which again $E^c_{1,3} < \min\{E^c_{1,2}, E^c_{2,3}\}$.

Definition 7.8. *For a given binary phylogenetic X-tree $\mathcal{T}$ with vertex set V and times $0 < T_1 < T_2 < T_3$ and for every $v \in \tilde{V} := V \cup S_1 \cup S_2 \cup S_3$ such that $v \neq b_0$, where b_0 is the root of the tree and S_i is the set of species present at time T_i (with $i \in \{1, 2, 3\}$), we define*

$$k_v := |\{x \in X | x \textit{ is a descendant of } v\ \}| - \delta_i(v),$$

where $\delta_i(v) = 1$ if $deg(v) \in \{1, 3\}$ and 0 else. For the root b_0, we set $k_{b_0} := |X| - 1$.

Then, we define the cladistic distance *$d^c : \tilde{V} \times \tilde{V} \to \mathbb{N}$ as follows:*

$$d^c(v, w) := \begin{cases} 0 & \textit{if } v = w \\ k_u & \textit{else (where u is the most recent common ancestor of v and w).} \end{cases}$$

Moreover, for every bifurcation b_i, we set $s_i := k_{b_i}$, and we define

$$k^i_v := |\{w | w \textit{ is a descendant of } v \textit{ present at time } T_i\}|.$$

Note that this definition assigns a distance of the clade size minus 1 to all taxa of a clade except for those which also lie together on a smaller clade. In particular, if two taxa form a cherry, their distance will be $2 - 1 = 1$, and if a clade consists of three taxa, the distance of the two taxa in the cherry to one another is still 1, whereas their distance to the third taxon will be 2, and so on. For an internal bifurcation b_i, the distance of this bifurcation to any node on the clade induced by b_i is one less than the number of leaves which are descendants of b_i, i.e., if b_i is the root of a cherry, it has distance $2 - 1 = 1$ to all its descendants. However, for lineages v present at times T_1 or T_2, we have to distinguish between v being a bifurcation (which has node degree 3), in which case v is treated like b_i, or v being a node of degree 2, in which case only one lineage descends from v and either ends in a taxon or in a bifurcation, in which case all taxa descending from v are also descendants of this bifurcation. Hence, the distance of v to its descending taxa is set to be one more than the distance of the corresponding bifurcation to these taxa.

In order to clarify the above cladistic distance definition, we now provide an example before we show that this distance is a tree metric (and even an ultrametric).

Example 7.9. Consider Figure 7.5. By Definition 7.8, we get $s_0 = 4$, $s_1 = 2$, $s_2 = 1$, $s_3 = 1$, $k_{v_1} = 3$, $k_{v_2} = 2$, $k_{w_1} = 2$, $k^1_{v_1} = 1$, $k^2_{v_1} = 3$, $k^3_{v_1} = 0$, $k^1_{v_2} = 1$, $k^2_{v_2} = 1$, $k^3_{v_2} = 2$, $k^2_{w_1} = 1$, $k^3_{w_1} = 2$. Using these values, we can calculate the cladistic distances, e.g. $d^c(A, C) = 2$, $d^c(D, E) = 1$, $d^c(A, D) = 4$, $d^c(b_1, A) = 2$, $d^c(v_1, A) = 2 + 1 = 3$ and $d^c(v_2, D) = d^c(w_1, D) = 1 + 1 = 2$.

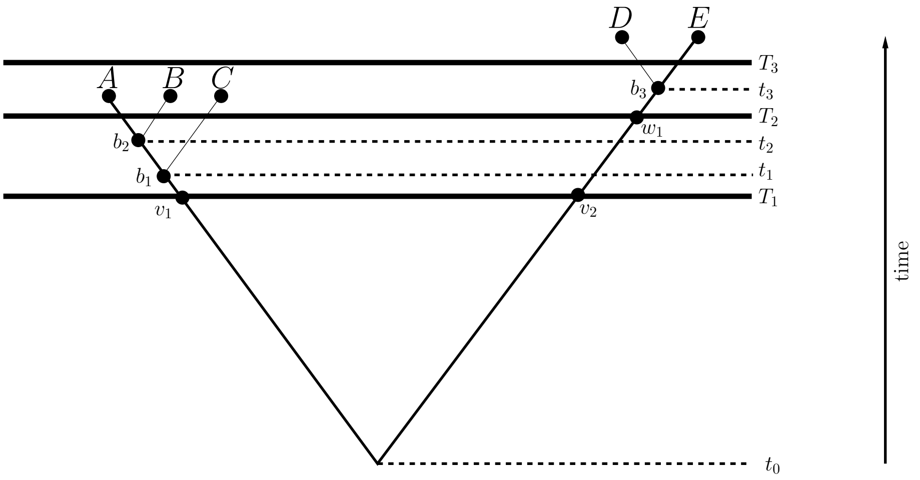


Figure 7.5: Illustration for the cladistic approach. Note that the cladistic distance from v_2 to D is equal to that of w_1 to D, as both v_2 and w_1 are only ancestors of b_3 and no other bifurcation.

Next we show that d^c is an ultrametric (and thus also a tree metric).

Proposition 7.10. *The cladistic distance function $d^c : X \times X \to \mathbb{N}$ as defined in Definition 7.8 is an ultrametric (and therefore also a tree metric) on $\mathcal{T}$.*

Proof. We show that for any distinct taxa A, B and $C \in X$ two of the distances $d^c(A, B)$, $d^c(A, C)$ and $d^c(B, C)$ are equal and not less than the third. Let A, B, $C \in X$. Two of these three taxa are on a smaller clade than the one they share with the third taxon (in order to see this, disregard all other taxa – then what remains is a 3-taxa tree, which has a cherry formed by two of the taxa). Wlog. we assume that the clade containing A and B is smaller than the clade containing A, B and C. Note that then by Definition 7.8, we have $d^c(A, C) = d^c(B, C)$. Then, if there are c_1 taxa additional to A and B on the smaller clade (see Figure 7.6), we have $d^c(A, B) = c_1 + 1$.

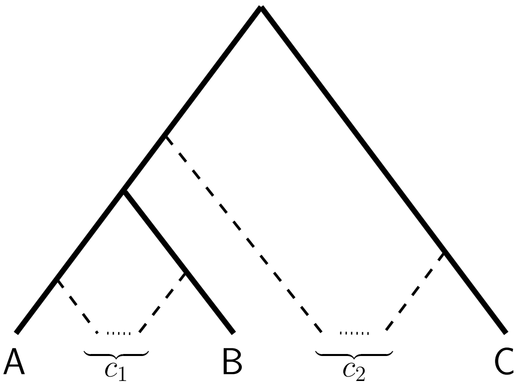


Figure 7.6: The cladistic distance function is an ultrametric. Two of the three taxa A, B and C (here: A and B) are on a smaller clade than the one they share with the third taxon. The dotted edges represent all other taxa belonging to a clade, e.g. there are c_1 taxa additional to A and B which belong to the clade induced by A and B.

Similarly, if there are c_2 taxa additional to A, B, C and the c_1 taxa of the small clade on the larger clade (see Figure 7.6), we have $d^c(A, C) = d^c(B, C) = c_1 + c_2 + 2 > c_1 + 1 = d^c(A, B)$. Therefore, d^c is an ultrametric. By [38], Chapter 7.2, this implies that d^c is also a tree metric. This completes the proof. □

As before, we now select uniformly at random one of the T_i-lineages as well as one of the T_j-lineages to get the expected cladistic distance $E^c_{i,j}$ of a species present at time T_i to one at time T_j in the underlying phylogenetic tree. Again, the expectation that a fossil from the intermediate time T_2 also will be an intermediate taxon of two taxa taken from T_1 and T_3, respectively, refers to the assumption that $E^c_{1,3} > \max\{E^c_{1,2}, E^c_{2,3}\}$. We will show in the following that this last inequality can fail and that in fact it is possible to get $E^c_{1,3} < \min\{E^c_{1,2}, E^c_{2,3}\}$ - that is the temporally most distant samples can, on average, be more similar than the temporally intermediate sample is to either of the two, even in the cladistic setting.

Lemma 7.11. *For a given binary phylogenetic X-tree and for times $0 < T_1 < T_2 < T_3$, let d^c be the cladistic distance function as defined in Definiton 7.8. Then, the expected cladistic distance $E^c_{j,l}$ for $j, l \in \{1, 2, 3\}$, $j < l$, can be written as follows:*

$$E^c_{j,l} = \frac{1}{n_j n_l} \cdot \left(\sum_{i:t_i<T_j} (P_i^{j,l} \cdot s_i) + \sum_{v \in S_j} (k_v^l \cdot k_v) \right),$$

where S_j is the the set of species present at time T_j.

Proof. In the tree, there are $n_j n_l$ distinct paths connecting elements of S_j with elements of S_l (as by definition, $n_j = |S_j|, n_l = |S_l|$). Thus, the expected value has to be divided by this product. Moreover, as in the previous section the number of such ways employing b_i is denoted by $P_i^{j,l}$, and for every b_i (including the root b_0), any path from T_j to T_l along b_i connects taxa which are on the clade induced by b_i and not on any subclade. Therefore, the cladistic distance between such species is by definition s_i. This explains the first sum. The second sum is the sumation induced by direct ancestry: every species present in time T_l has an ancestor v in T_j. The distance to this ancestor is by definition k_v, and there are by definition k_v^l species present in time T_l which have v as an ancestor. So altogether we have:

$$E^c_{j,l} = \frac{1}{n_j n_l} \cdot \left(\underbrace{\sum_{i:t_i<T_j} (P_i^{j,l} \cdot s_i)}_{\substack{\text{distances} \\ \text{contributed} \\ \text{by all } b_i\text{'s}}} + \underbrace{\sum_{v \in S_j} (k_v^l \cdot k_v)}_{\substack{\text{distances} \\ \text{contributed} \\ \text{by direct} \\ \text{ancestry}}} \right).$$

This completes the proof. □

Note that the expected values provided by Lemma 7.11 do not depend on the distance between the times $0 < T_1 < T_2 < T_3$. Rather, they depend on the tree topology. The times are only relevant in terms of the number of species present at each time. This means that in this cladistic setting, there is no direct analog to Theorem 7.6. More precisely, in order to find an example for which $E^c_{1,3} < \min\{E^c_{1,2}, E^c_{2,3}\}$, it is *not* necessary to make a certain choice of *values* for T_1, T_2, T_3. By making a suitable choice of a tree topology, it is nevertheless possible to find examples for which fossils from T_2 will on expectation be less related to fossils of times T_1 and T_3 than the latter two to one another. We now provide such an example.

Example 7.12. Consider again the tree given in Figure 7.5 and the values calculated in Example 7.9. The additional values needed to calculate the expected distances for this example are $n_1 = 2$, $n_2 = 4$, $n_3 = 2$, $P^{1,2}_0 = 4$, $P^{1,3}_0 = 2$, $P^{2,3}_0 = 6$, $P^{2,3}_1 = 0$, $P^{2,3}_2 = 0$. Then, by Lemma 7.11, we get

$$E^c_{1,2} = \frac{1}{n_1 n_2} \cdot \left(\sum_{i:t_i<T_1} (P^{1,2}_i \cdot s_i) + \sum_{v\in S_1} (k^k_v \cdot k_v) \right) = \frac{1}{2 \cdot 4} \cdot (4 \cdot 4 + 3 \cdot 3 + 1 \cdot 2) = \frac{27}{8}$$

$$E^c_{2,3} = \frac{1}{n_2 n_3} \cdot \left(\sum_{i:t_i<T_2} (P^{2,3}_i \cdot s_i) + \sum_{v\in S_2} (k^k_v \cdot k_v) \right) = \frac{1}{4 \cdot 2} \cdot (6 \cdot 4 + 0 \cdot 2 + 0 \cdot 1 + 2 \cdot 2) = \frac{28}{8}$$

$$E^c_{1,3} = \frac{1}{n_1 n_3} \cdot \left(\sum_{i:t_i<T_1} (P^{1,3}_i \cdot s_i) + \sum_{v\in S_1} (k^k_v \cdot k_v) \right) = \frac{1}{2 \cdot 2} \cdot (2 \cdot 4 + 0 \cdot 3 + 2 \cdot 2) = \frac{24}{8}$$

Therefore, $E^c_{1,3} < \min\{E^c_{1,2}, E^c_{2,3}\}$, and thus for the tree given in Figure 7.5, fossils from times T_1 and T_3 can be expected to be more related to one another than to a taxon from the intermediate time T_2.

7.4 Interpretation

The analysis of the fossil record provides an insight into the history of species and thus into evolutionary processes. Stochastic models can provide a good approach to infer patterns of diversification, and they form a useful link between molecular phylogenetics and paleontology [29]. Such models would greatly benefit from incorporation of potential fossil ancestors and other extinct data points to infer patterns of evolution. In this chapter, we have applied a simple model-based phylogenetic approach to study the expected degree of similarity between fossil taxa sampled at intermediate times.

'Gaps' in the fossil record are problematic [37] as they can be interpreted as 'missing links'. Therefore, numerous studies concerning the adequacy of the fossil record have been conducted (see, for example, [8], [30], [45]), and it is frequently found that even the available fossil record is still incompletely understood. This is particularly true for ancestor-descendant relationships (see, for instance, [9], [15]). For example, Foote [15] reported the probability that a preserved and recorded species has at least one descendant species that is also preserved and recorded is on the order of 1% – 10%. This number is much higher than the number of identified ancestor-descendant pairs. Thus, it remains an important challenge to recognize such pairs [2]. This is also essential with regard to ancestor-intermediate-descendant triplets, as it is possible that there are in fact fewer 'gaps' than currently assumed, i.e., that intermediates are present but not yet recognized. Such issues have an important bearing on any conclusions our results might imply concerning the testing of hypotheses of continuous morphological evolution, or concerning the shape of the underlying evolutionary tree based on the non-existence of certain intermediates.

Another challenge is to investigate different phylogenetic models for describing the expected degree of morphological separation between different fossil taxa sampled at different times. Our findings of Chapter 7.2 strongly depend on the assumption that morphological diversification is proportional to the distance in the underlying phylogenetic tree. This is justified if morphological difference is proportional to the number of differing discrete characters, if each of these characters changes at a constant rate over the time period considered, and if homoplasy is rare. This last assumption requires the rate of character change to be sufficiently small in relation to the time period of the sampling – the appearance of reverse or convergent character states will lead to a more concave (rather than linear) relationship between morphological divergence and path distance. A similar concave relationship might be expected for continuous morphological evolution as described by neutral Brownian motion.

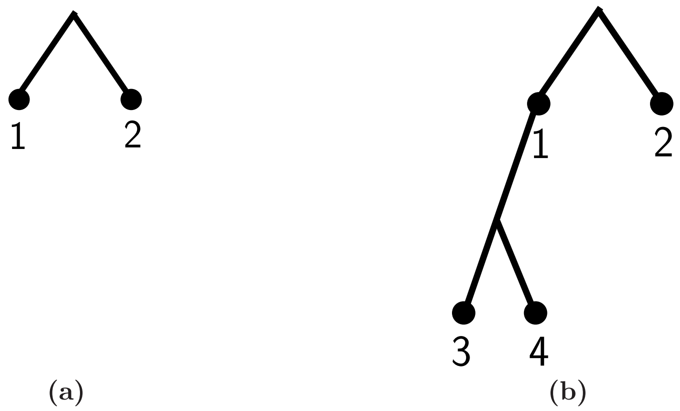


Figure 7.7: In the cladistic view, taxa 1 and 2 are more related in (a) than in (b), because in (a) they form a cherry, whereas in (b) the smallest clade containing 1 and 2 has three taxa (2, 3 and 4). However, the amount of evolution separating 1 and 2 is the same in both figures.

Our findings of Chapter 7.3, on the other hand, completely disregard the impact of the amount of evolutionary history that separates two species, as they consider clades to be the decisive factor responsible for the degree of relationship among different species. While this view might be biologically justified for some sets of species [5], it may cause problems when applied to intermediate species. Consider, for example, Figure 7.7. Here, the amount of history separating nodes 1 and 2 remains unchanged, but for the cladistic view it makes a difference if taxon 1 has descendants (and thus turns into an internal node). In particular, while the amount of evolution separating taxa 1 and 2 remains unchanged, their cladistic distance increases. This is somewhat counterintuitive and should preferrably be avoided. Therefore, an approach employing both the timewise distance as well as the cladistic information would probably describe the relationships amongst different species best. But since for both the pure cladistic as well as the pure history-based approach it was possible to construct examples in which the expected degree of relatedness of fossils from two distant times is less than to fossils of an intermediate time, we conjecture that even in a model that combines both approaches, such cases may exist. But this does not necessarily have to be bad news: in fact, gaps caused by a scenario as described in Chapter 7.2 may even give some faint hints on the shape and size of the underlying phylogenetic tree in the light of Theorem 7.6.

Chapter 8

Conclusion and Outlook

> I am still confused. But on a higher level.
>
> Enrico Fermi

The main purpose of this thesis was to present some very surprising properties of the most important tree inference techniques. Many of these 'surprises', such as presented in Chapters 4 and 5, deal with scenarios in which different methods lead to different trees. At first glance, such scenarios may seem more curious in a biological sense than in a mathematical one – as, from the mathematical point of view, one might wonder right away why different methods should lead to the same result. In this respect, examples in which Maximum Parsimony gives a tree which differs from the one induced by distance-based methods or by Maximum Likelihood might not seem too curious at all. Therefore, in this thesis we presented not only examples for such cases, but additionally the mathematical ideas which can be used to construct such examples for any number of taxa. Moreover, Chapter 5 shows that the likelihood support for a tree which is not most parsimonious can be made arbitrarily large under only slight modifications of the 'no common mechanism' N_r-model, even though without these modifications MP and ML are equivalent. Chapter 4 gives precise instructions on how to construct sequences which are homoplasy-free (and thus MP-wise 'best possible') on one tree and treelike (and thus distance-wise 'best possible') on a different tree – where both trees can be freely chosen. In this sense, our findings are indeed also mathematically surprising, because they show that arbitrarily many examples can be constructed and that these examples can make the gap between different methods arbitrarily large.

On the other hand, as mentioned above, some of the results presented in this thesis are even more surprising from a biological point of view. As demonstrated in Chapter 5, the often cited and well-known equivalence of MP and ML under the 'no common mechanism' N_r-model fails under certain additional assumptions of biological relevance – such as a molecular clock or bounded (and thus small) substitution probabilities. In the latter case, MP has traditionally been believed to be justified in the sense of agreement with ML (for a further discussion of this issue, the reader is referred to [12], pp. 100 ff.). Moreover, in Chapter 3 we showed that the Fitch algorithm sometimes provides a better estimation of the ancestral root state when some taxa are ignored – and that these taxa may even be arbitrarily close to the root. This is biologically counterintuitive as on short branches the probability of the conservation of the root state is relatively high, which is why it seems logical that MP should take them into account. Our example shows that, surprisingly, this is not in general true – in particular, if the rest of the tree fulfills certain properties, e.g. concerning balance, this part of the tree alone can provide a better estimate of the root state than when combined with other taxa.

Results that show how certain methods can disagree when inferring trees are not necessarily 'bad news' for phylogeneticists. They rather contribute to the general understanding of these methods and underline the importance of choosing the 'right' method and model for the analysis of particular data sets. In the same sense, Chapter 6 can be regarded as useful, as the sequence length bounds provided in this chapter may indicate that certain sequences simply are not long enough for a reliable tree reconstruction.

Chapter 7 is somewhat unrelated to the other chapters as it does not analyze existing methods or models but rather introduces a novel way to explain the degree of relatedness of different fossils using a simple stochastic model. Moreover, it is maybe the only chapter in this thesis whose results can be regarded as 'good news' already at first glance: here, we do not highlight disagreements of different methods, we rather explain why the perceived 'disagreement' of the patchy fossil record with the assumption of a continuous evolutionary process can in some cases (at least partially) be resolved. We show that some tree topologies and certain choices of times have the curious property that fossils from an intermediate time can be expected *not* to be morphologically intermediate. Additionally, we have shown that this property does not necessarily disappear when a more cladistic view is assumed. While the models used in this chapter are very simple, we are convinced that similar results can be achieved for more complex models, too – in particular if these combine the history-based and the cladistic view.

Solving mathematical problems may in a way be a thankless job: along with the satisfaction of having answered one question often comes the comprehension that ten more questions arise from the solution. For instance, finding a way to construct sequences of binary and ternary characters whose perfect phylogeny and perfect distance-based tree differ, was satisfactory in the sense that this result is best possible concerning the number of character states (as binary sequences are already known to be insufficient for this purpose). Similarly, finding a way of causing discrepancies between MP and ML for certain modifications of the N_r-model with no common mechanism was positive. And finally, the realization that these seemingly unrelated problems can be approached using similar techniques, and that even some sequences for which MP and and the derived Hamming distances disagree can be used to find an upper bound on the substitution probabilities such that MP and ML will also disagree for the same sequences, was astonishing. But, naturally, this cognition leads to many more questions: is there always an upper bound on the substitution probabilities such that for treelike sequences, ML will agree with the tree on which the distances are additive? What is the relationship of ML and distance-based methods like Neighbor-Joining in general? Are there binary sequences which lead, with an appropriate choice of an upper bound for the substitution probabilities or under the restriction of a molecular clock, to conflicting choices of MP and ML? These questions are beyond the scope of this thesis, but they certainly provide grounds for future research.

While we are confident that this thesis throws light on some surprising and curious aspects of phylogenetics, we also appreciate that even 35 years after Fitch's parsimony algorithm was first introduced and 30 years after the discovery of the Felsenstein zone, there are still properties even of the most frequently used and discussed tree inference methods which are unknown – which is why all these methods are still worth further investigation.

Appendix

Here we present in detail the analysis of the likelihood of the character $f_1 := aabb$ on tree $\mathcal{T}_1 := 12|34$ under the N_3-model (with state set $\{A, B, C\}$) with the help of the computer algebra system MAXIMA.

Notation: The pending edges of $\mathcal{T}_1$ are denoted by e_i, for leaves $i = 1, \ldots, 4$, with substitution probabilities p_i, respectively. We evaluate the likelihood by Felsenstein's postorder traversal [11] rooting the tree at leaf 1 (where the likelihood does not depend on the root position). The interior edge is denoted by $e_5 = (v, w)$, where w is the vertex at which the edges e_1 and e_2 are pending, and v is the vertex at which e_3 and e_4 are pending. The substitution probability for the interior edge e_5 is denoted by p_5. Moreover, we denote by Av, Bv and Cv the likelihood of character f_1 restricted to the vertices below v (that is, vertices 3 and 4) conditional on v being in state A, B or C, respectively. Similarly, Aw is the likelihood of the character restricted to 2,3,4 (which are the vertices below w) conditional on w being in state A, and so on. We denote by u the upper bound on the substitution probabilities. By Corollary 5.4 we know that the likelihood is maximized at a point where all substitution probabilities are either 0 or u, so these are the only cases for which we have to calculate the likelihood values.

MAXIMA CODE

```
Av(p3,p4) := p3*p4$

Bv(p3,p4) := (1-2*p3)*(1-2*p4)$

Cv(p3,p4) := p3*p4$

Aw(p3,p4,p5) := (1-2*p2)*((1-2*p5)*Av(p3,p4)+p5*Bv(p3,p4)+p5*Cv(p3,p4))$

Bw(p3,p4,p5) := p2*(p5*Av(p3,p4)+(1-2*p5)*Bv(p3,p4)+p5*Cv(p3,p4))$

Cw(p3,p4,p5) := p2*(p5*Av(p3,p4)+p5*Bv(p3,p4)+(1-2*p5)*Cv(p3,p4))$

A1(p1,p2,p3,p4,p5) := (1-2*p1)*Au(p3,p4,p5)+p1*Bu(p3,p4,p5)+p1*Cu(p3,p4,p5)$

for a:0 while a <= 1 do

for b:0 while b <= 1 do

for c:0 while c <= 1 do

for d:0 while d <= 1 do

for e:0 while e <= 1 do

display(ML(f_1,T_1,a*u,b*u,c*u,d*u,e*u) =

factor(expand(A1(a*u,b*u,c*u,d*u,e*u)/3)));
```

OUTPUT

1. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, 0, 0, 0, 0, 0\right) = 0$

2. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, 0, 0, 0, 0, u\right) = \frac{u}{3}$

3. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, 0, 0, 0, u, 0\right) = 0$

4. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, 0, 0, 0, u, u\right) = -\frac{u\,(2\,u-1)}{3}$

5. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, 0, 0, u, 0, 0\right) = 0$

6. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, 0, 0, u, 0, u\right) = -\frac{u\,(2\,u-1)}{3}$

7. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, 0, u, u, 0) = \frac{u^2}{3}$

8. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, 0, u, u, u) = \frac{u\left(3\,u^2-3\,u+1\right)}{3}$

9. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, u, 0, 0, 0) = 0$

10. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, u, 0, 0, u) = -\frac{u\,(2\,u-1)}{3}$

11. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, u, 0, u, 0) = 0$

12. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, u, 0, u, u) = \frac{u\,(2\,u-1)^2}{3}$

13. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, u, u, 0, 0) = 0$

14. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, u, u, 0, u) = \frac{u\,(2\,u-1)^2}{3}$

15. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, u, u, u, 0) = -\frac{u^2\,(2\,u-1)}{3}$

16. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, 0, u, u, u, u) = -\frac{u\,(2\,u-1)\left(3\,u^2-3\,u+1\right)}{3}$

17. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, 0, 0, 0, 0) = 0$

18. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, 0, 0, 0, u) = -\frac{u\,(2\,u-1)}{3}$

19. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, 0, 0, u, 0) = 0$

20. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, 0, 0, u, u) = \frac{u\,(2\,u-1)^2}{3}$

21. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, 0, u, 0, 0) = 0$

22. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, 0, u, 0, u) = \frac{u\,(2\,u-1)^2}{3}$

23. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, 0, u, u, 0) = -\frac{u^2\,(2\,u-1)}{3}$

24. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, 0, u, u, u) = -\frac{u\,(2\,u-1)\left(3\,u^2-3\,u+1\right)}{3}$

25. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, u, 0, 0, 0) = \frac{u^2}{3}$

26. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, u, 0, 0, u) = \frac{u\left(3\,u^2-3\,u+1\right)}{3}$

27. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, u, 0, u, 0) = -\frac{u^2\,(2\,u-1)}{3}$

28. $\mathrm{ML}\,(\mathrm{f\_1}, \mathrm{T\_1}, u, u, 0, u, u) = -\frac{u\,(2\,u-1)\left(3\,u^2-3\,u+1\right)}{3}$

29. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, u, u, u, 0, 0\right) = -\frac{u^2\,(2\,u-1)}{3}$

30. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, u, u, u, 0, u\right) = -\frac{u\,(2\,u-1)\,\left(3\,u^2-3\,u+1\right)}{3}$

31. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, u, u, u, u, 0\right) = \frac{u^2\,\left(9\,u^2-8\,u+2\right)}{3}$

32. $\mathrm{ML}\left(\mathrm{f\_1}, \mathrm{T\_1}, u, u, u, u, u\right) = \frac{u\,\left(9\,u^4-15\,u^3+14\,u^2-6\,u+1\right)}{3}$

done

FINDING THE MAXIMUM

We now have to consider all 32 outputs. We will show that for $0 \le u < \frac{1}{3}$, the maximum value is that given above by output 2, namely $\frac{u}{3}$.

- Clearly, $0 \le \frac{u}{3}$, so outputs 1, 3, 5, 9, 11, 13, 17, 19 and 21 are not optimal.
- $\frac{u^2}{3} < \frac{u}{3}$ for all $u \le \frac{1}{3}$, so outputs 7 and 25 are less than $\frac{u}{3}$ and thus not optimal.
- We have $-(2u-1) = (1-2u) \le 1$ for $u \ge 0$. Therefore, outputs 4, 6, 10, 12, 14, 15, 18, 20, 22, 23, 27 and 29 are less than $\frac{u}{3}$ and thus not optimal.
- As $u \le \frac{1}{3}$, we have $3u > 3u^2$, and thus $(3u^2 - 3u + 1) < 1$. Therefore, outputs 8, 16, 24, 26, 28 and 30 are less than $\frac{u}{3}$ and thus not optimal.
- $u^2(9u^2 - 8u + 2) < u$ for all $u \le \frac{1}{3}$. Therefore, output 31 less than $\frac{u}{3}$ and thus not optimal.
- For $u \le \frac{1}{3}$, we have $(9u^4 - 15u^3 + 14u^2 - 6u + 1) < 1$. Therefore, output 32 is less than $\frac{u}{3}$ and thus not optimal.

So the maximum likelihood of f_1 on $\mathcal{T}_1$ is $\frac{u}{3}$. Observe that the likelihood is maximized when $p_5 = u$ and all other substitution probabilities are 0 (see output 2). So in this case, ML assigns only the internal edge a non-zero substitution probability, which is exactly the edge where MP would suggest a substitution.

List of Symbols

Symbol	**Meaning**
AA	ambiguous reconstruction accuracy of Maximum Parsimony
$A\|B$	X-split
b_i	bifurcation at a time t_i
bp	base pairs
c	character state
C	set of character states
$d(x,y)$	time-based distance between two taxa x and y
$d^c(x,y)$	cladistic distance between two taxa x and y
$d_S(x,y)$	Hamming distance between taxa x and y induced by S
$E(\mathcal{T})$	edge set of a tree $\mathcal{T}$
$E_{i,j}$	expected time-based distance of species of time T_i to species of time T_j
$E^c_{i,j}$	expected cladistic distance of species of time T_i to species of time T_j
f	character
$i_{d_S}(\sigma)$	isolation index of an X-split σ for a sequence S
g	extension of a character
$l_\mathcal{T}(f)$	parsimony score of a character f on a tree $\mathcal{T}$
$l_\mathcal{T}(g)$	parsimony score of an extension g of a character on a tree $\mathcal{T}$
$p_{e,i}$	substitution probability of a character f_i on an edge e
$P_i^{j,k}$	number of paths from time T_j to time T_k along node b_i
r	number of character states
RA	reconstruction accuracy of Maximum Parsimony
S	character sequence
σ	X-split
$\Sigma(X)$	set of X-splits
$\Sigma^*(X)$	set of non-trivial X-splits
$\mathcal{T}$	phylogenetic tree
T, t	fixed time
$\mathcal{T}_Y$	restriction of an X-tree $\mathcal{T}$ to a subset $Y \subseteq X$
u	upper bound on the substitution probability
UA	unambiguous reconstruction accuracy of Maximum Parsimony
$V(\mathcal{T})$	vertex set of a tree $\mathcal{T}$
X	set of leaves of a tree $\mathcal{T}$
wlog.	without loss of generality

List of Figures

List of Tables

Index

Publications Resulting from this Thesis

[BF08] H.-J. Bandelt and M. Fischer. Perfectly misleading distances from ternary characters. *Syst. Biol.*, 57(4):540–543, 2008.

[FS08] M. Fischer and M. Steel. Expected anomalies in the fossil record. *Evol. Bioinf. Onl.*, 4:61–67, 2008.

[FS09] M. Fischer and M. Steel. Sequence length bounds for resolving a deep phylogenetic divergence. *J. Theo. Biol.*, 256:247–252, 2009.

[FT09a] M. Fischer and B. Thatte. Maximum parsimony on subsets of taxa. *Submitted to J. Theo. Biol.*, 2009.

[FT09b] M. Fischer and B. Thatte. Revisiting an equivalence between maximum parsimony and maximum likelihood methods in phylogenetics. *Submitted to Bull. Math. Biol.*, 2009.

Bibliography

[1] Alon, N., and Spencer, J. *The probabilistic method.* John Wiley and Sons, New York, 2000.

[2] Alroy, J. Continuous track analysis: a new phylogenetic and biogeographic method. *Syst. Biol. 44* (1995), 152 – 178.

[3] Atteson, K. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica 25* (1999), 251 – 278.

[4] Bandelt, H.-J., and Dress, A. Reconstructing the shape of a tree from observed dissimilarity data. *Adv. Appl. Math. 7* (1986), 309–343.

[5] Benton, M., and Harper, D. *Basic paleontology.* Prentice Hall, 1997.

[6] Churchill, G., von Haeseler, A., and Navidi, W. Sample size for a phylogenetic inference. *Mol. Biol. Evol. 9(4)* (1992), 753–769.

[7] Darwin, C. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London, 1859.

[8] Durham, J. The incompleteness of our knowledge of the fossil record. *J. Paleont. 41* (1967), 559 – 565.

[9] Engelmann, G., and Wiley, E. The place of ancestor-descendant relationships in phylogeny reconstruction. *Syst. Zool. 26* (1977), 1 – 11.

[10] Felsenstein, J. Cases in which parsimony or compatibility will be positively misleading. *Syst. Zool. 27* (1978), 401–410.

[11] Felsenstein, J. Evolutionary trees from dna sequences: a maximum likelihood approach. *J. Mol. Evol. 17* (1981), 368–376.

[12] Felsenstein, J. *Inferring phylogenies.* Sinauer Associates, Massachusetts, 2004.

[13] Fischl, W. Distances that perfectly mislead: a practical approach. Project report, 2007.

[14] Fitch, W. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool. 20(4)* (1971), 406–416.

[15] Foote, M. On the probability of ancestors in the fossil record. *Paleobiol. 22(2)* (1996), 141 – 151.

[16] Hendy, M. The relationship between simple evolutionary tree models and observable sequence data. *Syst. Zool. 38* (1989), 310–321.

[17] Hendy, M., and Penny, D. A framework for the qualitative study of evolutionary trees. *Syst. Zool. 38* (1989), 297–309.

[18] Huber, K., Moulton, V., and Steel, M. Four characters suffice to convexly define a phylogenetic tree. *SIAM J. discr. Math. 18(4)* (2005), 835–843.

[19] Huson, D., and Steel, M. Distances that perfectly mislead. *Syst. Biol. 53(2)* (2004), 327 – 332.

[20] Jukes, T., and Cantor, C. Evolution of protein molecules. *In "Mammalian Protein Metabolism", New York Academic Press* (1969), 21–132.

[21] Lecointre, G., Philippe, H., van Le, H., and Le Guyader, H. How many nucleotides are required to resolve a phylogenetic problem? the use of a new statistical method applicable to available sequences. *Mol. Phyl. Evol. 3(4)* (1994), 292–309.

[22] Li, G., Steel, M., and Zhang, L. More taxa are not necessarily better for the reconstruction of ancestral character state. *Syst. Biol. 57(4)* (2008), 647 – 653.

[23] Lockhart, P., Novis, P., Milligan, B., Riden, J., Rambaut, A., and T., L. Heterotachy and tree building: a case study with plastids and eubacteria. *Mol. Biol. Evol. 23(1)* (2006), 40–45.

[24] Marshall, C. Confidence intervals on stratigraphic ranges. *Paleobiol. 16(1)* (1990), 1 – 10.

[25] Marshall, C. The fossil record and estimating divergence times between lineages: maximum divergence times and the importance of reliable phylogenies. *J. Mol. Evol. 30* (1990), 400 – 408.

[26] McDiarmid, C. On the method of bounded differences. *Surveys in Combinatorics, J. Siemons ed., London Mathematical Society Lecture Note Series 141, Cambridge University Press* (1989), 148–188.

[27] Mossel, E., and Steel, M. A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci. 187* (2004), 189–203.

[28] Mossel, E., and Steel, M. How much can evolved characters tell us about the tree that generated them? *In: Olivier Gascuel (ed.), Mathematics of Evolution and Phylogeny, Oxford University Press* (2005), 384–412.

[29] Nee, S. Extinct meets extant: simple models in paleontology and molecular phylogenetics. *Paleobiol. 30(2)* (2004), 172 – 178.

[30] Newell, N. Adequacy of the fossil record. *J. Paleont. 33* (1959), 488 – 499.

[31] Neyman, J. Molecular studies of evolution: A source of novel statistical problems. *In "Statistical Decision Theory and Related Topics", New York Academic Press* (1971), 1–27.

[32] Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. Phylogenomics. *Annu. Rev. Ecol. Evol. Syst. 36* (2005), 541–562.

[33] Rokas, A., and Carrol, S. Bushes in the tree of life. *PLoS Biology 4(11)* (2006), e352.

[34] Rozanov, Y. *Probability theory: a concise course.* Dover Publications, New York, 1969.

[35] Saitou, N., and Nei, M. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. *J. Mol. Evol. 24* (1986), 189–204.

[36] Salisbury, B., and Kim, J. Ancestral state estimation and taxon sampling density. *Syst. Biol. 50(4)* (2001), 557–564.

[37] Schoch, R. Gaps in the fossil record. *Nature 299* (1982), 490.

[38] Semple, C., and Steel, M. *Phylogenetics.* Oxford University Press, 2003.

[39] Sober, E. *Evidence and Evolution: the logic behind the science.* Cambridge Unversity Press, 2008.

[40] Steel, M., and Charleston, M. Five surprising properties of parsimoniously colored trees. *Bulletin of Mathematical Biology 57(2)* (1995), 367–375.

[41] Steel, M., and Szekely, L. Inverting random functions ii: explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discr. Math. 15(4)* (2002), 562–575.

[42] Strauss, D., and Sadler, P. Classical confidence intervals and bayesian probability estimates for ends of local taxon ranges. *Math. Geol. 21(4)* (1989), 411–427.

[43] Townsend, J. Profiling phylogenetic informativeness. *Syst. Biol. 56(2)* (2007), 222–231.

[44] Tuffley, C., and Steel, M. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol. 59* (1997), 581–607.

[45] Valentine, J. How good was the fossil record? clues from the californian pleistocene. *Paleobiol. 15* (1989), 83 – 94.

[46] Wortley, A., Rudall, P., Harris, D., and Scotland, R. How much data are needed to resolve a difficult phylogeny? case study in lamiales. *Syst. Biol. 54(5)* (2005), 696–709.

[47] Xia, X., Xie, Z., Salemi, M., L., C., and Wang, Y. An index of substitution saturation and its applications. *Mol. Phyl. Evol. 26* (2003), 1–7.

[48] Yang, J. *Three mathematical issues in reconstructing ancestral genome.* PhD thesis, National University of Singapore, 2008.

[49] Yang, Z. On the best evolutionary rate for phylogenetic analysis. *Syst. Biol. 47(1)* (1998), 125–133.

[50] Yang, Z. *Computational Molecular Evolution.* Oxford University Press, 2006.

[51] Zahl, S. Bounds for the central limit theorem error. *SIAM J. Appl. Math. 14(6)* (1966), 1225–1245.

[52] Zhang, J., and Nei, M. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol. 44* (1997), 139 – 146.